The New Encyclopædia Britannica

Volume 27

MACROPÆDIA

Knowledge in Depth

FOUNDED 1768
15 TH EDITION

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

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

THE UNIVERSITY OF CHICAGO

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

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

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

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

Printed in U.S.A.

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

CONTENTS

San Francisco

San Francisco holds a secure place in the United States' romantic dream of itself—a cool, elegant, handsome, worldly seaport whose steep streets offer breathtaking views of one of the world's great bays. San Franciscans, according to the dream, are sophisticates whose lives hold full measures of such civilized pleasures as music, art, and good food. Their city is a magic place, almost an island, saved by its location and its history from the overpowering ugliness that, again according to the dream, afflicts so much of the rest of urban California.

For all the truth there may be in this picture, there is, of course, also another San Francisco in which the per capita consumption of alcohol is the highest of any U.S. city, the suicide rate is higher than the national average, and the divorce rate is several times that of New York City. San Francisco is clearly a city in which the tensions between the dream and the reality have been costly.

Furthermore, since World War II, San Francisco has had to come to grips with the common urban problems of pollution of both the air and the water; the ugliness that comes from rampant building, violence, and vandalism; and the decay of the inner city. San Francisco has been shrinking as families, mainly white and middle-class, have moved to its suburbs, leaving the city to a population that, viewed statistically, tends to be older and to have fewer married people and fewer whites than the stereotype has it. Almost one of every two San Franciscans is, in the sterile term of the census taker, "nonwhite"—in this case black, East Asian, Filipino, Samoan, or American Indian. Many others are immigrants from Spanish America. Their dreams increasingly demand a realization that has little to do with the romantic dream of San Francisco.

But both the dreams and the realities are important, for they are interwoven in the fabric of the city that might be called Paradox-by-the-Bay.

This article is divided into the following sections:

Physical and human geography

THE LANDSCAPE

The city site. Hilly, roughly square, and about 46 square miles (120 square kilometres) in area, San Francisco occupies the northern tip of a peninsula. To its south are the bedroom suburbs of San Mateo County; to the east and northeast is the bay; to the west and northwest lies the Pacific Ocean.

The hills and dizzying streets

The most prominent of San Francisco's hills are Twin Peaks, Mt. Davidson, and Mt. Sutro, all more than 900 feet (270 metres) in height. The best known are Nob Hill, where the wealthy "nobs" built extravagant mansions in the 1870s, and Telegraph Hill, which once looked down on the Barbary Coast, a neighbourhood alive with gaudy wickedness. Thanks to the pioneer planners' prejudice in favour of a squared-off grid, the downtown streets march intrepidly up precipitous slopes, terrifying newly arrived drivers, making the cable cars more than sentimental anachronisms, and providing splendid views of the bay.

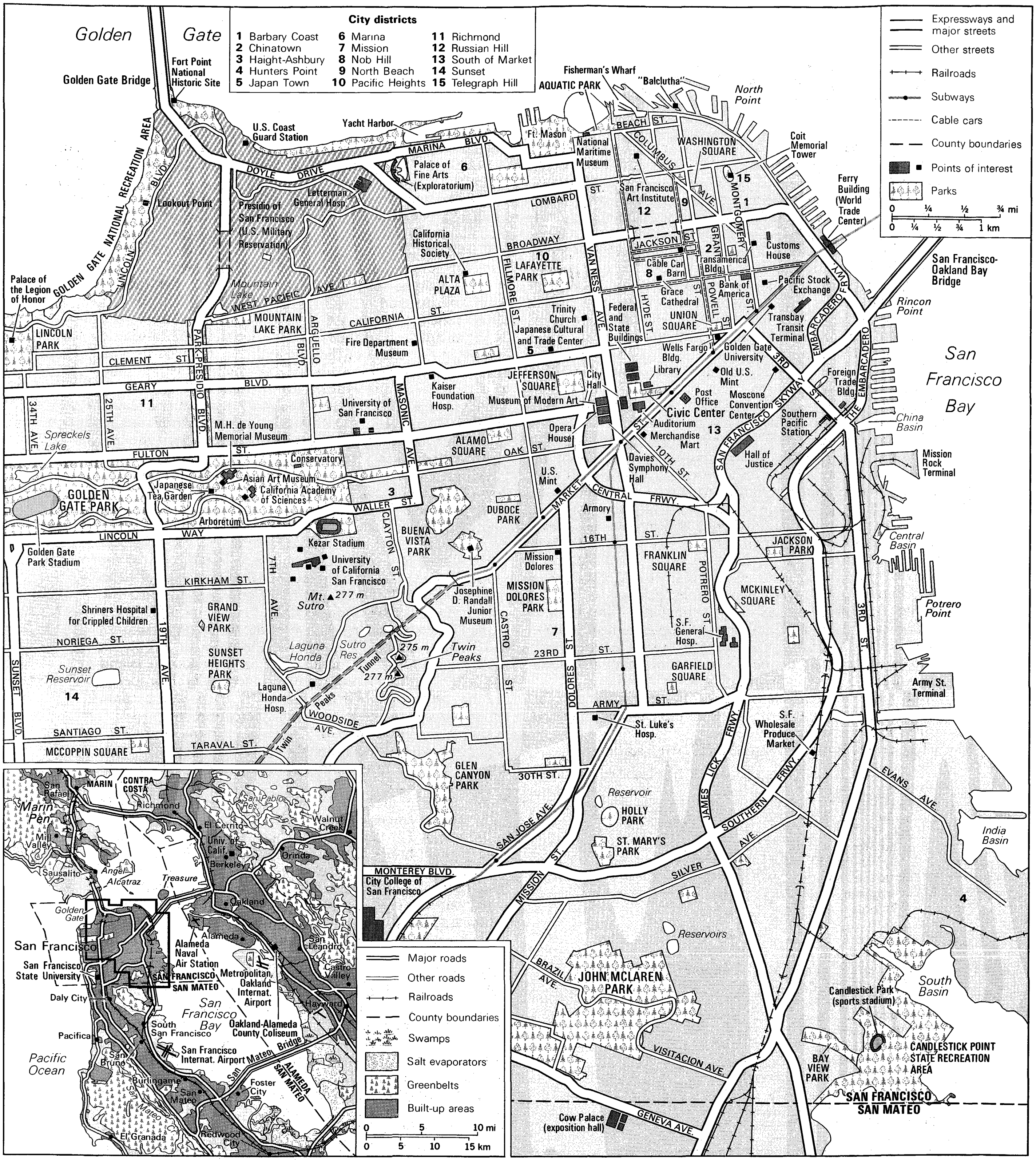
San Francisco Bay is a drowned river valley, submerged during the melting of the last glacial ice sheet. Enthusiastic and profitable filling of the tidelands has reduced its area at mean high tide from about 700 square miles in 1880 to a mere 435 square miles. More than half of the bay is still fillable, but in 1965 the state legislature created the Bay Conservation and Development Commission to control further landfill projects. At its widest point the bay measures 13 miles (21 kilometres) and at its deepest, in the Golden Gate channel, 357 feet. The maximum daily flow of water through the Golden Gate into the Pacific is seven times the flow of the Mississippi River at its mouth.

Within the portion of San Francisco Bay lying inside the city limits are the natural islands of Alcatraz and Yerba Buena and man-made Treasure Island, created for a world's fair in 1939 and later turned into a naval base. Alcatraz (Spanish: "Pelican") was from 1934 to 1963 the most notorious maximum-security, "escape-proof" prison in the United States. In 1969, after the decaying cell blocks had been given up by the Federal Bureau of Prisons, a multi-tribal group of American Indians invaded the island and asserted their rights to abandoned federal property, but they were forcibly evicted in 1971. The island became part of the Golden Gate National Recreation area in 1972.

Climate. Winter in San Francisco is rainy and mild, the spring sunny and mild, the summer foggy and cool, and the autumn sunny and warm. The average minimum temperature is 51° F (11° C), the average maximum, 63° F (17° C). The mean rainfall, almost all of which occurs between November and April, is about 21 inches (533 millimetres). The sun shines during two-thirds of the possible daylight hours. The most characteristic feature of the weather, however, is the summer fog, which lies low over the city until midday, creating consternation among shivering tourists. This fog is a phenomenon of temperature contrasts created when warm, moist ocean air comes in contact with cold water welling up from the ocean bottom along the coast.

The city layout. The central business district, the financial district, North Beach, and Chinatown occupy the site of the Gold Rush city, expanded by progressive fillings along the waterfront. The bones of many a ship deserted in 1849 now lie under business buildings several blocks inland. To the west, at the approach to the Golden Gate Bridge, lies the Presidio, a U.S. military reservation remarkable for its parklike lawns and stands of trees. South of the Presidio is Golden Gate Park, reclaimed from a one-time sandy desert. The rest of San Francisco is largely composed of the residential neighbourhoods, from Pacific Heights, in which the old, moneyed families live, to Hunter's Point, which is predominantly black.

A great change, which has been described as the Manhattanization of San Francisco, became apparent after the late 1960s, and it has been both welcomed and excoriated. In the financial district in particular, one tall building after another has been built in a city in which, for generations, few buildings were higher than 20 stories. Among the modern skyscrapers are the Bank of America Center (52 stories) and the Transamerica Corporation building (48 stories and a 212-foot spire), which rises to a point like an elongated pyramid. (K.La./G.C.Ha.)



Central San Francisco and (inset) its metropolitan area.

The great earthquake and fire of 1906

Concern has also arisen from the experience that San Francisco shares with no other U.S. city—destruction by earthquake. Severe quakes have been felt in 1868, 1898, 1900, 1906, and 1989. But it was the 1906 earthquake that did the most damage and that has become identified with the city. That quake, which occurred on April 18, was followed by a fire that destroyed the centre of town and burned for four days, until the ashes were wetted down by rain. Four square miles, making up 512 blocks in the centre of town, were gone, along with 28,000 buildings and a total property value of about $350,000,000. Approximately 700 people died; 250,000 were left homeless. Survivors camped in Golden Gate Park. An Eastern newspaperman, celebrating the survival of a local distillery, composed the verse, "If, as some say, God spanked the town / For being over frisky, / Why did he burn the Churches down / And save Hotaling's Whisky?"

Since the 1906 earthquake, seismologists and engineers have warned that it could happen again. Several relatively strong earthquakes (measuring more than 5.0 on the Richter scale) have since hit the city and caused little damage, but the 1989 quake (7.1 on the Richter scale) did destroy some structures within the city and even more in the surrounding areas. Modern office towers, however, were largely unaffected, indicating that new building methods may provide some protection for the city.

(K.La./G.C.Ha./Ed.)

THE PEOPLE

The pattern of immigration into San Francisco during the latter half of the 19th century was significantly different from that of anywhere else in the United States. The waves of newcomers included not only native-born Americans moving west but also Europeans arriving directly by ship without previous Americanization along the Eastern Seaboard. The demography of the gold-rush city was summed up concisely by a real-estate firm that advertised that it could "transact business in the English, French, German, Spanish and Italian languages." San Francisco remains one of the two most European of American cities—New Orleans is the other—and surely the most Mediterranean. Italians have remained the dominant European minority, followed by Germans, Irish, and British.

Ethnic minorities

The blacks. Before World War II about 20,000 blacks lived in the entire Bay Area, about 4,000 of them in San Francisco. The 24-fold increase during the next 30 years was set in motion by the war, which brought at least 500,000 war workers to the Bay Area's shipyards and other industries. Among them were tens of thousands from the South, who settled mainly in San Francisco, Oakland, and Richmond. In San Francisco they moved into the old Carpenter Gothic houses in the blocks around Fillmore Street, vacated when the Japanese were driven into wartime internment camps. By the 1980s the character of the district was shifting again as the renovation of these houses and the high cost of property caused rents to rise. Poorer black residents were being forced out of their neighbourhoods and into slum housing in the city's already crowded southeastern sector. An increasing number of black men and women have become prominent in the city's life, however, and blacks have won many elective offices.

The Chinese. Chinatown, which is said to be the largest Chinese community outside of Asia, is also probably the least understood minority community in the city. The colourful shops and restaurants of Grant Avenue mask a slum of crowded tenements and sweatshops that has the highest population density in an already densely populated city. Increasingly, Chinese have moved into North Beach, hitherto predominantly Italian, onto the nearby slopes of Russian Hill, or into the Richmond district north of Golden Gate Park. Many of those who reside in Chinatown are more recent immigrants, from Hong Kong in particular, who prefer their native tongue and way of life.

The Japanese. Never as large as Chinatown, the Japanese community of San Francisco was wiped out at a single stroke by the infamous Executive Order 9066 of 1942, which sent them, foreign-born and native alike, into "relocation centres" that were, in all but name, concentration camps. The present centre of the Japanese community is Japan Town (Nihonmachi), a few blocks east of Fillmore Street. There an ambitious trade and cultural centre has risen, with restaurants, a hotel, shops, and business establishments. Though the rising generation of Japanese-Americans go to Japan Town as visitors, bound for celebrations or to buy imported goods, their own roots are elsewhere.

The Spanish-speaking. Few visitors see the Mission district, which is a great irony because the historic origins of the city extend to Spain and Mexico and the Spanish-speaking population rivals the Chinese as the second greatest ethnic minority.

Before World War II the Mission district, named for the Mission Dolores, was principally blue-collar and Irish. The Irish have largely been replaced by Spanish-speaking immigrants, mainly from Central America and Mexico. Living among them are pockets of American Indians and Samoans.

The Filipinos. The Filipino community in San Francisco has grown remarkably since World War II and spread to all areas of the city, especially the South of Market area. Filipinos have become an active ethnic group, particularly in the fields of politics, education, and business. They are known for their high literacy rate and their love of the arts and music.

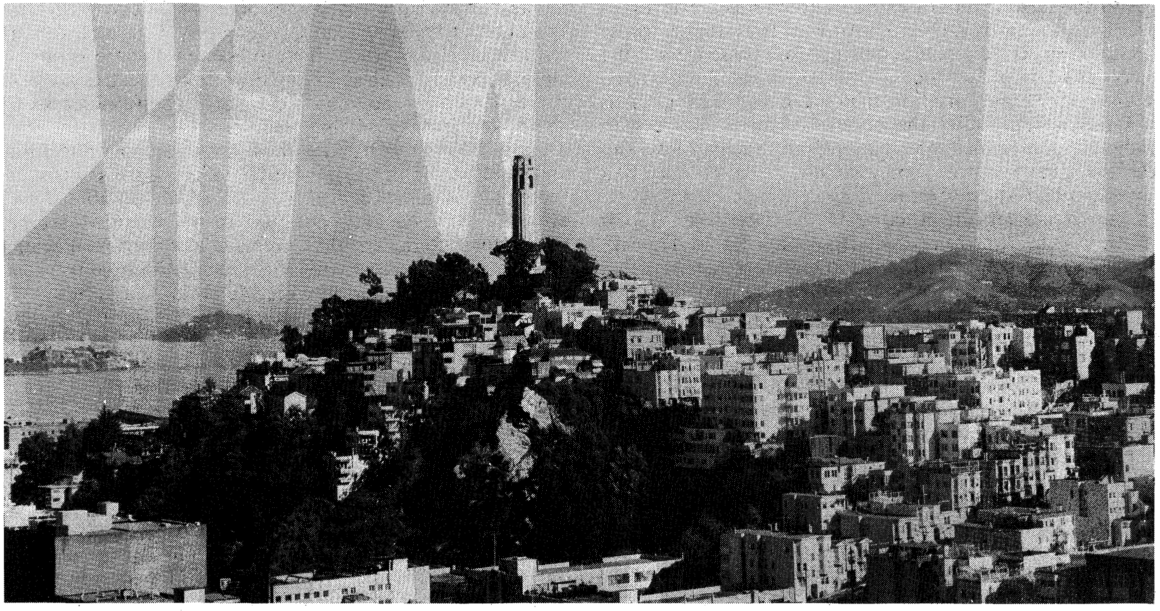
THE ECONOMY

The gold rush to California (1848–49) and other western territories and states in the mid-1800s established San Francisco as the premier city of the West, known from the Oregon border to the pueblo of Los Angeles simply as the City. It is still a great port, the financial and administrative capital of the West, and a substantial centre for commerce and manufacturing.

Components of the economy

A large portion of the city's employed work in the area of finance. Other leading areas of employment include business services (personnel supply, building maintenance, security, computers and data processing, advertising), retail trade, the tourist and convention industry, and professional service. Many companies have chosen to locate their national headquarters in San Francisco.

Colour Library International



The Coit Memorial Tower on Telegraph Hill, San Francisco; at left is Alcatraz Island in San Francisco Bay.

The port. From its beginnings as a port of call in the hide-and-tallow trade and, later, as the home port of the Pacific whale fishery, San Francisco has always been acutely conscious of the importance of shipping. In the 19th century ships sailed around Cape Horn or from the Isthmus of Panama, and "steamer day" was a civic institution; after 1914 cargo and passenger vessels arrived from the East by way of the Panama Canal. In 1867 the Pacific Mail Steamship Company opened the first transpacific service, sailing from San Francisco to Yokohama and Hong Kong.

Imports and exports passing through the San Francisco Customs District make the combined ports of San Francisco Bay—San Francisco, Oakland, Alameda, Sacramento, and Stockton—one of the most active international ports in the country.

Industry. Manufacturing is the main source of income in the Bay Area. In San Francisco, in which manufacturing is a lesser source of income, the principal industries are apparel and other textile products, food processing, and shipbuilding, while the aerospace and electronics industries are strong in the cities of the peninsula.

Finance. A financial centre since the first pinch of gold dust was exchanged for hard cash, San Francisco is the seat of the Pacific Stock Exchange as well as the headquarters of many banks, among them the Bank of America and the Wells Fargo Bank. In banking activity San Francisco ranks second only to New York City.

Transportation. Los Angeles long ago surrendered unconditionally to the automobile; San Francisco has on occasion rebelled and found it possible to oppose "progress." A monument to this revolt is the elevated Embarcadero Freeway, which stops dead as if sliced off by a guillotine. Again, angry citizens in 1964 defeated the state highway commission when it threatened to run a freeway through the "panhandle" of Golden Gate Park.

Among the serious problems that remain is periodic smog, produced mainly by the cars in the area. (An air-pollution-control district was formed in 1955.) Another problem is that access to San Francisco from its commuter towns is largely by a network of freeways that are highly congested at rush hours. Travel from the East Bay cities of Oakland and Berkeley and from Marin County to the north is over two great but overburdened bridges. The 4½-mile-long San Francisco–Oakland Bay Bridge, completed in 1936, consists of two back-to-back suspension bridges, a connecting tunnel on Yerba Buena Island, five truss spans, and a cantilever span. The orange-red Golden Gate Bridge, leading north to Marin County, was completed in 1937. It is a pure suspension bridge with a 4,200-foot centre span.

The once-busy ferries

Until the ferries were doomed by the bridges, San Francisco was served by a great network of ferry routes whose splendid vessels were said to deliver more passengers to the Ferry Building at the foot of Market Street than arrived at any other transportation depot except Charing Cross in London. Only after the bridges began to choke with traffic did San Franciscans realize what they had lost, and the ferries have returned, on a smaller scale, between San Francisco and Marin County.

A much greater undertaking is the interurban rapid-transit system known as BART (Bay Area Rapid Transit), which began operating in 1972. Operating between San Francisco and the East Bay communities through an underwater tube more than 3½ miles long, BART was the first system of its sort, part subway and part elevated, to be built in half a century, but it does not resemble the older systems in all respects. Comfortable, computerized automatic trains running at speeds averaging 80 miles an hour are a major feature.

San Francisco is situated at the head of a peninsula, and it has always been a dead end for rail traffic, with the transcontinental trains (the first westbound train arrived over the tracks of the Central Pacific on September 6, 1869) discharging their passengers in Oakland, whence they were carried to San Francisco by ferry and by bus. As in the rest of the country, the railroad's importance has declined since World War II. For carrying goods, trucking has largely taken over.

San Francisco International Airport is located seven miles south of the city–county limits, on a filled site on the southwestern shore of the bay. It is owned and operated by the municipal government.

ADMINISTRATION AND SOCIAL CONDITIONS

Government. Unlike any other California city, San Francisco (incorporated 1850) has a consolidated city–county government. The 1932 freeholders' charter, under which the city–county still operates, provides the mayor with strong executive powers but delegates substantial authority to a chief administrative officer (appointed by the mayor) and a controller. The legislative authority is lodged with an elected board of supervisors. The other key officials are the superintendent of schools and the manager of utilities, both appointed.

Public utilities. Since 1934 San Francisco's principal source of water has been the Hetch Hetchy Reservoir, 167 miles away, in the High Sierra. Other sources are the Calaveras Reservoir in Alameda County and reservoirs in San Mateo County to the south. The Hetch Hetchy project required the damming of a valley in Yosemite National Park almost as splendid as the Yosemite Valley itself and the construction of tunnels, one 25 miles long, through the Coast Range. In 1902 the first high-voltage line transmitting hydroelectric power was completed between a powerhouse on the Mokelumne River and San Francisco, some 180 miles away. Since then, the Bay Area has developed a network of hydroelectric plants on the rivers of the interior, as well as a steam-powered plant on Monterey Bay.

Education. Although strictly speaking they cannot be counted as San Francisco institutions, the greatest universities in the Bay Area are the University of California, located across the bay in Berkeley, and Stanford University, down the peninsula in Palo Alto. Within San Francisco itself noted institutions of higher education are the University of San Francisco (Jesuit) and San Francisco State University (formerly San Francisco State College), which was founded as a normal school in 1899 but has achieved national prominence for its academic excellence. Other institutions include Golden Gate University, the City College of San Francisco (a two-year public college), and the San Francisco Art Institute.

CULTURAL LIFE

San Francisco is the home of two major musical organizations. The San Francisco Symphony performs in the Louise H. Davies Symphony Hall and gives pop concerts in the summer. The San Francisco Opera enjoys an early season so that its leading singers may fulfill their commitments at New York City's Metropolitan Opera. With the exception of the American Conservatory Theatre (ACT), a resident repertory group, the professional theatre is virtually nonexistent. The surviving downtown theatres are occupied largely by the touring casts of Broadway successes.

San Franciscans feel that their city is a haven for the artist. While this is true for those who value architecture and public sculpture, the painting collections do not rival those of Los Angeles and the East Coast. Memorable, however, are the jades and porcelains in the Asian Museum, the Rodin sculptures at the California Palace of the Legion of Honor, and the many treasures in small museums such as the Fire Department Pioneer Memorial Museum.

The Beat poets

One of the greatest writers to make San Francisco his home, although only for a short period in the mid-1860s, was Mark Twain. Other notables of the past century have included Frank Norris, Ambrose Bierce, Bret Harte, Jack London, and Robert Louis Stevenson, who lived in great poverty in a boarding house. During the mid-1950s San Francisco was one of the major gathering places for the Beat poets, including Lawrence Ferlinghetti (publisher and cofounder of the City Lights bookstore, which became a shrine of the movement), Gregory Corso, Allen Ginsberg, Jack Kerouac, and Gary Snyder. The current population of writers includes many reputable practitioners, a number of whom are affiliated with the creative writing program on the campus of San Francisco State University.

A most vital part of San Francisco culture from the beginning has been found in its restaurants, hotels, and drinking places. To this must be added the popular culture of the ethnic enclaves—Chinatown, the Italian community of North Beach, the black culture of the Fillmore district, Japan Town, the Russian colony along Clement Street, and the Spanish-speaking Mission district.

San Francisco's greatest contribution to the nation's life, however, has had nothing to do either with its ethnic cultures or with its officially anointed institutions—or, indeed, as has been suggested, with the emergence of the "topless" dancer in a North Beach nightclub in 1964. Instead, it was the unheralded appearance in 1967 of the "flower children," a generation of young people, long of hair and usually grubby in appearance, who declared themselves in headlong flight from the Great Society and preached the saving graces of peace and love. Unfortunately, by the 1970s the main street of their capital, the Haight-Ashbury district, had turned into an ugly and dangerous marketplace for drugs.

San Francisco also emerged, for a period in the late 1960s, as a capital of rock music, which achieved national prominence for the San Francisco sound of such groups as the Jefferson Airplane, the Grateful Dead, and the Quicksilver Messenger Service, as well as such individual performers as Janis Joplin. The city has become a centre for advocates of rights for homosexuals and has one of the largest homosexual communities in the country.

History

EXPLORATION AND EARLY SETTLEMENT

It is extraordinary that the site of San Francisco should have been explored first by land instead of from the sea, for San Francisco Bay is one of the most splendid natural harbours of the world, and great captains and explorers sailed unheeding past the entrance—Juan Rodríguez Cabrillo (1542–43), Sir Francis Drake (1579), and Sebastián Vizcaíno (1602). In 1769 a scouting party from an expedition led by the Spanish explorer Gaspar de Portolá looked down from a hilltop onto a broad body of water, the first white men known to have seen San Francisco Bay. It was not until August 5, 1775, that the first Spanish ship, the "San Carlos," commanded by Lieut. Juan Manuel de Ayala, turned eastward between the headlands, breasted the ebbing tide, and dropped anchor just inside the harbour mouth. (Though it is possible that Drake may have entered the bay, most evidence is against it.)

Settlers from Monterey, under Lieut. José Joaquin Moraga and the Rev. Francisco Palóu, established themselves at the tip of the San Francisco peninsula the next year. The military post (which remains in service as the Presidio of San Francisco) was founded September 17, 1776, and the Mission San Francisco de Asis (popularly called the Mission Dolores) was opened on October 9.

Almost half a century later, a village sprang up on the shore of Yerba Buena Cove, two miles east of the mission. The pioneer settler was an Englishman, Capt. William Anthony Richardson, who in 1835 cleared a plot of land and erected San Francisco's first dwelling—a tent made of four pieces of redwood and a ship's foresail. In the same year, the United States tried unsuccessfully to buy San Francisco Bay from the Mexican government, having heard reports from whalers and captains in the hide-and-tallow trade that the great harbour held bright commercial possibilities. Richard Henry Dana, whose ship entered the bay in 1835, wrote in *Two Years Before the Mast* (1840) that "If California ever becomes a prosperous country, this bay will be the centre of its prosperity."

The Americans had to wait only another 11 years. After fighting began along the Rio Grande, Capt. John B. Montgomery sailed the sloop of war "Portsmouth" into the bay on June 3, 1846, anchored in Yerba Buena Cove, and later went ashore with a party of sailors and marines to raise the U.S. flag in the plaza. On January 30, 1847, Yerba Buena was renamed San Francisco, which was regarded as a more propitious name for the town's future development.

The permanent white population of Yerba Buena in 1844 did not exceed 50 persons. By 1846 the settlement had a white population of 375 in addition to 83 blacks, American Indians, and Sandwich Islanders (Hawaiians). Two years later, just before the discovery of gold on the American River, the town had grown to about 200 shacks and adobes inhabited by about 800 whites.

THE GROWTH OF THE METROPOLIS

The discovery of gold

The city of the '49ers. With the discovery of gold, San Francisco picked up pace and direction. The modest village was at first almost deserted as its population scrambled inland to the Mother Lode, and then it exploded into one of the most extraordinary cities ever constructed. Some 40,000 gold hunters arrived by sea, another 30,000 plodded across the Great Basin, and still another 9,000 moved north from Mexico. By 1851 more than 800 ships rode at anchor in the cove, deserted by their crews.

Everybody except the miners got rich. Eggs sold for $1 apiece, and downtown real estate claimed prices that would almost hold their own against the appreciated values of the late 20th century. Until the bubble burst in the Panic of 1857, 50,000 San Franciscans became rich and went bankrupt, cheated and swindled one another, and took to the pistol and knife all too readily. As *The Sacramento Union* noted in 1856, there had been "some fourteen hundred murders in San Francisco in six years, and only three of the murderers hung, and one of these was a friendless Mexican." Two vigilance committees (1851 and 1856) responded to the challenge with crude and extralegal justice, hanging four men apiece as an example to the others.

In 1859 silver was discovered in the Nevada Territory. The exploitation in Nevada of the Comstock Lode, which eventually yielded $300,000,000, turned San Francisco from a frontier boomtown into a metropolis whose leading citizens were bankers, speculators, and lawyers who dressed their ladies in Paris gowns and ate and drank in splendid restaurants and great hotels.

San Francisco then was by all accounts an intoxicating city whose many charms moved the historian-moralist B.E. Lloyd to advise parents in 1876

> to look closer to their daughters, for they know not the many dangers to which they are exposed . . . and to mildly counsel their sons, for when upon the streets of this gay city they are wandering among many temptations.

The 1860s and 1870s marked the birth of the modern San Francisco, which has for more than 100 years laid claim with some justice to being the Athens, Paris, and New York City of the West but which never completely lost the mark of its wild beginning. As Rudyard Kipling was to observe after he visited the city in the 1890s, "San Francisco is a mad city, inhabited for the most part by perfectly insane people whose women are of remarkable beauty."

The 20th century. San Francisco's growth in the first half of the 20th century was shaped by two cataclysmic events. The first of these was the great earthquake and fire of 1906, which destroyed the central business district. Between 1906 and 1910, however, the city rebuilt atop its ashes. The other cataclysm was World War II, during which hundreds of thousands of servicemen passed through San Francisco on their way to the war in the Pacific or were stationed in and around the city; in addition, some 500,000 people came to work in the war-related industries in the area and settled temporarily in the cities around the bay. This was the beginning of the great postwar surge in the San Francisco area's permanent population.

BIBLIOGRAPHY. For accounts of the early history of San Francisco, the serious reader should not be discouraged by the voluminousness and variousness of the works of the historians HUBERT HOWE BANCROFT and THEODORE H. HITTELL, the latter of whom is particularly readable. FRANK SOULÉ, JOHN H. GIHON, and JAMES NISBET, *The Annals of San Francisco* (1855), is invaluable; while BENJAMIN E. LLOYD, *Lights and Shades of San Francisco* (1876), is both vivid and divertingly moralistic. HERBERT ASBURY, *The Barbary Coast* (1934, reprinted 1968), is a classic account of the underworld; while JULIA COOLEY ALTROCCHI, *The Spectacular San Franciscans* (1949), is a useful social history. JOHN HASKELL KEMBLE, *San Francisco Bay: A Pictorial Maritime History* (1957), contains a splendid collection of drawings and photographs. Of the many books about the earthquake and fire of 1906, WILLIAM BRONSON, *The Earth Shook, the Sky Burned* (1959), is a first-rate work of historical reconstruction. The growth of San Francisco is treated in GUNTHER BARTH, *Instant Cities: Urbanization and the Rise of San Francisco and Denver* (1975); JOHN B. MCGLOIN, *San Francisco: The Story of a City* (1978), is a popular history; and FREDERICK M. WIRT, *Power in the City: Decision Making in San Francisco* (1974, reissued 1978), is a study of local politics. The WRITERS' PROGRAM, CALIFORNIA, *San Francisco: The Bay and Its Cities,* new rev. ed. (1973), is a popular city guide; *San Francisco Bay* (1957), written by the naturalist-conservationist HAROLD GILLIAM, is both authoritative and evocative; and MELLIER G. SCOTT, *The San Francisco Bay Area: A Metropolis in Perspective* (1959), is a good systematic description of the metropolitan area. See also GLADYS HANSEN, *San Francisco Almanac: Everything You Wanted to Know About the City* (1980); and LAWRENCE FERLINGHETTI and NANCY J. PETERS, *Literary San Francisco: A Pictorial History from the Beginnings to the Present Day* (1980).

(K.La./G.C.Ha.)

São Paulo

The largest city of Brazil and dynamic capital of the state of the same name, São Paulo is the foremost industrial centre in Latin America. With one of the world's fastest growing metropolitan populations, it is also the largest city of the Southern Hemisphere and one of the largest conurbations in the world. Sometimes called the "locomotive that pulls the rest of Brazil," São Paulo has a vibrant and energetic urban core characterized by an ever-growing maze of modern steel, concrete, and glass skyscrapers. The city is located in the hills of the Serra do Mar, which forms part of the Great Escarpment that extends between the Brazilian Highlands and the Atlantic Ocean. It lies about 220 miles (354 kilometres) southwest of Rio de Janeiro and about 30 miles inland from the port of Santos. The city's name derives from its having been founded by Jesuit missionaries on January 25, 1554, the anniversary of the conversion of St. Paul.

The article is divided into the following sections:

Physical and human geography

THE LANDSCAPE

The city site. The Brazilian Highlands are composed of ancient crystalline rocks, which in the vicinity of São Paulo form a surface of gently rounded hills mantled with a reddish clay soil. Rivers such as the Tietê, on which São Paulo is located, rise near the edge of the Great Escarpment and flow generally westward to the Rio Paraná. In their course, they cross stratified sandstones and limestones overlaying the crystalline base, as well as sheets of volcanic rock that form the Paraná Plateau. Here, there are rapids and waterfalls, as well as dams and reservoirs that supply great quantities of hydroelectric power.

Located at an elevation of 2,690 feet (820 metres) above sea level, the city is surrounded by open country, valleys, and foothills. The higher terrain constitutes the preferred residential areas; the lower parts are on alluvial land along the banks of three rivers (the Tietê, the Pinheiros, and the Tamanduateí), and these are occupied by working-class residences, manufacturing establishments, and commercial enterprises. The area of the city is 576 square miles (1,493 square kilometres), but including suburban communities, such as Santo André, Diadema, São Bernardo do Campo, São Caetano do Sul, Osasco, Guarulhos, Mairiporã, Barueri, Santana do Parnaíba, Franco da Rocha, and Mogi das Cruzes, metropolitan São Paulo sprawls over an area of 3,070 square miles. Open spaces on the perimeter of the city, where there are clay soils mixed with sandy deposits, are used for intensive market gardening. A forest reserve of about 39 square miles is maintained in the nearby Serra da Cantareira, while the beaches of Santos and Guarujá provide pleasant resort areas.

Climate. The Tropic of Capricorn, at about 23°27′ S, passes through São Paulo and roughly marks the boundary between the tropical and temperate areas of South America. Because of its elevation, however, São Paulo enjoys a distinctly temperate climate. July is the coldest month, with an average temperature of 57.9° F (14.4° C) and occasional frost. Warmest is February, which averages 69.1° F (20.6° C). Rainfall is abundant, particularly during the summer season from October through March, averaging 56 inches (1,422 millimetres) per year. Humidity and air pollution combine to form a mist that often hangs over the city.

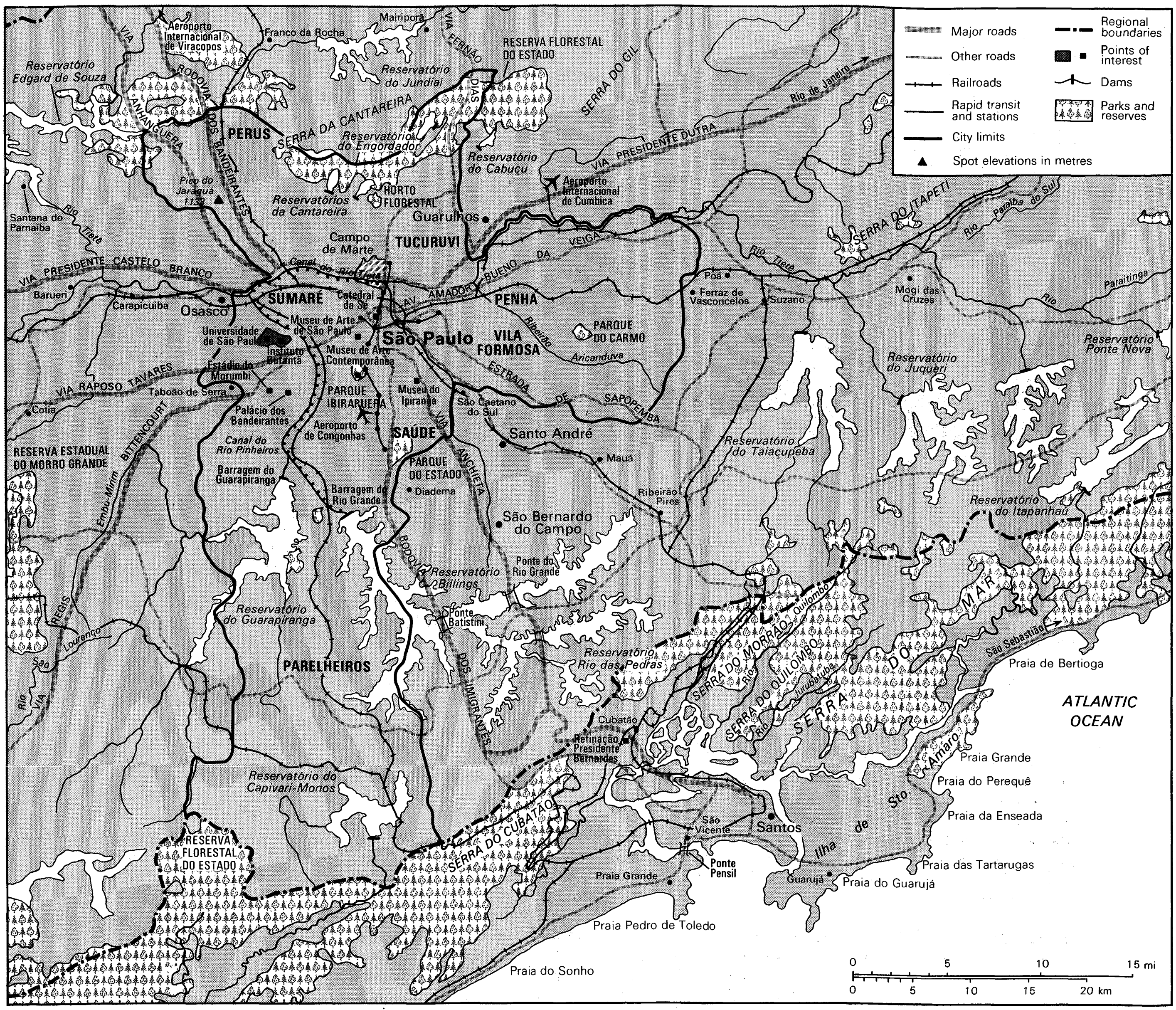
The city plan. The central business district of São Paulo focusses on the famous Triângulo, where the 42-story Edificio Itália, inaugurated in 1975, rises to a height of 558 feet. In 1947 there were only three skyscrapers in all of São Paulo, among which the newly constructed, 36-story Banco do Estado de São Paulo was the tallest. Now, the entire city is studded with modern buildings whose construction reflects a variety of architectural styles and materials.

Surrounding the central business district are extensive areas devoted to manufacturing, wholesale and retail trade, and repair and maintenance services. Most extensive are the residential areas characterized by low, red-roofed houses, interspersed with high-rise apartments or office complexes, singly or in clusters. Suburban shopping centres, like suburban neighbourhoods, have become commonplace.

São Paulo had no city plan until 1889, and no zoning law was passed until 1972. Until well into the 19th century, therefore, this state capital retained a colonial aspect, with narrow, unpaved streets, shabby buildings, and old churches and convents of Jesuit and Franciscan styles. Successive city administrations since then have attempted to stimulate more rational urban growth and to modernize the city's transportation system. Projects have included the straightening and rechannelling of the Tietê and Tamanduateí rivers, the widening and relocation of streets and avenues, the development of new parks and lakes, and the construction of superhighways and of a 40-mile subway network. Still, the city's street pattern shows little overall coherence, and the problems of intraurban traffic congestion and pollution have reached monumental proportions.

THE PEOPLE

The original settlers of São Paulo were relatively poor and largely from southern Portugal. They were, however, a restless people who sought actively to improve their status in life. Among them were the *bandeirantes* ("explorers") who formed expeditions that pushed far into the interior of South America in search of slaves and mineral wealth and, in the process, expanded the frontiers of what has become modern-day Brazil.



São Paulo metropolitan area.

With the great expansion of coffee cultivation in São Paulo state after 1880 came a massive immigration of Europeans. Italians and Portuguese were the most numerous, but there were also many Spaniards, Germans, and eastern Europeans. Other settlers came from Japan and the Middle East. Today, there are more Japanese in São Paulo than in any other community outside Japan, and Japanese farmers supply much of the city's market for fruit and vegetables. Even more numerous are internal migrants, primarily from the Northeast of Brazil. These include many blacks, who are the descendants of African slaves. Overall, the population is more than half European and about one-third black and mulatto, with the remainder made up of small groups of Asians and others. Roman Catholicism is the near-universal religion, and the archdiocese of São Paulo is the world's largest in number of adherents. Various other religions are represented by smaller numbers, and many Paulistas, as the inhabitants of the city are known, also attend the rites of local cults. Portuguese is the predominant language, although other languages are commonly spoken.

THE ECONOMY

Industry and commerce. Industrial development, beginning in the late 19th century, but especially since World War II, has transformed metropolitan São Paulo into the foremost industrial centre in Latin America. This city has often been referred to as the "Chicago of South America," but it actually is a leader of Brazilian commerce and industry to a much greater degree than is Chicago in the United States. The value of its industrial production is by far the country's largest. Its leading industries produce textiles, mechanical and electrical appliances, furniture, foodstuffs, and chemical and pharmaceutical products. There are also heavy metallurgical plants located at nearby Taubaté, oil refineries and chemical plants at Cubatão, and plants manufacturing motor vehicles and farm machinery in São Bernardo do Campo, Santo André, and other suburban communities. The several thousand manufacturing establishments in São Paulo provide employment for some 15 percent of the population. Despite its rapid growth in recent decades, however, the industrial sector has been able to absorb only a small fraction of the growing labour force. Hence, unemployment and underemployment are continuing problems.

Commerce, both wholesale and retail, is well developed and is spread over the city by zones according to specialty. Banks are concentrated in the central Triângulo of the city but maintain branches in almost every district. In addition to important Brazilian banks, there are banking institutions representing interests in North and South America, Europe, Asia, and Africa. No less important, in

Museu Paulista da Universidade de São Paulo (Paulista Museum) in São Paulo.
Colour Library International

terms of employment, are street vending, peddling, and neighbourhood stores.

Transportation. Major arteries of transportation radiate in all directions from São Paulo. Three major airports—Congonhas within the city, Cumbicas 15 miles east, and Viracopos 60 miles northwest—together with several smaller ones, provide São Paulo with both international and domestic service. The Viaçao Aérea São Paulo (VASP), with headquarters in the city, is Brazil's second largest airline and is owned by São Paulo state. Marine transport is provided through the port of Santos. São Paulo is also a hub of railroads, which include a transcontinental line from Santos to Antofagasta, Chile. Modern highways connect with inland cities, Santos, and Rio de Janeiro, and almost all the states of Brazil. Within the city, the first freeway was opened in 1969, and the subway system was inaugurated in 1976. Automobile traffic in the city and suburbs is heavy, and, despite street and highway improvements, congestion is a major and growing problem, which adds to the industrial city's serious conditions of air and noise pollution.

ADMINISTRATION AND SOCIAL CONDITIONS

Government and services. São Paulo is governed by a mayor and city council. It is also the seat of state government, headquartered in the Palácio dos Bandeirantes in the southwestern district (neighbourhood) of Morumbi. In addition to the state offices and departments that have headquarters in the capital, many branches of the federal government are represented there. More than 50 consulates represent countries in the Americas, Europe, Asia, and Africa.

The city has built a chain of reservoirs, tunnels, and canals to supply fresh water to an urban population that it is estimated may top 20,000,000 by the year 2000. The Cantareira water supply project, begun in 1969, increased water supplies greatly, but demand from the burgeoning population continuously outstrips supply. Pollution is an ever-present danger because of the slowdown of dammed streams that carry industrial waste. Electricity has been available in abundance since 1900. First, the waters of the Rio Tietê were dammed and dropped through penstocks from the Great Escarpment to generators below. Subsequently, dams on many rivers to the west, including Itaipú, a joint project with Paraguay, have been built to sustain the city's industries and residential districts.

Public and private health facilities are numerous, including hospitals for civil servants, maternity hospitals, and hospitals specializing in the treatment of cancer, tuberculosis, and other diseases.

Education. São Paulo has a well-developed system of primary and secondary education, both public and private, and a variety of vocational-technical schools. Among the institutions of higher education the largest and most esteemed in all of Brazil is the state-supported Universidade de São Paulo, established in 1934, which incorporated the historic Faculdade de Direito (College of Law) in the old São Francisco Square and preexisting polytechnical schools, as well as schools of pharmacy, dentistry, agriculture, and medicine. Economics, architecture, and engineering were added later. Affiliated institutions include a school of sociology and politics, founded in 1933, and the Instituto Butantã, a world-famous centre for research on snakes and the production of antitoxins and antivenins. The Pontifícia Universidade Católica de São Paulo was established in 1946 and the Universidade Mackenzie in 1952. Also well known is the Escola de Administração de Emprêsas of the Getúlio Vargas Foundation, established after World War II to train administrators.

CULTURAL LIFE

São Paulo early became a prominent cultural and intellectual centre, due largely to the opening in 1827 of the Faculdade de Direito, one of the first two in Brazil, where many of the nation's most eminent leaders were educated. The Instituto Histórico e Geográfico de São Paulo, founded in 1894, is one of the oldest cultural associations in the state. The city is also a leading centre for libraries, publishing houses, and theatres. The municipal library is housed in one of São Paulo's skyscrapers. In 1922 São Paulo's Modern Art Week, celebrated by a group of young writers, artists, and musicians in the Teatro Municipal, introduced modernism in the arts of Brazil. The Museu de Arte de São Paulo, founded in 1947, is perhaps the best in South America, and the Museu de Arte Contemporânea is also outstanding. São Paulo's symphony orchestra is similarly advanced in the field of music.

Publishing and broadcasting have long been established in São Paulo. Several of the nation's largest and most influential newspapers are published in the city, including *O Estado de São Paulo* and the *Diário Popular,* both more than 100 years old. Television was introduced in 1950, and the city is headquarters for some of the most important Latin-American radio stations.

The Paulistas are noted for their enthusiasm in sports. Soccer is the predominant sports attraction, as evidenced by the 150,000-seat Morumbi stadium and the Pacaembu stadium, which seats 70,000. Also popular are swimming, tennis, volleyball, basketball, and auto racing, for which São Paulo has one of the world's largest tracks, at Indianópolis on the city's south side. There are also countless parks, plazas, and playgrounds. The São Paulo zoo, with 3,500 animals, is the largest in Latin America.

History

São Paulo was the first highland settlement established in Brazil. It began as a small Indian settlement in 1554 under the direction of Portuguese Jesuit missionaries and occupied the lower terraces of the Rio Tietê in the midst of tall grasses and scattered scrub trees. The community grew slowly and had only 300 inhabitants by the end of the 16th century. Yet, São Paulo became a township in 1560 and had a town council that could enact and enforce laws. In 1683 it succeeded São Vicente as seat of the captaincy, or hereditary fief, and the inhabitants already had become known as Paulistanos or Paulistas.

Throughout the 17th century São Paulo was a base for expeditions (*bandeiras*) of armed pioneers (*bandeirantes*) who penetrated the remote hinterlands in search of Indian slaves, gold, silver, and diamonds. In the process they expanded the frontiers of what was to become modern-day Brazil. In 1711 São Paulo attained the status of a city, yet it remained an agrarian town that had yet to experience any significant prosperity.

Brazil's independence was declared on September 7, 1822, by Dom Pedro I, the Portuguese emperor in Brazil, at a site in São Paulo marked by the museum and monument of Ipiranga. Nevertheless, São Paulo continued to retain its colonial character until the latter part of the 19th century. Then, rather suddenly, coffee cultivation spread across the state of São Paulo, providing employment for many of the European and other immigrants who began arriving in great numbers. Included were Italians (who came to outnumber native Brazilians), Portuguese, Spaniards, Germans, eastern Europeans, Syrians, Lebanese, and Japanese. An era that was to transform São Paulo into a modern world-class city had begun.

Between 1885 and 1900 the São Paulo region was transformed from an isolated frontier to a new and independent region that focussed on the city of São Paulo and the port of Santos. With the spread of coffee, the major centre of urban activities was shifted to São Paulo, and the growth of the city was spectacular. New industries by 1905 included textile mills, shoe factories, and others using local raw materials. Cotton textile mills alone employed 39,000 workers.

In the late 19th century São Paulo had only one-tenth the population of Rio de Janeiro; by 1970 it had become the largest city in Brazil and one of the largest in the world. It included almost one-half of the population of São Paulo state, Brazil's most populous political subdivision, and accounted for about one-third of the nation's total industrial employment, and both proportions have continued to grow. Immigrants were pouring into the city at a rate of 300,000 per year, especially from the impoverished Northeast. São Paulo is a dynamic city, and continuous progress and official estimates indicate that by the year 2000 São Paulo will have surpassed Shanghai in population and will be the second largest urban agglomeration in the world, surpassed only by Mexico City.

BIBLIOGRAPHY. STEFAN GEYERHAHN (ed.), *São Paulo,* trans. from the Portuguese (1977), is a picture book and general description; BENEDITO LIMA DE TOLEDO, *São Paulo: Três Cidades em um Século* (1981), describes architecture, urban evolution, and history; N. LECOCQ MÜLLER, "Demographic Growth and Urban Expansion in the Metropolitan Area of São Paulo," *Revista Geográfica* (1983), is an informative study; JOSEPH L. LOVE, *São Paulo in the Brazilian Federation: 1889–1937* (1980), is a study of the state's geography, population, local politics, and economy; JOHN D. WIRTH and ROBERT L. JONES (eds.), *Manchester and São Paulo: Problems of Rapid Urban Growth* (1978), focusses on the economic conditions of the city; VINOD THOMAS, *Pollution Control in São Paulo: Costs, Benefits, and Effects on Industrial Location* (1981), studies one of the city's major problems; PAULO CURSINO DE MOURA, *São Paulo de Outrora: Evocações de Metrópole* (1980), is a history of the streets of the city; and ERNANI SILVA BRUNO, *História e Tradições de Cidade de São Paulo,* 3 vol., 2nd ed. (1954), is a general history of the city.

(A.Le./C.W.M.)

Scandinavian Literature

Scandinavian literature consists of those writings in the North Germanic group of the Germanic languages, the modern forms of which include Swedish, Norwegian, Icelandic, Danish, and Faeroese. The literary works written in these languages, though manifesting certain differences reflective of distinct national institutions, exhibit strong similarities stemming from deep-seated common linguistic and cultural ties. Some authorities include Finland among the Scandinavian countries on geographical and economic grounds, but the literature of the Finnish-speaking people, like their language, stands apart in a number of respects. (Finnish belongs to the Baltic-Finnic branch of the Finno-Ugric language family and is most closely related to Estonian, Livonian, Votic, and Karelian.) The present article does, however, devote some attention to various notable Finnish authors who wrote in Swedish. (Ed.)

The article is divided into the following sections:

The Middle Ages

The literature of Scandinavia and, in particular, of Iceland has reflected two extraordinary features of the social and cultural history of pagan Europe and of Iceland. The way in which names such as Siegfried, Brunhild, and Attila cropped up again and again in different European literatures has borne witness to the dissemination of legends and traditions common to the early Germanic tribes of Europe, starting from the great movements westward in the 4th, 5th, and 6th centuries. The literature of Iceland provides not only the most detailed descriptions available of the life-style of early Germanic peoples but constitutes the most complete account of their literature and literary traditions. Although the sagas and poems were first written down by Christian scribes, they present a picture of a pre-Christian European culture that reached its heights in the new settlements in Iceland.

A second feature directly concerns the peoples of Scandinavia. A remarkable characteristic of Scandinavian literature was the accuracy with which it described the geography of northern Europe, accuracy that was born of actual knowledge. From the late 8th century until well into the Middle Ages, the history of the Norsemen was one of unceasing movement toward western and central Europe. The Norsemen discovered Iceland, as early Icelandic historians had it, when their ships were blown off course about 860. The next century found the Vikings pushing west by way of Britain, Ireland, and France to Spain and then through the Mediterranean to North Africa and east to Arabia. Across land they reached the Black Sea, by sailing north they came to the White Sea, and finally, turning westward again, they reached America long before Columbus.

NORWEGIAN AND ICELANDIC LITERATURE

The roots of Norwegian literature reach back more than 1,000 years and become inextricably intertwined with early Icelandic literature. Although a large part of this early literature was composed either in Iceland or elsewhere in Scandinavia by Icelanders, the Norwegian element in it is considerable and indisputable, even though this cannot always be isolated and defined. In many instances, it is obvious that some of the literature derives from a time before the Scandinavian settlement of Iceland in the 9th century. In other cases, it appears that the composers of the works had resided for long periods in the mother country of Norway.

The classical period in Iceland. The best known Icelandic literature belongs to the classical period, roughly equivalent to the early and medieval periods in western European literature. Icelandic manuscripts yield much knowledge of European myth and legend, which is in part common to all Germanic peoples. Stories of the Norse gods and myths—of Odin, god of war; Balder the Beautiful; Thor, god of thunder; and Valhalla, hall of the slain—form the nucleus of early Icelandic literature.

Almost all extant early Scandinavian poetry was recorded in Icelandic manuscripts, although some was clearly composed before the Scandinavian peoples reached Iceland in the late 9th century. Much of the oldest poetry was recorded in the Codex Regius manuscript, which contained the *Sæmundar Edda* (*c.* 1270), commonly designated by scholars as the *Poetic Edda*, or *Elder Edda*. The poetry is sometimes called Eddaic and falls into two sections: heroic lays, which, broadly speaking, dealt with the world of men; and mythological lays, which dealt with the world of the gods.

Eddaic poetry

The heroic lays. The heroic lays followed the mythological in the Codex Regius and were probably the earlier of the two. Many of the legends on which they were based originated in Germany or even among the Goths. Oldest of all was perhaps the *Hamdismál* ("Lay of Hamdir"), which forcefully expressed the heroic ideals of Germanic tribal life. The story closely resembled one told by Jordanes, a Gothic historian of the mid-6th century, and his account suggested that his source was an even earlier poem about Hamdir. Another of the older lays in the *Poetic Edda* was the *Atlakvida* ("Lay of Atli"), which referred to events that took place in 5th-century western Germany, Atli (or Attila) being king of the Huns from 434 to 453. Nearly all heroic lays were associated with the story of Sigurd (or Siegfried), the valiant hero, and his ill-fated love for Brunhild, who, too, figured to varying extent in different lays. Many scholars hold that the lays concerned with the spiritual conflict of the heroines Brunhild and Gudrun, which tend to be romantic and sentimental, were later compositions than the austere heroic lays. The *Poetic Edda* contained only a small portion of the poetry known in Iceland in the Middle Ages and now lost. Fragments of ancient lays appeared in 13th- and 14th-century sagas such as the *Hlödskvida* ("Lay of Hlöd") in the *Heidreks saga*, as did mention of Danish and Swedish heroes in some fragments that must also have been known to the author of the Old English epic poem *Beowulf.*

The mythological lays. Mythological lays about the Norse gods made up the first half of the *Poetic Edda.* It is unlikely that any of these originated outside Norway, Iceland, and Norse colonies in the British Isles. The *Völuspá* ("Sibyl's Prophecy") was a striking poem on the history of the world of gods, men, and monsters, from the beginning until the "twilight of the gods." Many passages in the poem are obscure, but most modern scholars agree that it was composed in Iceland about the year 1000, when the people were turning from the old religion to the new. An interesting story of the gods was told in the *Skírnismál* ("Words of Skírnir"): sitting in "Gate Tower," throne of Odin, the god Freyr, lord of the world, gazes into the world of giants and falls in love with a giant maiden; to win her, he sends his messenger Skírnir, who first offers gifts and then threatens the maiden until she agrees to make a tryst with Freyr. Scholars have seen an ancient fertility myth in this story, and it was certainly one of the older mythological poems in the *Poetic Edda* and probably originated in Norway before Iceland was settled by Norwegians.

Didactic element in Eddaic poetry

The mythological poems so far mentioned were all narrative, but many of those in the *Poetic Edda* were didactic. The *Hávamál* ("Words of the High One"; *i.e.,* Odin) consisted of fragments of at least six poems. In the first section, the god speaks of relations between man and man and lays down rules of social conduct; in other sections he discourses on relations between men and women and tells how love of women may be lost or won; the last two sections are about runes and magic power. Most of the poems were probably composed in Norway in the 9th and 10th centuries. Another didactic poem, the *Vafþrúdnir* ("Words of Vafthrúdnir"), related a contest between Odin and a giant.

Some important mythological lays appeared in other manuscripts. *Baldrs draumar* ("Balder's Dreams") described how the god Balder dreamed that his life was threatened and how his father, Odin, rode to the grave of a prophetess to force her to reveal the fate in store for Balder.

The Eddaic verse forms. Three metres are commonly distinguished in Eddaic poetry: the epic measure, the speech measure, and the song measure. Most narrative poems were in the first measure, which consisted of short lines of two beats joined in pairs by alliteration. The number of weakly stressed syllables might vary, but the total number of syllables in the line was rarely fewer than four. In these respects it resembled the measure used by Anglo-Saxon and early Germanic poets. The speech measure used in the *Atlamál* ("Words of Atli") differed little from the epic measure, though its lines usually had a greater number of weakly stressed syllables. The song measure was the most irregular of the Eddaic verse forms. It was chiefly in didactic poems and generally consisted of strophes of six lines divided into half strophes of three lines.

Characteristics of skaldic poetry

Skaldic verse. Norwegians and Icelanders of the 9th to the 13th century also composed skaldic poetry (from the Icelandic word *skáld,* "poet"). It was not composed in the free variable metres of the *Poetic Edda* but was strictly syllabic: every syllable had to be counted and every line had to end in a given form. Like Eddaic lines, the skaldic lines were joined in pairs by alliteration, often using internal rhyme or consonance; but this poetry differed in syntax and choice of expression. Word order is freer than in Eddaic poetry, and a highly specialized poetic vocabulary employed periphrases, or kennings, of such complexity that the poetry resembles riddles. Little is known about skaldic verse forms, but they are thought to have been developed in Norway during the 9th century and could have been influenced by the forms and diction of Irish poets of the period. The earliest known poet was Bragi the Old, who probably wrote in Norway in the latter half of the 9th century. Harald I (died *c.* 940) of Norway was eulogized by several poets, among them Thörbjörn Hornklofi, whose poem the *Haraldskvaedi* ("Lay of Harald") was partly Eddaic and partly skaldic in style.

The distinction between Icelandic and Norwegian literature at this period is difficult to make. Skaldic verse seems to have originated in Norway and to have been developed by Icelandic poets who either, like Egill Skallagrimsson, spent much time in Norway or wrote in praise of Norwegian kings, as did Sigvatr, counsellor and court poet of Olaf II of Norway. Although its complexity means that skaldic poetry is now less appreciated than it deserves, the orally transmitted poems of the 10th and 11th centuries were valuable sources for Icelandic historians in the following centuries.

Influence of Christian missionaries on Iceland

Prose. Iceland's adoption of Christianity in 1000 opened the way for powerful influences from western Europe. Missionaries taught Icelanders the Latin alphabet, and they soon began to study in the great schools of Europe. One of the first was Ísleifr, who after being educated and ordained as a priest was consecrated bishop. His school at Skálholt in southern Iceland was for many centuries the chief bishopric and a main centre of learning. The earliest remembered historian was Saemundr the Wise, but Ari Thorgilsson is regarded as the father of historiography in the vernacular. A short history, *Íslendingabók* (or *Libellus Islandorum, c.* 1125; *The Book of the Icelanders*), and the more detailed *Landnámabók* ("Book of Settlements") are associated with his name. Extant works of the period are few or anonymous. Annals of contemporary events date from the 13th century and the oldest religious manuscripts, consisting of homilies and saints' lives, from *c.* 1150. Larger collections of religious literature appeared in late 12th- and early 13th-century manuscripts. As elsewhere, the most popular books were often lives of the Apostles and saints.

The sagas. The word saga is used in Icelandic for any kind of story or history, whether written or oral. In English it is used to refer to the biographies of a hero or group of heroes written in Iceland between the 12th and 15th centuries. These heroes were most often kings of Norway, early founders of Iceland, or legendary Germanic figures of the 4th to the 8th century. The oldest saga is the fragmentary *Oldest Ólafs saga helga* ("First Saga of St. Olaf"), written about 1180. In form it is a hagiographic narrative, laying emphasis on miracles worked through the agency of the saint. It was probably written in the monastery of Thingeyrar, which played an important part in cultural life in the late 12th and early 13th centuries.

Several sagas about King Olaf Tryggvason, at whose instigation the Icelanders adopted Christianity, were also written at Thingeyrar, where the work of the monks was fanciful rather than realistic. A more critical style of history was established in the south by Saemundr and Ari, and several notable works were written at Skálholt or nearby in the 13th century, such as the *Hungrvaka* ("The Appetizer"), a short history of the bishops of Skálholt from Ísleifr to Kloengr. In the late 12th century several short histories of Norwegian kings were brought from Norway to Iceland, where they influenced Icelandic historians. The *Ágriþ,* a summary of the histories, or sagas, of Norwegian kings, written in the vernacular in Norway, was particularly influential. The *Fagrskinna* ("Fine Skin") covered the same period in more detail, while the *Morkinskinna* ("Rotten Skin"), probably written earlier, covered the period from Magnús the Good (1035–47) to the late 12th century.

The role of Snorri Sturluson

Snorri Sturluson wrote many kinds of works and played an important role in political wrangles in his time. Among works ascribed to him was the *Snorra Edda* (*c.* 1225), a handbook of prosody and poetic diction commonly referred to as the *Prose Edda*, or *Younger Edda*. He twice visited Norway, and a large part of his work consisted of lives of its early kings: he combined his *Ólafs saga* with lives of other Norwegian kings to form the *Heimskringla* (*c.* 1220; "Orb of the World"). The value of these as historical sources has long been debated. Snorri was certainly well read in vernacular history and attempted to write faithful accounts of what he had read in earlier records. He did not aim to write scientific history; his work was creative and therefore portrayed his heroes imaginatively. The stirring *Egils saga* (on the skald Egill Skallagrimsson) is attributed to Snorri.

The Icelanders', or family, sagas. These sagas were about heroes who had supposedly lived in the 10th and 11th centuries. Their origins are unclear, and it is debatable whether they were faithful records of history. One theory is that they were composed in the 11th century and transmitted orally until written down in the 13th century; though researchers now reject this view, it is true that the sagas owed much to oral tales and the tradition of oral verse. Their historicity is difficult to verify, since their content and form were shaped both by the sources used and by the author's intentions.

It is also difficult to determine the date of many of the sagas. The obviously early works were somewhat crudely structured and expressed Norse ideals of loyalty and heroism. The *Gísla saga,* written before the middle of the 13th century, showed a development of artistic skill and contained rich descriptions of nature and verses of considerable beauty and tragic feeling. The *Laxdaela saga* ("Saga of the Men of Laxárdal"), written a few years later, was a delicately worked tragedy in which the author showed an unusual appreciation of visual beauty. One work that was clearly the author's creation was the *Hrafnkels saga Freysgoda* ("Saga of Hrafnkell, Freyr's Priest"): despite realistic detail, the saga contained little historical fact. As the century progressed, a taste for fantastic and romantic elements grew. The *Grettis saga* ("Saga of Grettir the Strong") included several motifs from folklore and portrayed a hero fighting against trolls and ghosts.

The saga of Njáll and Gunnar

The greatest of Icelanders' sagas, the *Njáls saga,* had in fact two heroes, Njáll and Gunnar. Gunnar is young and inexperienced and Njáll is a wise and prudent man endowed with prophetic gifts; he embodies traditional Norse ideals of loyalty and bravery, yet faces his death by burning with the resignation of a Christian martyr.

The heroic sagas. The fantastic element was further developed in the *fornaldar sögur,* literally "the sagas of antiquity," whose heroes were supposed to have lived in Scandinavia and Germany before Iceland was settled. The best known, the *Völsunga saga* (*c.* 1270), retold in prose stories from heroic lays of Sigurd, the Burgundians, and Jörmunrekr, and the *Hrólfs saga kraka* (*c.* 1280–1350) incorporated ancient traditions about Danish and Swedish heroes who also appeared in the Old English poems "Widsith" and *Beowulf.*

Many of the works on contemporary history were combined about 1300 in the *Sturlunga saga,* including the *Íslendinga saga* by Sturla Thórdarson.

Translations from Latin. A quantity of secular literature was translated from Latin between the 12th and 14th centuries. The "Prophecies of Merlin," already translated in verse by a Thingeyrar monk, were combined with a complete translation of Geoffrey of Monmouth's history and titled *Breta sögur* ("Stories of the Britons"). In one 14th-century manuscript this was preceded by the *Trójumanna saga* ("Story of the Trojans"), translated from Dares Phrygius. A Norwegian translation of the Bible was begun in the reign of Haakon V Magnusson (1299–1319).

Romances. Romances were also translated or adapted from continental romances. Interest in romance began in Norway and soon took root in Iceland. The earliest romance was probably the *Tristrams saga* (1226), derived from the Anglo-Norman poet Thomas. This was followed by the *Karlamagnús saga* ("Saga of Charlemagne"), a collection of prose renderings of French chansons de geste, including a Norse version of the *Chanson de Roland.* Romances in Icelandic were numerous, and their effect on the style of later writers is evident in such sagas as the *Laxdaela saga* and *Grettis saga.*

Post-classical literature in Iceland. In the period following the classical age, little was written that attracted attention outside Iceland. Realism and detached objectivity declined, and sentimentality and fantasy gained the upper hand. The decline in literary standards is sometimes attributed to Iceland's loss of independence in 1262 and the changes that followed. Interest in earlier manuscripts continued, and many 14th- and 15th-century manuscript collections of 13th-century material were made. The most beautiful of all Icelandic manuscripts, the *Flateyjarbók* (*c.* 1390), included versions of sagas of Olaf Tryggvason and St. Olaf, together with texts from other sagas or about heroes associated with Iceland.

Decline in literary standards

Prose. Prose literature of the 14th century included several sagas. Among them were the *Finnboga saga ramma* ("Saga of Finnbogi the Strong"), about a 10th-century hero, and another telling the love story of its hero Víglundr. Sagas about bishops, already a theme in the 13th century, became more numerous, as did lives of foreign saints. A large collection of exempla (moral tales) was also made, each short tale illustrating some moral precept.

Ásgrímsson's *Lilja*

Poetry. Much poetry was written up to the time of the Reformation, and many new forms were devised. The best poems were religious pieces, in honour of the Virgin, the Apostles, or other saints. The well-known *Lilja* (*c.* 1350; "The Lily") by Eysteinn Ásgrímsson, a monk from Thykkvabaer, gave an account of the fall of Satan, the creation, the first sin, and the birth, life, and Passion of Christ. The term *rímur*—rhymes—is used of the narrative poetry developed after 1500 that consisted of mainly four-line strophes: the lines had end rhyme. The metrical forms, although apparently derived from Latin hymns, inherited the alliterative system of earlier poetry. Ballads written in Icelandic never attained the popularity of Danish ballads in Denmark nor achieved the high standard of the Norwegian *Draumkvaede* ("Dream Ballad"). Most of those preserved dated from the 14th to the 16th century and were free translations of Danish and Norwegian originals.

SWEDISH LITERATURE

Swedish literature proper began in the late Middle Ages when, after a long period of linguistic change, Old Swedish emerged as a separate language. The foundations of a native literature were established in the 13th century. The oldest extant manuscript in Old Swedish was the *Västgötalagan* ("Law of West Gotland"), part of a legal code compiled in the 1220s. These legal documents often employed concrete images, alliteration, and a solemn prose rhythm suited to their proclamatory nature.

Early Swedish ballads

The poetry of chivalry was first represented in *Eufemiavisorna* ("The Songs of Euphemia"), written in doggerel between 1303 and 1312, which included a translation of Chrétien de Troyes' romance *Yvain.* Anonymous ballads probably dating from the 14th and 15th centuries also reflected a new interest in romance. These ballads, though mostly derived from foreign sources and combining the imported ideals of courtly love with native, pagan themes and historical events, formed the most accessible genre of what can be called Swedish medieval literature.

DANISH LITERATURE

Denmark's first literature appeared in the runic inscriptions scratched on stone or carved in metal, mainly epitaphs of warriors, kings, and priests that occasionally had short, unrhymed alliterative verses in the Viking spirit. Runic inscriptions were used in Denmark from about 250, but most of those preserved date from 800 to 1100. With the introduction of Christianity, Latin became the predominant literary language, and Denmark's first important contribution to world literature, Saxo Grammaticus' *Gesta Danorum* (written between 1185 and 1222; "The Deeds of the Danes"), which contained, for example, the Hamlet story, was written in Latin. The medieval ballads of Denmark are among the most important in Europe; 539 are known in more than 3,000 versions, but nearly all were written down after the end of the Middle Ages, the first printed edition appearing in 1591.

The 16th century

THE IMPACT OF THE REFORMATION ON SWEDISH LETTERS

Two dates mark the beginning of modern Swedish history: 1523—the breach with Denmark and Gustav I Vasa's accession; and 1527—the breach with Rome and the establishment of a national Lutheran Church. The political revolution that eventually brought Sweden to the position of a European power had no considerable effect on literature until a century later, but the Reformation wholly dominated Swedish letters in the 1500s.

The Swedish translation of the Bible

The most important literary event of this period was the translation of the Bible in 1541, which inaugurated modern Swedish and provided an inexhaustible source for poets of subsequent times. Closely involved in the Bible translation were the apostles of the Swedish Reformation, Olaus Petri and his brother Laurentius. Olaus Petri's vigorous approach was revealed in his published sermons and in a Swedish chronicle, the first historical Swedish work based on critical research. Olaus Petri may also have written the biblical *Tobie comedia* (published 1550), the first complete extant Swedish play.

As a consequence of the Reformation, two of Sweden's most distinguished scholars of the period, Johannes Magnus and his brother Olaus, were driven into exile. In his history of all the kings of the Goths and Swedes, Johannes provided Sweden with a number of valiant kings unknown to critical historians. Olaus wrote the first geographical and ethnographical account of Scandinavia, *Historia de gentibus septentrionalibus* (1555; "History of the Northern Peoples").

DEVELOPMENTS IN DANISH LITERATURE

In 1536 the Lutheran Reformation was carried through in Denmark, and the beginning of the 16th century was characterized by many pamphlets for or against the Roman Catholic Church. European humanism and the Renaissance made their influence felt also in Denmark, where Christiern Pedersen was the most prominent humanist who supported the Reformation. He edited *Gesta Danorum* by the 13th-century historian Saxo Grammaticus, translated the New Testament, adapted Martin Luther's pamphlets into Danish, and participated in a translation of the Bible (1550). Poul Helgesen was the most gifted opponent of the Lutheran Reformation and Hans Tausen its most talented spokesman. The *Visitation Book* by the Lutheran bishop Peder Palladius is an important literary document. The two most important historians were Anders Sørensen Vedel and Arild Huitfeldt.

Sixteenth-century Danish poetry was religious or polemical, with fine love poetry and hymns. The earliest plays date from the beginning of the century. The most important playwright of the period was Hieronymus Justesen Ranch, whose farce *Karrig Nidding* ("The Miserly Rascal") was his best play.

ICELANDIC LEARNING AND LITERATURE

The chief political figure and poet of the Reformation was Jón Arason, last Catholic bishop of Hólar, beheaded in 1550. By his life Jon showed that he was a Viking as well as a martyr, although most of his surviving poetry is religious.

The impact of the Reformation

The effect of the Reformation on Icelandic learning and literature was that Catholic poetry was discarded and attempts were made by the first Lutheran bishops to replace it with hymns poorly translated from Danish and German.

Lutheran teachers instructed the people in Protestant dogma, and several translations of sermons and books of instruction by German Lutherans were printed in Icelandic from as early as 1540. Gudbrandur Thorláksson was the most energetic of the Lutheran teachers. In translating the Bible he used earlier Icelandic versions of some books of the Old Testament and Oddur Gottskálksson's Icelandic translation of the New Testament. In his psalmbook he showed appreciation of Icelandic poetic tradition and adhered to Icelandic alliteration and form.

The 17th century

SWEDISH POETRY AND PROSE

Georg Stiernhielm and his followers

In the first half of the 17th century, Swedish literature remained limited in scope and quantity. A unique contribution, however, was made by Lars Wivallius, whose lyrics revealed a feeling for nature new to Swedish poetry. With its intervention in the Thirty Years' War, Sweden established itself as a European power, and this led to a development of national pride and culture, as revealed in literature of this epoch. The outstanding work was the allegorical epic *Hercules* (1658) by Georg Stiernhielm, which reflected many of the social and political problems of the time. Stiernhielm's followers included the two brothers Columbus, one of whom, Samuel, wrote *Odae sueticae* (1674; "Swedish Odes") and the prose *Mål-roo eller roomål,* a charming collection of anecdotes that illumine Stiernhielm's character. A rival to Stiernhielm was the unidentified "Skogekär Bärgbo," whose *Wenerid* (1680) was the first sonnet cycle in Swedish.

Stiernhielm aimed at an integration of Sweden's cultural heritage with the accepted ideals of continental classicism. His *Hercules* is full of old Swedish words that he was eager to revive. Columbus also demanded a more vigorous, flexible language as did "Skogekär Bärgbo" in *Thet swenska språkets klagemål* (1658; "The Lament of the Swedish Language"). National pride and religious feeling are combined in the works of the bishops Haquin Spegel and Jesper Swedberg, father of the Swedish mystic Emanuel Swedenborg. Spegel contributed to Swedberg's new hymnbook of 1695, which became the poetry book of the Swedish people and was of lasting influence. Even Lucidor was represented in it, giving intense expression to the contrasting moods of the period: in his love songs and, above all, in his drinking songs, he was as pagan and reckless as he was devout in his hymns and funeral poems.

At Uppsala, meanwhile, the scholar Petrus Lagerlöf attempted to impose purer classical standards on native literature, and Olof Verelius edited and translated Icelandic sagas. It was Olof Rudbeck, however, who became interested in Verelius' work and developed a theory that Sweden was the lost Atlantis and had been the cradle of Western civilization. He proposed this idea in *Atland eller Manheim* (1679–1702), which, translated into Latin as *Atlantica,* attained European fame.

Baroque and classicist tendencies ran parallel in late 17th-century Swedish literature. Gunno Eurelius (Gunno Dahlstierna) wrote an elaborate epic, *Kungaskald* ("Hymn to the King"), for King Charles XI's funeral in 1697. Simpler in style was Johan Runius, who expressed a Christian stoicism of the kind found among Swedes during the disastrous early decades of the 18th century. Jacob Frese was a gentler and more intimate poet; his lyrics and hymns contained some of the emotional pietism that became a feature of 18th-century thought.

THE LITERARY RENAISSANCE IN DENMARK

The revival of interest in Scandinavian antiquity

The literary Renaissance reached Denmark in the 1600s, giving rise to a strict adherence to classical patterns and blind belief in authority in political, religious, and literary matters. In religious literature Latin dogmatics and pamphlets reflecting the superstitions of the century were dominant. It was, however, a great era of scholarship. Ole Worm is famous for his book on the runic inscriptions, *Monumenta Danica* (1643). Thormod Torfaeus and Árni Magnússon introduced the study of Old Norse literature; Peder Hansen Resen edited and translated some of the poetry of the Old Norse *Edda;* and Erik Pontoppidan and Peder Syv introduced the linguistic study of Danish.

Danish poetry in the 17th century tended to follow the classics slavishly, and the favourite forms were the hexameter, the Alexandrine, and the sonnet. Simplicity is deliberately avoided; the style is precious; allegories, euphemisms, and metaphors abound. Anders Arrebo translated the Psalms and wrote *Hexaëmeron* (1661), a Danish version of the 16th-century French poet Guillaume du Bartas' *La Semaine.* The century was rich in occasional poetry; didactic and pastoral poems were also common. Anders Bording, an interesting exponent of Danish Baroque poetry, was also the founder of the first Danish newspaper, *Den danske Mercurius* (from 1666), in which the news appeared in rhymed Alexandrines. The only truly great poet was Thomas Kingo, a supreme master in almost every kind of poetry. His hymns reflect a violent, passionate character, wörldly and yet deeply religious.

Characteristic poetic forms

Of special interest among Danish works of the 17th century were the memoirs of Leonora Christina, daughter of Christian IV, a fascinating document about her 20 years' imprisonment in the Blue Tower of Copenhagen.

RENEWED LITERARY ACTIVITY IN NORWAY

Political union between Denmark and Norway started in 1380, and the Danish language eventually became the official and the literary medium. Copenhagen, with its university, established itself as the cultural capital of the two countries. Not until after the Reformation were there signs of renewed literary activity in Norway itself; *e.g.,* in the nostalgic apologia for Norway, *Om Norgis rige* ("Concerning the Kingdom of Norway"), written in 1567 by Absalon Pederssøn Beyer. The most original and most conspicuously Norwegian writer of this age was Petter Dass, whose *Nordlands trompet* (*The Trumpet of Nordland*) gives a lively picture in verse of the life of a clergyman; although probably completed before the turn of the century, this work was not printed until 1739.

ICELANDIC LETTERS

In Iceland the foremost poet of the 17th century was Hallgrímur Pétursson, a Lutheran pastor who struggled against poverty and ill health. His *Passíusálmar* (1666; "Hymns on the Passion") is among the most popular books in Iceland. Another interesting poet was Stefán Ólafsson, remembered for both religious and secular works, the latter notable for exuberantly humorous portrayals of contemporaries and satirical observations of manners and customs.

As in other countries, interest in antiquity was stirred during the 17th century, and modern learning may be said to date from that period. Arngrímur Jónsson called the attention of Danish and Swedish scholars to Icelandic traditions and literature in a series of works in Latin, some containing abstracts of sagas now lost. Later in the century Árni Magnússon systematically collected the early Icelandic manuscripts.

The 18th century

SWEDISH CLASSICISM AND ENLIGHTENMENT

After the death of Charles XII (1718) and the collapse of his empire, a utilitarian attitude to life and letters gradually developed in Sweden. Olof von Dalin was the outstanding popularizer of the new ideas of the French and English Enlightenment. Educated at Lund, he later went to Stockholm and began to publish, anonymously, *Then swänska Argus* (1732–34; "The Swedish Argus"), a weekly periodical modelled on that of the Englishman Joseph Addison. One of the first serious journalistic ventures in Sweden, it marked the beginning of a new era, in which orthodoxy gave way to Skepticism and Enlightenment, Baroque to Classicism, and German influence to English and French; at this time the middle class began gradually to take over the function of chief upholder of literature. In *Argus* Dalin ridiculed the foibles of the capital and in *Sagan om hästen* (1740; "The Story of the Horse") he showed himself a master of allegorical satire. He also produced some pseudo-Classicist plays that, like many dramatic ventures of the early and mid-18th century, are academic and lifeless. The one notable exception is *Den Svenska sprätthöken* (1740; "The Swedish Fop"), a comedy by Count Carl Gyllenborg.

English and French influence

With the second phase of the Enlightenment, marked by the influence of Rousseau, are associated Hedvig Charlotta Nordenflycht, the epicurean Gustav Philip Creutz, and his stoic friend Gustaf Fredrik Gyllenborg. In *Den Sörgande turturdufwan* (1743; "The Sorrowing Turtledove"), Fru Nordenflycht laments the death of her husband in highly personal lyrics. Creutz was a more sophisticated personality. He wrote little, but his few writings, of which the pastoral *Atis och Camilla* (1762) is the most important, reveal a mastery of form and versification.

Prose—particularly the novel—developed more slowly. The first genuine novel, *Adalrik och Giöthildas äfventyr* (1742–44; "The Adventures of Adalrik and Giöthilden"), by Jacob Mörk and Anders Törngren, shows the influence of the Icelandic sagas. Only two 18th-century Swedish writers were of European reputation, and both were scientists: Carl von Linné (Linnaeus) and Emanuel Swedenborg.

The first Swedish novel

The Gustavian period takes its name from King Gustav III (1746–92), a brilliant man and a patron of art and letters. He was especially interested in drama and opera and, thanks to his patronage, a proper theatrical tradition developed. Gustav himself sketched out some works, the best of which was a historic opera, *Gustaf Vasa,* which was finished in collaboration between Johan Henrik Kellgren and the composer J.G. Naumann. Kellgren, an academic poet and arbiter of taste, ruled that Swedish literature should be modelled on Classicist French patterns, but, beginning as a Rationalist and satirist after the fashion of Voltaire, he reluctantly accepted pre-Romantic ideas later. In *Stockholmsposten,* the main organ of literary opinion in the capital, Kellgren used his polemical wit against Thomas Thorild, a truculent champion of individual genius. After Kellgren's death the controversy was carried on by Carl Gustaf af Leopold, who imposed pseudo-Classical standards on the academy and applied them in his own rhetorical odes and tragedies. Johan Gabriel Oxenstierna did his most original work while a diplomat in Vienna; his *Skördarne* (1796; "Harvests") reveals pre-Romantic feeling for the beauty of nature. Bengt Lidner was the chief exponent of pre-Romanticism in poetry. His most successful work was the ode *Grefvinnan Spastaras död* (1783; "The Death of Countess Spastara").

Carl Michael Bellman stands apart from the conflicting ideals of the time. A poet and musician, he combined stylized realism with humour and the most uniquely delicate sense of language and rhythm. He was the greatest Swedish lyricist of the 18th century.

The dissertation *Om upplysning* (1793; "On Enlightenment") by Nils von Rosenstein, the first secretary of the Swedish Academy, expressed the ideals of the Gustavian epoch. Memoirs by G.J. Adlerbeth, G.J. Ehrensvärd, and others evoke the witty but artificial atmosphere of Gustav III's court. Gustav IV, who followed, did not encourage literature; nevertheless, Anna Maria Lenngren wrote some of her best verse satires between 1795 and 1800, many aimed at aristocratic foibles. The sentimental idylls of Frans Mikael Franzén are full of pre-Romantic idealism from German and English sources.

LITERARY ACTIVITY IN DENMARK, NORWAY, AND ICELAND

Denmark. The 18th century was a fertile period in Danish literature. The great name in the first half of the century was that of Ludvig Holberg, a Norwegian by birth. His most important contributions, written for the Danish theatre, which opened in 1722, were 32 comedies of character and manner, including some moral allegories in his old age. His aim was to create a modern Danish literature on European lines and to make people laugh at their own follies. Influenced by English and French thinking, he was a Rationalist and a moderate. He also wrote satire, a mock-heroic poem, and *Nicolai Klimii iter Subterraneum* (Latin, 1741; *Journey of Niels Klim to the World Underground*). His *Moralske tanker* (1744; "Moral Thoughts") and his *Epistler* (1748–54; "Letters") are the finest examples of a Danish political essay form.

The comedies of Ludvig Holberg

Among Holberg's contemporaries the finest lyrical poets are H.A. Brorson, a mystic whose pietist hymns often have a background of personal sorrow or agony; and

Ambrosius Stub, whose poems are mainly religious and moralizing verses, witty epigrams, or drinking songs. A satirist, Christian Falster, was a conservative counterpart to Holberg; Friedrich Eilschov and Jens Schelderup Sneedorff, the latter of whom edited *Den patriotiske Tilskuer* ("The Patriotic Spectator"), a Danish *Spectator,* were both Rationalist disciples of Holberg.

Revival of Danish literature

A significant revival of Danish literature took place toward the end of the century. In 1772 the Norwegian Johan Herman Wessel, one of the greatest humorists to use the Danish language, wrote *Kaerlighed uden strømper* ("Love Without Stockings"), a parody of the Danish imitations of Italian operas and French tragedies that had superseded Holberg's comedies.

The works of Johannes Ewald

At the same time a revival of emotional poetry was taking place, influenced by German and English literature. Johannes Ewald, perhaps Denmark's greatest lyrical poet, was the first to discover the poetic wealth of Scandinavian antiquity in the *Gesta Danorum* of Saxo Grammaticus and in the myths, sagas, and ballads. He wrote verse dramas and deeply personal and descriptive poems. *Fiskerne* (1779; "The Fishermen") was the first serious Danish drama in which ordinary people were treated heroically. His memoirs, *Levnet og meninger* (posthumously published in 1804; "Life and Opinions"), were influenced by Laurence Sterne and Jean-Jacques Rousseau. Jens Baggesen at first imitated the satires of Holberg and Wessel but gradually developed as a poet of distinction. In *Labyrinten* (1792–93; "The Labyrinth"), he described his travels in Europe in the manner of Sterne.

Norway. Several of Denmark's leading writers of the 18th century were of Norwegian birth, preeminently Ludvig Holberg and the members of Det Norske Selskab (the Norwegian Society). Established in Copenhagen in 1772 by a group of resident Norwegians, it looked to French rather than to German and English literature for models. Within Norway itself there was little overt literary activity, though the establishment in 1760 of a Royal Norwegian Society of Learning in Trondheim was evidence that Norway was beginning to assert its cultural aspirations.

Iceland. *Húss-Postilla* (1718–20; "Sermons for the Home"), by Jón Vídalín, bishop of Skálholt, is the best example of early 18th-century prose. Among important later writers, Eggert Ólafsson carried out a comprehensive geographical field survey (published in Danish, 1772; partial Eng. trans.) of Iceland's country and its people. In his poetry he expressed 18th-century Rationalism combined with Romantic patriotism. Jón Thorláksson, poet and scholar, translated John Milton's *Paradise Lost* and Alexander Pope's *Essay on Man.*

Finnur Jónsson, bishop of Skálholt, wrote *Historia Ecclesiastica Islandiae* (1772–78), which covers the history of Christianity in Iceland. Jón Espólín published *Íslands árbaekur* (1822–55; "Annals of Iceland"), a history of Iceland from 1262.

The 19th century

SWEDISH LITERATURE

Romanticism. Political changes in Sweden up to 1804 meant that ardent nationalism emerged as a characteristic of Swedish Romanticism. The idealism at the core of this movement was laid by the Kantian teaching of Benjamin Höijer and the impact of Friedrich Schiller, Johann Wolfgang von Goethe, and the German Romantics on Swedish literature. Student societies and their periodicals, such as *Polyfem* (1809–12) and *Phosphorus* (1810–13), led the attack on the traditional school. Most gifted of the Forforister, or Phosphorists, Per Daniel Atterbom, wrote a verse "Prolog" (1810) to *Phosphorus* revealing both talent and commitment to Romanticism.

The return to the literary past

Meanwhile, another society, Götiska Förbundet (Gothic Society), advocated, from its start in 1811, that study of the "Gothic" past could morally improve society. One of its members, Esaias Tegnér, wrote a most popular poem, *Frithiofs saga* (1825), based on an Old Norse theme. Tegnér valued old Northern mythology for the patterns he discerned in it—patterns also found in Greek mythology and Romantic metaphysics, in which religion, philosophy, and poetry appeared to be one and the same. Nevertheless, Tegnér's ideals of clarity of thought and formal perfection led him sometimes to side with traditionalists in their struggle against obscurities and formal innovations.

Several leading Romantics were learned men whose poetry strove to embody a philosophical system or an interpretation of history. The most ambitious attempt of this kind was P.D.A. Atterbom's *Lycksalighetens ö* (1824–27; "The Isle of Bliss"), an allegory dealing with adventures of a legendary king, Astolf, and a history of poetry as an illustration of man's alienation from the divine. The greatest poet was perhaps Erik Johan Stagnelius, who held aloof from schools and coteries. The recurrent theme in his *Liljor i Saron* (1821; "Lilies of Sharon") was the lament of the human soul, imprisoned in a world of darkness and sin.

In prose the most complex personality among the later Romantics was a novelist, Carl Jonas Love Almqvist, who combined an extravagant imagination with realism. A master of prose style, he was at his best in the long short story, in which he foreshadowed Strindberg's method of raising problems for debate. The novel was established by Fredrika Bremer, author of *Grannarna* (1837; "The Neighbours"), whose "sketches from ordinary life" appeared from 1828. Sophie von Knorring wrote chiefly about aristocratic families, and Emilie Flygare-Carlén produced stories dealing with west-coast life, including *Rosen på Tistelön* (1842; *The Rose of Tistelön*).

Scandinavism

Emergence of Realism and Poetic Realism. Realism made only slow headway in spite of the example of the Finno-Swedish poet Johan Ludvig Runeberg (see below). Literature of the 1840s and 1850s was mainly an aftermath of Romanticism. A movement known as Scandinavism produced a good deal of verse: Carl Vilhelm August Strandberg (pseudonym "Talis Qualis"), fieriest poet of this type, later made excellent translations from Byron. Popular reading was provided by August Blanche in *Bilder ur verkligheten* (1863–65; "Pictures of Real Life"), short stories depicting Stockholm life with humour and vivacity, while Frans Hedberg wrote pompous historical plays.

Poetic Realism became an official program of the "pseudonym poets" of the 1860s, including Carl David of Wirsén, Edvard Bäckström, Pontus Wikner, and Carl Snoilsky. Only Snoilsky had the temperament and poetic gift needed to carry out the program. Wirsén, on the other hand, as secretary of the Swedish Academy, launched formidable opposition against innovators; and Viktor Rydberg fell between Idealism and Naturalism. His important early work consisted of an ideological novel, *Den siste athenaren* (1859; *The Last Athenian*), and a treatise, *Bibelns lära om Kristus* (1862; "The Teaching of the Bible About Christ"), which prepared the way for scientific Rationalism.

August Strindberg

Sources of modern Swedish literature. Four influences combined to free Swedish literature from petrifying conventions: the English writings of Charles Darwin, Herbert Spencer, and John Stuart Mill; the French Naturalism of Émile Zola; the drama of the Norwegians Henrik Ibsen and Bjørnstjerne Bjørnson; and the criticism of the Dane Georg Brandes. The modern literature growing out of this was first and best represented in the work of August Strindberg, Sweden's greatest writer. Modern drama has dated from his play *Mäster Olof* (1872), and the modern novel from *Röda rummet* (1879; *The Red Room*). Strindberg overshadowed all the writers of the 1880s, including Gustaf af Geijerstam, author of *Erik Grane* (1885), Anne Charlotte Edgren-Leffler, and the gifted Victoria Benedictsson; the latter two wrote about the adverse position of women in society. Benedictsson's stories, such as *Från Skåne* (1884; "From Skåne"), revealed the regional character of the new prose literature. Regional poetry was written by Albert Bååth and Ola Hansson, both of Skåne.

In 1888 Verner von Heidenstam began the reaction against Utilitarianism and Naturalism with a volume of verse, *Vallfart och vandringsår* ("Pilgrimage and Wander Years"). His later poetry and historical tales won him the Nobel Prize for Literature in 1916. Oscar Levertin, stimulated by Heidenstam's example, wrote poetry full of colour and lore of the past and as a critic was influential in molding contemporary taste. Gustaf Fröding was also influenced by Heidenstam, and his verse constantly mingles the melancholy and gay. Regionalism entered Neoromantic poetry with Fröding, who was from Värmland, and with the work of Erik Axel Karlfeldt the province of Dalarna came into its own. Karlfeldt's mature poetry won him the Nobel Prize in 1931.

Selma Lagerlöf

Meanwhile, Selma Lagerlöf, the first Swede to win a Nobel Prize for Literature (1909), had developed the prose tale; her long series of novels and short stories, beginning with *Gösta Berlings saga* (1891), reached an international public through translation. Per Hallström was a more skillful writer of short stories than of novels. Romantic, too, in his love for the skerries (rocky isles) was Albert Engström, a great humorist.

Finno-Swedish literature. A significant literature in the Swedish language developed in Finland during the 19th century. Its emergence can be traced to the works of Johan Ludvig Runeberg. His epic poems, *Elgskyttarne* (1832; "The Moose Hunters") and *Hanna* (1836), won him a major place in Swedish letters. Notable, too, were the writings of Zacharias Topelius, which contributed much to the development of the Finnish historical novel. Topelius is perhaps best remembered for *Fältskärns berättelser* (1853–67; *The King's Ring and the Surgeon's Stories*), a romanticized account of 17th- and 18th-century Finno-Swedish history.

NORWEGIAN LITERATURE

The Age of Wergeland. After 1814 a new, exciting, and difficult age began for Norway: an opportunity seemed to be offered to develop an independent Norwegian culture and way of life, but there were deep differences of opinion as to how this could best be achieved. A poet and critic, Johan Sebastien Welhaven was chief representative of those who insisted that the existing Danish element in the culture should not be neglected. Henrik Wergeland was a spokesman for those whose nationalistic pride led them, on the other hand, to demand a complete break with Denmark. Welhaven stood for a coolly intellectual approach, for restraint and control, and for conscious artistry, as his own sonnet cycle *Norges daemring* (1834; "The Dawn of Norway") exemplifies. Wergeland was more passionate and revolutionary, and his enormous epic, *Skabelsen, mennesket og messias* (1830; "Creation, Humanity and Messiah"), typified the spirit he admired.

The ideological conflict between Wergeland and Welhaven

Wergeland dominated the age as poet, orator, and social reformer, and the clash between him and Welhaven and between the two factions associated with them—the "patriots" and the "intelligentsia"—began an ideological conflict that has continued to persist in modified forms.

National Romanticism. The literature of the mid-19th century, known as Norway's "national Romanticism," continued to reflect the country's larger aspirations. The compilation and publication, between 1841 and 1844, of *Norske folkeeventyr* ("Norwegian Folk Tales") by Peter Christen Asbjørnsen and Jørgen Engebretsen Moe, and the 1853 collection by Magnus Brostrup Landstad, *Norske folkeviser* ("Norwegian Folk Ballads"), indicated a lively interest in the past, as did Peter Andreas Munch's eight-volume history of the Norwegian people (1857–63). Ivar Aasen was the creative spirit behind the Landsmål movement to establish a literary language based on rural dialects linked with Old Norse. Many publications of these years, including earlier works of Ibsen and Bjørnson, turned consciously to Norway's heroic past and its peasants. To these years belonged also the lyric poetry of Aasmund Olafsson Vinje, founder of the periodical *Dølen,* who adopted Nynorsk (New Norwegian) as his literary language.

The problem dramas of Ibsen and Bjørnson

In 1855 Camilla Collett, Wergeland's sister, published *Amtmandens døttre* ("The Governor's Daughters"), which, by considering the place of women in society, marked a beginning of a trend that, encouraged by the immensely influential Danish critic, Georg Brandes, culminated in the 1870s and the 1880s in the realistic "problem" literature of Ibsen, Bjørnson, and their contemporaries. *Samfundets støtter* (1877; *Pillars of Society*) was the first of a succession of problem dramas by Ibsen to win him worldwide fame. By then he had already written two verse dramas, *Brand* (1866) and *Peer Gynt* (1867), and his long "double drama" *Kejser og Galilaeer* (1873; *The Emperor and the Galilean,* 1876). The first substantial drama of this type by Bjørnson was *En fallit* (1875; *The Bankrupt*). Although never the world figure that Ibsen became, Bjørnson was a leading personality of his age in Norway, as novelist, dramatist, and lyric poet and in public affairs.

The novelists Jonas Lie and Alexander Kielland, together with Ibsen and Bjørnson, were the major figures of modern Norwegian literature and were responsible for a remarkably large body of important work between 1870 and 1884, as the following titles illustrate: Ibsen's works *Et dukkehjem* (*A Doll's House*), *Gengangere* (*Ghosts*), *En folkefiende* (*An Enemy of the People*), and *Vildanden* (*The Wild Duck*); Bjørnson's dramas *Det ny system* (*The New System*), *En handske* (*A Gauntlet*), and *Over aevne* (*Beyond Human Power I*) and his novel *Det flager i byen og på havnen* (*The Heritage of the Kurts*); Lie's novels *Gaa Paa!* ("Go Ahead!"), *Livsslaven* (*One of Life's Slaves*), and *Familjen paa Gilje* (*The Family at Gilje*); and Kielland's *Skipper Worse* (1882), *Gift* (1883; "Poison"), and *Fortuna* (1884; *Professor Lovdahl*). The foremost stylist of his age, Kielland was an elegant, witty novelist with a strong social conscience and an active reforming zeal stemming from an admiration for John Stuart Mill.

The literature of the 1870s emphasized individual development and expression in keeping with the optimistic attitude of the times to social change and improvement. In the following decade, growing skepticism and disillusionment made writers more bitter in their attacks on "established" social institutions. The publication of *Fra Kristiania-Bohêmen* ("From the Christiania Bohemia") in 1885 by Hans Henrik Jaeger created, by its seeming advocacy of sexual license, a public scandal. The most extreme exponent of Naturalism was Amalie Skram, especially in a four-volume novel, *Hellemyrsfolket* (1887–98; "The People of Hellemyr"). Arne Garborg, poet, novelist, dramatist, and critic, was a much superior writer whose work reflected successive movements of Romanticism, Realism, Naturalism, and Neoromanticism. His wider reputation was first established with a novel, *Bondestudentar* (1883; "Peasant Students"), but perhaps his greatest achievement was the poem cycle *Haugtussa* (1895).

DANISH LITERATURE

The works of Adam Oehlenschläger

The Romantic period. The Romantic movement came to Denmark from Germany, inspired partly by the German Jena Romantics and partly by the Neoclassicism of Goethe and Schiller. Friedrich Schelling's philosophy was interpreted in Denmark by the Norwegian Henrik Steffens, but the leading Danish Romantics gave it a form very different from the original. The leader of the Romantic movement in Denmark was Adam Oehlenschläger, whose unparalleled versatility in poetry, drama, and prose showed the influence of certain works of Goethe and Schiller and of German Romanticism. His plays *Sanct Hansaften-Spil* (1802; "Play for Midsummer Eve") and *Aladdin; Hakon Jarl* (1857), one of his many Northern tragedies; and a cycle of dramatic poems, *Helge* (1814), were outstanding. The popular and historical songs and hymns of the poet N.F.S. Grundtvig, as well as his personal poetry, have given him a lasting place in Danish literature. Sharing the Romantic enthusiasm for antiquities of Scandinavia, he translated the 13th-century historians Saxo Grammaticus and Snorri Sturluson (see above) and translated *Beowulf* even before it had appeared in English. Bernhard Severin Ingemann wrote historical novels and a poetic cycle, *Holger Danske* (1837; "Holger the Dane"), around the themes of chivalry and nationalism as well as his unsophisticated *Morgen og aftensange* (1837–39; "Morning and Evening Songs"). Johannes Carsten Hauch wrote tragic and philosophic dramas, novels, and contemplative poetry.

Romantic Realism. New elements of reason and realism appeared after the first quarter of the century in the works of Poul Møller, who wrote the first Danish novel on contemporary life, *En dansk students eventyr* (1824; "The Adventures of a Danish Student"), and dramatic poems and fables, sometimes showing personal disillusionment, and of Steen Steensen Blicher, who, in *Traekfuglene* (1838; "The Birds of Passage"), interpreted human nature with sad resignation. Some of his best poems were in the Jutland dialect. His many *noveller,* or short stories, beginning in 1824 with the masterly "En landsbydegns dagbog" ("The Journal of a Parish Clerk"), struck notes varying from sorrow and resignation to humour and irony.

Minor writers of the same period were Thomasine Gyllembourg-Ehrensvärd, whose novel *En hverdagshistorie* (1828; "A Story of Everyday Life") was much admired; Andreas de Saint-Aubin, who wrote novels under the nom de plume of "Carl Bernhard"; and Carl Bagger, whose novel *Min broders levned* (1835; "My Brother's Life") shocked the literary world by its bold realism.

Basic characteristics

Poetic Realism. About 1830, early Romanticism gave way to a less naive poetic realism, more contemplative and more concerned with form than with content. Johan Ludvig Heiberg, who led this movement, attempted to revivify Danish drama by importing French vaudeville, and in his serious romantic plays *Elverhøj* (1828; "The Elfinhill") and *Syvsoverdag* (1840; "Day of the Seven Sleepers") he juxtaposed poetic and pedestrian reality. His finest achievement was a verse comedy, *En sjael efter døden* (1841; "A Soul After Death"). He was the leading literary critic of his time, profoundly influenced by the philosophy of G.W.F. Hegel. Henrik Hertz also regarded the perfection of poetic form as more important than its content, as was clearly expressed in *Gjenganger-breve* (1830; "Letters of a Ghost"). He also wrote comedies and serious Romantic plays, including *Kong Renés datter* (1845; *King René's Daughter*).

An upsurge of interest in lyrical poetry occurred in the 1830s and 1840s, led by poets concerned with the aesthetic treatment of love and nature. Christian Winther, best known for a long verse novel, *Hjortens flugt* (1885; "The Flight of the Stag"), sang the praises of his native island, Zealand, and of woman. Ludvig Bødtcher wrote delicate and sensitive poetry, some of which was inspired by the Italian scene. Emil Aarestrup treated erotic themes. Frederik Paludan-Müller became an uncompromising moralist; *Adam Homo* (1841–48), a poetic epic, was a bitter contemporary satire. Hans Christian Andersen was most important for his fairy tales, the majority of whose plots were his own invention, though he also wrote novels, plays, travel books, and poems.

Kierkegaard

Søren Kierkegaard holds a position entirely isolated in Danish literature. His highly personal religious philosophy was expressed in such works as *Enten Eller* (1843; *Either/Or: A Fragment of Life*) and *Stadier paa livets vei* (1845; *Stages on Life's Way*). He spent his last years in a violent and passionate attack on "official Christianity."

Meïr Aron Goldschmidt edited a rebellious, anti-royalist weekly, *Corsaren* ("The Corsair"), while many of his novels and short stories were concerned with Jewish life in the Danish community. The 1850s and 1860s produced few new Danish writers of importance: the most original was Hans Egede Schack, whose novel *Phantasterne* (1857; "The Daydreamers") revealed great psychological gifts.

FAEROESE LITERATURE

Emergence of a written literature

Modern Faeroese literature emerged during the second half of the 19th century. Until this time, the literary tradition of the Faeroese was almost exclusively oral. It consisted principally of ballads, epic and fantastic in style and centred on legends such as that of Sigurd. A new, national written literature in Faeroese became possible only after the orthography was normalized by means of rules introduced in 1846 by the linguist and folklorist Venceslaus Ulricus Hammershaimb. Its development was promoted by nationalist agitation, which hastened the restoration of the old Faeroese parliament in 1852 and the end of the Danish royal trade monopoly in 1856.

Much of the writings of these early formative years consisted of patriotic poetry. The most memorable examples of such emotional songs were produced by Fríðrikur Petersen, Rasmus Effersøe, and Jóannes Patursson.

ICELANDIC LITERATURE

The literary and linguistic renaissance in Iceland at the start of the 19th century was fostered by three men in particular: a philologist, Hallgrímur Scheving; a poet and lexicographer, Sveinbjörn Egilsson; and a philosopher and mathematician, Björn Gunnlaugsson. The principal movement in this renaissance was Romanticism. Inspired by the German philosopher Henrik Steffens, Bjarni Thorarensen produced nationalistic poetry that became a model for 19th-century lyrical poetry. Jónas Hallgrímsson, however, surpassed Thorarensen as a metrist. He was one of four involved in the periodical *Fjölnir* ("The Many-Sided"), which aimed to revolutionize literary theory and practice. The Fjölnismenn were anti-traditional and rejected the use of rhymes.

Literary and linguistic renaissance

The group was replaced in the 1840s by another group of poets, of whom the most outstanding were Benedikt Gröndal, Steingrímur Thorsteinsson, and Matthías Jochumsson. Gröndal wrote powerful lyric poetry, two prose fantasies, and an autobiography, *Daegradvöl* (1923; "Day-Spending"). Thorsteinsson wrote nature poetry and satirical epigrams but is best remembered as translator of *King Lear* (1878) and *A Thousand and One Nights* (1857–64). Jochumsson's *Hallgrímur Pétursson* (1874) and hymn *Fadir andanna* (*c.* 1884; "Father of Spirits") established him as the greatest lyric poet of the three. He, too, translated Shakespeare in addition to Ibsen's *Brand.* A poet, Grímur Thomsen, was contemporary with but distinct from this group; his poetry was less lyrical, more austere and rugged, as *Hemings flokkur Áslákssonar* (1885; "The Story of Heming Aslakssonar") exemplifies.

The latter part of the century produced three talented poets: Thorsteinn Erlingsson, author of *Aldaslagur* (1911; "Sound of the Ages"); Einar Benediktsson, a Neoromantic; and Stephan G. Stephansson, an embittered expatriate whose irony passed in Iceland for Realism.

The 19th century also saw a renaissance in imaginative prose. Jón Thoroddsen wrote two novels that have acquired a position not incommensurate with that of the medieval sagas: *Piltur og stúlka* (1850; *Lad and Lass*) and the incomplete *Madur og kona* (1876; "Man and Woman"), distinguished in prose style, narrative skill, wit, and perceptive observation of peasant and small-town life.

The 20th century

NORWEGIAN

In the 1890s established Norwegian writers came under fire from the new generation. The manifesto of new ideas was an essay published in 1890 in the periodical *Samtiden* ("The Present Age") by Knut Hamsun, "Fra det ubevidste Sjaeleliv" ("From the Unconscious Life of the Mind"), which demanded attention to what was individual and idiosyncratic rather than typical. Hamsun was impatient with contemporary emphasis on social problems, and his early novels—*Sult* (1890; *Hunger*), *Mysterier* (1892; *Mysteries*), and *Pan* (1894)—exemplified these ideas; his later novels, such as *Markens grøde* (1917; *Growth of the Soil*), were less extreme but still showed a strong, sometimes savage irony. Hamsun won the Nobel Prize for Literature in 1920.

Knut Hamsun's emphasis on the individual

Lyric poetry at this time flourished with Sigbjørn Obstfelder, who had a close affinity with the Symbolist movement, and Nils Collett Vogt, who produced some of the best lyrics of the 1890s. In drama Gunnar Heiberg, who combined a sharply satirical wit with a lyric deftness, expressed the new spirit in *Kong Midas* (1890), *Gerts have* (1894; "Gert's Garden"), *Balkonen* (1894; "The Balcony"), and *Kjaerlighetens tragedie* (1904; "The Tragedy of Love"). Sharing Hamsun's preoccupation with the irrational side of human conduct was Hans E. Kinck, a writer of considerable power and penetration. In his verse drama *Driftekaren* (1908; "The Drover") and long novel *Sneskavlen brast* (1918–19; "The Avalanche Broke"), Kinck showed himself to be a more reflective and analytical writer than Hamsun.

The real achievements of Norwegian literature in the first half of the 20th century were in the novel and lyric poetry. Drama was not conspicuous, except for the plays of Gunnar Heiberg and Nordahl Grieg. In the early decades of the century, regionalism was a strong element, particularly in the novel; and authors adopted language coloured by dialect, thus becoming identified with their

Central theme of Sigrid Undset's novels

region. Kristofer Uppdal, of the mid-north region of Trøndelag, wrote a remarkable work—a 10-volume novel cycle, *Dansen gjenom skuggeheimen* (1911–24; "The Dance Through the Shadow World"). The novel also treated of conflicts arising from the spread of industrialism, which Norway underwent later than did other European countries. The most proletarian writer was Oskar Braaten, but superior as an artist was Johan Falkberget, who wrote with understanding and historical insight about the miners in Røros in *Christianus Sextus* (1927–35) and in *Nattens brød* (1940; "Bread of Night"). Sigrid Undset, who won the Nobel Prize for Literature in 1928, set her novels in many different ages, and their concern was to examine women's loyalties within the framework of their role in society. A long historical novel, *Kristin Lavransdatter* (1920–22), was a masterpiece of Norwegian literature. Her later novels, *Gymnadenia* (1929; *The Wild Orchid*) and *Den braendende busk* (1930; *The Burning Bush*), were greatly influenced by her conversion to Roman Catholicism. Olav Duun, again of the mid-north region, revealed his insight into life as endless conflict in a six-volume novel cycle about the development of a peasant family through four generations—*Juvikfolke* (1918–23; *The People of Juvik*).

Shortly before World War I, there were several good lyric poets: Herman Wildenvey, Olaf Bull, Tore Ørjasaeter, and Olav Aukrust. Between World Wars I and II, there emerged many socially committed writers: the poet Arnulf Øverland; a novelist and critic, Sigurd Hoel; a dramatist and critic, Helge Krog; and Nordahl Grieg. After World War II, Tarjei Vesaas wrote a remarkable series of novels, including the symbolic *Huset i mørkret* (1945; "The House in the Darkness") and *Bruene* (1966; "The Bridges"). Cora Sandel, who had made a major contribution with her "Alberte" trilogy (1926–39), continued to write, as did Aksel Sandemose, an experimental writer, and Johan Borgen, who won acclaim for his early short stories, the *Lillelord* trilogy (1955–57), and the autobiographical *Barndommens rike* (1965; "Childhood's Realm"). Borgen later became the leading novelist in Norway and maintained this standing until his death in 1979. Since then, Terje Stigen, Knut Faldbakken, and Bjørg Vik have become the dominant figures in prose fiction. Stigen's works are basically realistic narratives that variously treat historical and contemporary subjects. Faldbakken has demonstrated much fantasy and ingenuity, most recently having completed a series of novels that portray the collapse of technological society. An excellent short-story writer, Vik centres her attention on middle-class family life and often portrays it from a mildly feminist viewpoint.

SWEDISH

Mood of decadence and pessimism

The early years of the 20th century were a period of decadence and pessimism in Swedish literature. Representative of this mood were Hjalmar Söderberg and Bo Bergman. Söderberg's forte was the short story (*Historietter* [1898]), in which psychological subtlety and irony were happily combined and in which, as in his novels *Martin Bircks ungdom* (1901; "Martin Birck's Youth") and *Doktor Glas* (1905), he appeared as a master of Swedish prose. Bergman also produced memorable short stories, but his real medium was the lyric; he developed his talent in a series of collections from *Marionetterna* (1903; "The Marionettes") to *Riket* (1944; "The Kingdom").

The modern Swedish novel. The development of the novel was associated with Gustaf Hellström, Ludvig Nordström, Elin Wägner, and Sigfrid Siwertz. Hellström's work as a journalist in Europe, the United States, and England greatly influenced him. Irony and careful detail emerged in his best known novel, *Snörmakare Lekholm får en idé* (1927; *Lacemaker Lekholm Has an Idea*). Siwertz was a more elegant stylist, and a decisive influence upon him was the philosophy of Henri Bergson, reflected in *En flanör* (1914; "An Idler"); but his weightiest work was a family saga, *Selambs* (1920; *Downstream*), a novel of Stockholm during World War I. Nordström, overflowing with vitality and keen but grotesque humour, accomplished some of his best work in *Landsorts-bohème* (1911; "Small-Town Bohemia") and in his short stories—*e.g., Fiskare* (1907; "Fishermen") and *Öbacka-bor* (1921). Elin Wägner was an ardent pacifist and feminist; her most powerful work was a peasant novel, *Åsa-Hanna* (1918). The outstanding novelist of the 1920s was Hjalmar Bergman: with vivid imagination and restless energy, Bergman wrote a long series of stories, many set in "Wadköping" (his native Örebro), others in Italy. In *Loewenhistorier* (1913) he depicted an irrational, impulsive, unsuccessful hero; in *Farmor och vår Herre* (1921; *Thy Rod and Thy Staff*) he portrayed one of the dominating female personalities that fascinated him. The satire *Markurells i Wadköping* (1919; *God's Orchid*) and *Swedenhielms* (performed 1925), one of the few Swedish comedies, were his most widely known works.

The "proletarian" novel

Meanwhile, the "proletarian" novel had been developed by writers concerned with the miseries of the working class, particularly Martin Koch and Ivar Lo-Johansson. There was particularly harsh criticism of working class conditions in stories by Jan Fridegård. Vilhelm Moberg wrote novels of peasant life but achieved his greatest success with the four-part prose epic about a group of Swedish emigrants to North America, *Utvandrarna* (1949–59; *The Emigrants*). The development of the Swedish autobiographical novel was helped by Eyvind Johnson, with the series "Romanen om Olof" (1934–37); Harry Martinson, with *Nässlorna blomma* (1935; *Flowering Nettle*) and *Vägen ut* (1936; "The Way Out"); and Agnes von Krusenstjerna. In her novel cycles, the "Tony" trilogy (1922–26) and the "Fröknarna von Pahlen" series (1930–35), Krusenstjerna described her own aristocratic environment and analyzed a degenerate psychology. Harry Martinson was one of a group of five primitivist writers formed about 1930. He later developed into one of the finest lyricists of the century. Sensuous imagery and a feeling for nature characterized his work. He attempted to revive the verse epic in his *Aniara* (1956), a symbolical story of a voyage of a spaceship.

Pär Lagerkvist's work

The internationally best known Swedish writer of the 20th century was Pär Lagerkvist, who won the Nobel Prize for Literature in 1951. In his youth a bold innovator, he later developed an admirably pure prose style, as in the allegorical novel *Dvärgen* (1944; *The Dwarf*). His collections of poems, *Ångest* (1916; "Anguish"), and early plays—for instance, *Himlens hemlighet* (1919; "The Secret of Heaven")—were Expressionistic in style. The dominant theme throughout Lagerkvist's work was a search for vital, often outspokenly religious values.

Development of lyric poetry. Several of the best Swedish writers were connected with the development of lyric poetry. One of the most notable, Vilhelm Ekelund, was in his youth the chief exponent of Symbolism in Sweden and later, as an author of aphorisms, exerted much influence on the development of literary modernism. Among the most popular poets were Dan Andersson, Birger Sjöberg, and Hjalmar Gullberg. In Gullberg's poetry, religious commitment and classical learning are balanced by irony and wit. A more esoteric style in modernism was introduced by Bertil Malmberg and developed by the group of poets called the generation of the 1940s, which included Erik Lindegren and Karl Vennberg. Stylistically influenced by T.S. Eliot, they often expressed an anguish and disbelief that approached French Existentialism. Lindegren's *Mannen utan väg* (1942; *The Man Without a Way*) was typical of this generation's search for meaning in life. The most distinguished novelist of the 1940s was Lars Ahlin, who was concerned with man's search for grace through love and humiliation in works such as *Min död är min* (1945; "My Death Is Mine").

The greatest lyric poet of the century was Gunnar Ekelöf. His first collection of poems, *Sent på jorden* (1932; "Late on Earth"), was heralded as the first specimen of Surrealism in Swedish literature. Ekelöf's later development passed through successive phases of Romanticism and anti-poetic Skepticism resolved in a trilogy of books blending autobiography and Eastern mysticism.

Emphasis on political and social issues

Contemporary trends. In reaction to the literature of the 1940s and 1950s, which was much concerned with artistic form and the individual approach to life, the 1960s was a period of political and social commitment in poetry and fiction alike. Recurrent topics were the war in Vietnam and bitter onslaughts on the Swedish welfare state. Independent lyric poetry, however, continued to be produced by writers such as Östen Sjöstrand and Thomas Tranströmer, and a tortured experience of life, coloured by Roman Catholicism, was forcefully expressed in the novels of Birgitta Trotzig.

The aforementioned movement toward and subsequently away from profoundly politically committed literature is exemplified by the work of Sara Lidman. During the 1950s she was one of Sweden's most creative novelists but then ceased producing fiction in order to take part in the political debate of the time. Since the late 1970s, however, Lidman has returned to creative literature with a series of novels centred on life in an isolated Swedish community. Sven Lindqvist went through a similar process; after a period of committed writing, he returned in *En älskares dagbok* (1981; "A Lover's Diary") to a more or less autobiographical novel of his own youth. Political writing persists in Sweden, but it has become more imaginative and less tied to immediate events. P.C. Jersild, for example, has painted a chilling picture of civilization after a devastating nuclear war in *Efter floden* (1982; "After the Flood"); he had earlier demonstrated his talent in allegories set in a state veterinary institution and in a hospital. Sven Delblanc also has made use of allegory, and there is sometimes an almost mystical intensity apparent in his work.

Developments in Finno-Swedish literature. The second flourishing of Finno-Swedish literature occurred in the 1920s, with the development of modernism in lyric poetry. This trend was initiated by Edith Södergran, whose visionary, dreamlike poems proved influential throughout much of Scandinavia. After her came such poets as Gunnar Björling, noted for his impressionistic pictures of nature; Rabbe Enckell, a key theoretician of the movement; and Elmer Diktonius, devoted to political concerns.

Among the most talented prose writers of the period was Runar Schildt, whose short stories dealt with such questions as the relation of the intellectual and the artist to life. Several outstanding writers appeared somewhat later in the century, as, for example, Tito Colliander, who treated themes of guilt and atonement from a religious standpoint; Christer Kihlman, whose novels combined social criticism with psychological commentary; and Tove Jansson, internationally famous for her imaginative portrayals of the fairy-tale realm of Moomintrolls.

DANISH

Georg Brandes' radical ideas

The influence of Georg Brandes. About 1870 there arose in Denmark a new movement, led by Georg Brandes, from which a modern (*i.e.,* a Naturalistic or Realistic) literature emerged. His *Hovedstrømninger i det 19de aarhundredes litteratur* (1872–90; *Main Currents in 19th Century Literature*), describing the growth and defeat of reaction, caused a great sensation. As noted earlier, he influenced Ibsen and Strindberg and wrote many scholarly and critical works illustrating radical ideas. His later biographies of Shakespeare, Goethe, Voltaire, Julius Caesar, and Michelangelo revealed how he was influenced by Nietzsche into developing a philosophy of aristocratic radicalism. Among his followers were Jens Peter Jacobsen, whose short story "Mogens" (1872) and novel *Fru Marie Grubbe* (1876) are the supreme examples of Danish Naturalism, while his other novel *Niels Lyhne* (1880) and some of his short stories dealt with dream as against reality; and Holger Drachmann, greatest lyric poet of the period, who later reacted strongly against Brandes and whose poetry and prose were often about the sea.

Henrik Pontoppidan, one of Denmark's greatest novelists, dealt at first with social injustices and contemporary political, moral, and religious problems in his short stories. The Denmark of his day was also the subject of his greatest work, three long novel cycles, *Det forjaettede land* (1891–95; *The Promised Land*), *Lykke-Per* (1898–1904; "Lucky Peter"), and *De dødes rige* (1912–16; "The Realm of the Dead"); and in these he makes penetrating, if unflattering, analyses of Danish national character. Herman Bang was another novelist interested in the outsiders of life and in insignificant people. His skillful, mainly impressionistic technique was displayed in his best novels, *Ved vejen* (1886; "By the Way-Side"), *Tine* (1889), and *Det hvide hus* (1898; "The White House").

Other notable writers at the end of the century were Gustav Wied, whose "satyr plays" and whose novels *Livsens Ondskab* (1899; "Life's Malice") and *Knagsted* (1902) were full of malicious humour; Vilhelm Topsøe, a conservative realist; Peter Nansen, who wrote stories reminiscent of those of Guy de Maupassant; Carl Ewald, whose nature stories were based on Darwinian philosophy; Karl Larsen, who caught the atmosphere of Copenhagen and its inhabitants with fine precision; and several playwrights, including Edvard Brandes, Otto Benzon, Gustav Esmann, Sven Lange, Einar Christiansen, and Henri Nathansen.

Symbolist mode

Neoromantic revival. In the 1890s a Neoromantic poetic revival occurred, reinstating the value of emotion and fantasy. The leader of these Symbolist poets was Johannes Jørgensen, whose finest works show a simplicity of style and intensity of feeling. Other poets of the time included Viggo Stuckenberg, who expressed sad resignation; Sophus Claussen, whose poems, often obscure, show sensuality, pantheistic love of nature, and sophisticated aestheticism; and Helge Rode, a mystic who also wrote plays and criticism attacking intellectualism.

20th-century literary trends. Several women contributed to literature at the turn of the century: Gyrithe Lemche, who wrote a novel cycle, *Edwardsgave* (1900–12); Agnes Henningsen, a brilliant writer who was often concerned with experiences of the emancipated woman; and Karin Michaëlis, a fine psychologist, best known for her novel *Den farlige alder* (1910; *The Dangerous Age*).

Nexø's epics of proletarian life

The two greatest early 20th-century novelists were Martin Andersen Nexø and Johannes Vilhelm Jensen. Nexø's works described the lives of poor people; *Pelle Erobreren* (4 vol., 1906–10; *Pelle the Conqueror*) and *Ditte Menneskebarn* (3 vol., 1917–21; *Ditte: Daughter of Man*) were great epics of proletarian life, and his reminiscences were among the finest in the language. Jensen, who was also a great and original lyric poet and prolific essayist, wrote *Den lange rejse* (6 vol., 1908–22; *The Long Journey*), an ambitious epic of man from the baboon stage to the discovery of America; he was also noted for *Himmerlandshistorier* (1904, revised 1910; "Tales from Himmerland"), based on his childhood memories of North Jutland; *Kongens fald* (1900–01; *The Fall of the King*); and nine volumes of *Myter* (1907–44; "Myths"). Other novelists of this period included Jakob Knudsen, whose interest, taking account of the inequality of man and the need for authority, was with Christian and moral problems; Harald Kidde, an introspective and melancholy writer; and Knud Hjortø, a keen and intelligent writer of psychological novels.

Regional literature of the early 1900s was produced chiefly by Jutland writers. Prominent among them were three poets: Jeppe Aakjaer, Johan Skjoldborg, and Thøger Larsen. Also, Marie Bregendahl and Harry Søiberg drew upon Jutland settings for their novels.

Isak Dinesen

Significant poets of the post-World War I generation were Tom Kristensen, Otto Gelsted, Emil Bønnelycke, Kai Friis Møller, and Per Lange. Among interesting novelists, Jacob Paludan wrote some very good fiction—*Fugle omkring fyret* (1925; *Birds Around the Light*) and *Jørgen Stein* (1932–33)—as also did Hans Kirk, whose *Fiskerne* (1928; "The Fishermen") was social realism at its best. Harald Herdal, a disciple of Nexø, exposed society's hypocrisy in his proletarian novels. Jørgen Nielsen's themes were suppressed hatred, sin, and fear among Jutland peasants. H.C. Branner, an important writer of novels, plays, and short stories, spoke of the loneliness of men and the danger of power. Another writer of these three genres was Knud Sønderby, who had a brilliant style and deep understanding. Nis Petersen, poet and novelist, was famous for *Sandalmagernes gade* (1931; *The Street of the Sandalmakers*) and *Spildt maelk* (1934; *Spilt Milk*). Isak Dinesen (Karen Christence Dinesen, Baroness Blixen-Finecke), an aristocratic writer with subtle irony and unusual sensitivity, wrote both in Danish and in English. Her first notable work, a collection of short stories featuring a strong fairy-tale-like quality, was in fact written in English. Entitled *Seven Gothic Tales*, it was published in 1934 in the United States and subsequently translated by the author into Danish as *Syv fantastiske Fortællinger*. Other major works of Dinesen include her memoir *Den afrikanske Farm* (1937; *Out of Africa*) and two more collections of finely crafted stories, *Vinter-Eventyr* (1942; *Winter's Tales*) and *Sidste Fortællinger* (1957; *Last Tales*).

Two Faeroese novelists also made important contributions to modern Danish prose fiction: Jørgen-Frantz Jacobsen, with his novel *Barbara* (1939), which provides a fascinating portrait of a capricious woman; and William Heinesen, with his masterpiece *De fortabte spillemænd* (1950; *The Lost Musicians*). Here, as in the rest of his varied writings, Heinesen renders Faeroese life as a microcosm illustrative of social, psychological, and cosmic themes. Other distinguished novelists of the time were Hans Scherfig, a great humorist and social satirist; and Martin Alfred Hansen, a psychological novelist, whose best known novel is *Løgneren* (1950; *The Liar*).

The modern Danish theatre

Danish playwrights of the post-World War I period such as Sven Clausen and Svend Borberg were influenced by German Expressionism, Symbolism, Luigi Pirandello, and Sigmund Freud. Kaj Munk, who revived the heroic drama of Shakespeare and Schiller, showed unusual qualities in his best plays—*En idealist* (1928; *Herod the King*) and *Ordet* (1932; *The Word*)—which were concerned with problems of God and man. The work of Kjeld Abell marked a severance from Naturalist drama, and a radical perspective underlay his witty dialogue. His most important plays were *Melodien, der blev vaek* (1935; *The Melody That Got Lost*), *Anna Sophie Hedvig* (1939), *Dage på en sky* (1947; *Days on a Cloud*), and *Skriget* (1961; "The Scream"). C.E. Soya was an important playwright of the period, and also a novelist and fine short-story writer; some of his daring experiments with the theatre have been very successful.

Many postwar poets found an aesthetic manifesto in *Fragmenter af en dagbog* (1948; "Fragments of a Diary"), by Paul la Cour, who was influenced by contemporary French poetry. Jens August Schade, a sophisticated naivist, also had an important influence, as did the prematurely deceased poets Gustaf Munch-Petersen and Morten Nielsen. A revival of poetry followed the liberation (1945), and the Existentialist periodical *Heretica* (1948–53) became the voice of a group of young writers who regarded a Christian philosopher, Vilhelm Grønbech, as their spiritual progenitor. Two outstanding poets apart from the *Heretica* group were Halfdan Rasmussen, who also wrote excellent nonsense verse, and Erik Knudsen, also a brilliant satirical playwright. Both studied contemporary problems and reacted against the anti-rationalism and anti-intellectualism of the *Heretica* movement. Tove Ditlevsen was another important poet, as well as novelist and short-story writer, unattached to any group; his often intensely personal work reflects the loneliness of life in the poorer quarters of Copenhagen. Klaus Rifbjerg has been the dominant novelist in Denmark since the publication of his *Den kroniske uskyld* ("Chronic Innocence") in 1958. He is a writer of great inventiveness and linguistic originality, who has analyzed modern society and its problems, both public and private, in realistic novels. In a more satirical vein, Leif Panduro has examined the place of the individual in society, paying special attention to the problems of middle age and the emptiness of a welfare-state society in novels and television dramas. A more philosophical approach is found in Villy Sørensen, whose Kafkaesque stories place him in the sphere of Absurd literature. There has been, at the same time, a vogue of the documentary novel, the principal exponent of which has been Thorkild Hansen.

The modern poetry that became a hallmark of Danish literature after World War II gave rise to the mature works of Frank Jæger, Thorkild Bjørnvig, and Ivan Malinovski. Further experimentation resulted in Structuralist works such as Inger Christensen's *Det* (1968; "It"). Henrik Nordbrandt has been the leading Danish poet of the 1970s and '80s. A pessimistic sense of isolation pervades his poems, which have been influenced by Oriental writings.

The 1970s saw an upsurge of women authors, some consciously writing for the feminist cause and others content to portray life as experienced by women. Outstanding among them are Dea Mørch, Kirsten Thorup, and Dorit Willumsen.

FAEROESE

Faeroese literature came into its own after the turn of the century. Jens H.O. Djurhuus, who created rhetorical poetry of splendid resonance, was the first to emerge as a writer of international stature. His brother, Hans Andrias Djurhuus, wrote in a more naive manner, producing poems, fairy tales, and plays, which were based on native historical traditions and legends.

Faeroese golden age writers

Five writers dominated the Faeroese literary scene from about the 1930s through mid-century. Of these so-called Faeroese golden age authors two, Jørgen-Frantz Jacobsen and William Heinesen, wrote in Danish (see above), while the other three, Christian Matras, Heðin Brú (Hans Jakob Jacobsen), and Martin Joensen, in Faeroese. The works of Matras reveal a profound lyric poet seeking to interpret the essence of Faeroese culture. A fine stylist, Brú did much to create a Faeroese literary prose in his portrayals of village life in a time of transition (*e.g., Feðgar á ferð* [1940; *The Old Man and His Sons*]). Joensen's novels and short stories are of a similar character, but their emphasis is on psychological realism rather than on style. A prose writer of a distinctly more modern bent is Jens Pauli Heinesen. His works reflect an approach to Faeroese life that is generally more international than that of Brú or Joensen and that is infused with a certain satirical element.

Poetry continues to attract many writers. Karsten Hoydal was the first Faeroese writer to compose verse directly influenced by modern foreign poets; he also translated many of their works, especially those of Edgar Lee Masters of the United States. Regin Dal and Steinbjørn B. Jacobsen have gone much further in their modernism, the latter adopting a style somewhat akin to the Imagism of the U.S. poet Ezra Pound. Other writers whose works exhibit a modernist tendency include Guðrið Helmsdal, the foremost Faeroese woman poet of the contemporary scene.

ICELANDIC

Modern Icelandic prose writing did not really develop until the late 1870s, when a group of young men, influenced by the theories of the Danish critic Georg Brandes, began their literary careers. Unfortunately, they had absorbed Brandes' ideas uncritically, which resulted in introspective, self-pitying works believed by their authors to be realistically written. The early works of Einar Kvaran suffered from this flaw, but Kvaran later developed into a novelist of skill and power.

The works of Halldór Laxness

Notable 20th-century prose writers. Several writers of this time showed a keen eye for character and an understanding of human feelings and of the stark life of rural Iceland: Jón Trausti (Gudmundur Magnússon), who wrote the cycle *Heidarbýlid* (4 vol., 1908–11; "The Mountain Cot"); Gunnar Gunnarsson, whose *Kirken på bjerget* (1923–28; "The Church on the Mountain") was written in Danish; and Gudmundur Hagalín. The outstanding modern prose writer was Halldór Laxness, who was awarded the Nobel Prize for Literature in 1955. His mature works were influenced by his conversion to Roman Catholicism and his identification with the basic ideas of Communism. His major works were *Salka Valka* (1936), *Sjálfstaet fólk* (1935; *Independent People*), *Íslandsklukkan* (1943; "The Bell of Iceland"), and *Gerpla* (1952). He helped restore Icelandic as a sensitive medium for storytelling.

Among more recent prose fiction writers, Guðbergur Bergsson has proved himself one of the most talented and forceful. Reflective of the growing social and political consciousness of the 1960s, some of his novels from that period—*Ástir samlyndra hjóna* (1967; "The Love of a Harmoniously Married Couple") and *Anna* (1969)—subjected contemporary Icelandic society and the military relations of the nation with the United States to biting satirical attacks. His later works, the collection of short stories *Hvað ereldi guðs* (1970; "What Does God Eat") and a series of novels produced in the mid-1970s, were decidedly experimental in character, revealing an attempt by the author to go beyond ordinary reality to expose some of the more disgusting and grotesque aspects of life.

Major poets. At the beginning of the 20th century, poetry had lyricists in Thorsteinn Erlingsson, whose early delicacy later developed into a more powerful note in

"Aldaslagur" (1911; "Sound of the Ages") and in an incomplete epic, "Eidurinn" (1913; "The Oath"); in Einar Benediktsson, who wrote in an ornate style sometimes capable of greatness, as "Í dísarhöll" ("In the Hall of the Muses") shows; and in Stephan G. Stephansson, an expatriate farmer in Canada, who was a more bitter poet, influenced by the "realism" that passed for Georg Brandes' ideas in Icelandic literature—but *Andvökur* (1909–38; "Sleepless Nights") revealed a sensitive spirit.

Prominent poets of the next generation included Davíd Stefánsson, a traditionalist who expressed deep personal feelings in straightforward language and simple verse forms. His approach was shared by Tómas Gudmundsson and Jón Helgason, whose book *Úr landsudri* (1939, revised 1948; "From the South") was outstanding. Steinn Steinarr (Adalsteinn Kristmundsson), who was deeply influenced by Surrealism, experimented with abstract styles and spearheaded modernism in Icelandic poetry with his collection *Ljóð* (1937; "Poems").

Since mid-century several poets have distinguished themselves. The early works of Hannes Pétursson showed great sensitivity and skill in adapting Icelandic to new, European metres. Pétursson's more recent poems (those in the collection *Ur hugskoti* [1976; "Recollections"]), however, reveal a movement away from innovative forms to more traditional verse. Still other contemporary poets of merit include Thorsteinn frá Hamri and Sigurður Pálsson. Hamri's poems *Veðrahjálmur* (1972; "Sun Rings") grapple with questions about lasting values, particularly with the possibility of realizing human fellowship in the modern world. Pálsson's *Ljóð vega salt* (1975; "Poems on the See-Saw") combine autobiographical elements with philosophical questioning about the nature of contemporary life.

The development of the Icelandic drama. Icelandic drama really started to develop through Jóhann Sigurjónsson, whose first success was *Fjalla-Eyvindur* (1911; *Eyvind of the Hills*), followed by *Galdra-Loftur* (1915; "Loftur the Sorcerer"); both plays were based on powerful folktales. Gudmundur Kamban's *Hadda-Padda* (1914) was highly praised by Georg Brandes, and he remained important in Scandinavian drama for the next quarter of a century. After Kamban, there were few plays of lasting value, though Davíd Stefánsson's *Gullna hlidid* (1941; "The Golden Gate"), Jakob Jónsson's *Tyrkja-Gudda* (published 1948), and Agnar Thórdarson's satirical comedy of modern Reykjavík life, *Kjarnorka og kvenhylli* (1957; "Nuclear Force and Female Popularity"), had considerable merit. In *Ganksklukkan* (1962; *The Cuckoo Clock*) the latter produced a powerful play on the dehumanizing effect of modern life.

BIBLIOGRAPHY. ELIAS BREDSDORFF, BRITA MORTENSEN, and RONALD POPPERWELL, *An Introduction to Scandinavian Literature, from the Earliest Time to Our Day* (1951, reprinted 1977), a useful short survey; HARALD BEYER, *A History of Norwegian Literature* (1956, reprinted 1979; originally published in Norwegian, 1952), probably the best survey in English; ALRIK GUSTAFSON, *A History of Swedish Literature* (1961), an excellent critical history, with bibliographical appendix; PHILLIP M. MITCHELL, *A History of Danish Literature,* 2nd ed. (1971); STEFÁN EINARSSON, *A History of Icelandic Literature* (1957).

Studies that focus on the works of specific periods of Scandinavian literary history include GABRIEL TURVILLE-PETRE, *The Heroic Age of Scandinavia* (1951, reprinted 1976), and *Origins of Icelandic Literature* (1953, reissued 1975); CAROL J. CLOVER, *The Medieval Saga* (1982); S.B.F. JANSSON, *The Runes of Sweden* (1962); ANTON BLANCK, *Den nordiska renässansen i sjuttonhundratallets litteratur* (1911); R. AHLÉEN, *Swedish Poets of the Seventeenth Century* (1932); ANTON NILSSON, *Svensk romantik: den platonska strömningen* (1916); BRIAN W. DOWNS, *Modern Norwegian Literature, 1860–1918* (1966), the first book in English about the "classic" period—an excellent survey, with bibliography; RICHARD BECK, *History of Icelandic Poets, 1800–1940* (1950); STEFÁN EINARSSON, *History of Icelandic Prose Writers, 1800–1940* (1948, reissued 1966); J.W. MCFARLANE, *Ibsen and the Temper of Norwegian Literature* (1960, reissued 1979); HELGE G. TOPSOE-JENSEN, *Scandinavian Literature from Brandes to Our Day* (1929, reprinted 1971; originally published in Danish, 1928); SVEN HAKON ROSSEL, *A History of Scandinavian Literature: 1870–1980* (1982); KARIN ELKJAER (ed.), *Danske og udenlandske forfattere efter 1914,* 4th ed. (1977); ELIAS BREDSDORFF (ed.), *Contemporary Danish Plays* (1955, reprinted 1970), and *Contemporary Danish Prose* (1958, reprinted 1974); M.S. ALLWOOD (ed.), *20th Century Scandinavian Poetry* (1951); *Seven Swedish Poets* (Ekelöf, Gullberg, Lagerkvist, Lindegren, Malmberg, Martinson, and Södergran), ed. and trans. by F. FLEISHER (1963).

Specialized studies include: JOHN M. WEINSTOCK and ROBERT T. ROVINSKY (eds.), *The Hero in Scandinavian Literature: From Peer Gynt to the Present* (1975), film treated along with literary genres; JANET MAWBY, *Writers and Politics in Modern Scandinavia* (1979), which emphasizes the impact of the German occupation and the U.S. involvement in Vietnam; JOHN L. GREENWAY, *The Golden Horns: Mythic Imagination and the Nordic Past* (1977), a study of myth and its effect on Scandinavian literature and life.

(B.S.B./S.Be./J.W.McF./B.Mo./E.O.G.T.-P./W.G.J.)

Classical Scholarship

Classical scholarship comprises the study, in all its aspects, of ancient Greece and Rome. In continental Europe this field is known as "classical philology"; the use, in some circles, of "philology" to denote the study of language and literature—the result of abbreviating the 19th-century "comparative philology"—has lent an unfortunate ambiguity to the term. During the 19th century Germans evolved the concept of *Altertumswissenschaft* ("science of antiquity") to emphasize the unity of the various disciplines of which the study of the ancient world consists. Broadly speaking, the province of classical scholarship in time is the period between the 2nd millennium BC and AD 500 and in space the area covered by the conquests and spheres of influence of Greece and Rome at their widest extent.

This article surveys the history of classical scholarship thus defined from antiquity until the late 20th century. For full treatment of the revival of classical culture and thought during the Renaissance, see HUMANISM; see also GRECO-ROMAN CIVILIZATION, CLASSICAL; and HISTORY, THE STUDY OF.

The article is divided into the following sections:

ANTIQUITY AND THE MIDDLE AGES

Until the Renaissance, Greek scholarship in the East and Latin scholarship in the West tended to follow different courses, and it is therefore convenient to treat them separately during this period.

Exposition of Homeric poetry

Greek scholarship. *Beginnings.* Greek epic poetry was recited in early times by professional performers known as rhapsodists, or rhapsodes, who sometimes offered interpretations of the works as well. In the 6th century BC Theagenes of Rhegium is said to have "searched out Homer's poetry and life and date," to have offered an allegorical interpretation of the battle of the gods in the 20th book of the *Iliad,* and to have been cited for a variant in Homer's text. The Sophists of the 5th century BC—paid writers, lecturers, and teachers such as Protagoras, Prodicus, Gorgias, and Hippias—gave ethical instruction in the form of the exposition of poetry, particularly that of Homer, which from this time formed the staple of Greek education. Some of them were interested in etymology, phonetics, the exact meanings of words, correct diction, and the classification of the parts of speech. Hippias laid the foundations of ancient chronography by making a list of victors in the Olympic Games, and Alcidamas (*c.* 400 BC) wrote a book on Homer. However, the efforts of the Sophists in this direction, considerable as they were, had a more or less casual and arbitrary character.

Plato (*c.* 428/427–348/347 BC) strongly resisted the claim that the poets were reliable interpreters of religion and morality. In his dialogue *Cratylus* he rejected the theory that the study of words can reveal the meaning of things, insisting that things themselves must be studied. Plato's pupil Aristotle (384–322 BC) defended poetry against his master; he valued highly the *Iliad* and the *Odyssey,* which from his time were regarded (together with the mock-epic *Margites*) as the genuine works of an individual Homer. He took a similar view of tragedy, which he believed effected a purification (*katharsis*) of the emotions upon which it played. Aristotle wrote about linguistic, dramatic, and other problems in Homer, refuting such detractors of the poet as Zoilus, compiled lists of Olympic and Pythian victors, collected details about the Athenian tragic and comic festivals, and supplemented his *Politics* with a collection of 158 studies of the constitutions of various Greek states. He also carried further the discussion of the constituent parts of a sentence and discussed the nature of synonyms, compounds, and rare words in early poetry.

The school of Aristotle

The school of Aristotle, known as the Lyceum, or Peripatos, continued to make this kind of learned work an adjunct to its philosophical activities. Aristotle's successor, Theophrastus (*c.* 372–*c.* 287 BC), collected the opinions of earlier philosophers. Dicaearchus (flourished *c.* 320 BC) wrote about the life of Greece, and Aristoxenus (flourished late 4th century BC) about the history and the theory of music. Heracleides Ponticus (*c.* 390–*c.* 322 BC) wrote one book on Archilochus and Homer and another on the dates of Homer and Hesiod. Clearchus collected proverbs, and Demetrius of Phaleron fables. All these philosophers were guided by Aristotle's teleological concept of intellectual activity, according to which philosophy is the culminating element of civilization. A 4th-century commentary on an Orphic poem, discovered in 1963 on a papyrus from a grave in Derveni, Macedonia, deserves mention as the earliest known commentary on a text; it is not a linguistic commentary but offers an allegorical interpretation that is doubtless very different from what the poet had intended.

The Library of Alexandria

During the Hellenistic Age (usually reckoned to extend from the death of Alexander the Great in 323 BC to the 1st century AD) scholarship flourished nowhere more than in the great city of Alexandria, the capital of the Ptolemies, the kings of Egypt. Early in the 3rd century BC Ptolemy I founded the famous Mouseion (Museum) of Alexandria, a community of learned men organized along the lines of a religious cult and headed by a priest of the Muses; part of the Museum was a splendid library that became the most celebrated of the ancient world. In its establishment the king is said to have had the assistance of the eminent Peripatetic scholar and statesman Demetrius of Phaleron, who left Athens about 300 BC; unfortunately, the evidence about the part he played is scanty and unreliable. The Museum community included both poets and scholars, as well as several individuals who combined these pursuits. From the time of the poet-scholar Philetas, or Philitas (*c.* 330–*c.* 270 BC), the tutor of Ptolemy II, the scholars there were much concerned with the collection and interpretation (*glossae*) of rare poetic words. Philetas' pupil Zenodotus of Ephesus (*c.* 325–260 BC) was the first librarian at Alexandria; using the manuscripts collected for the Library but also trusting to his own judgment, sometimes in a manner that seemed to later critics dangerously subjective, he made the first critical edition of Homer, marking passages of doubtful authenticity with critical signs in the margins. Zenodotus also edited Pindar and Anacreon and perhaps other lyric poets; at about the same time the epic and elegiac poet Alexander Aetolus is said to have edited the tragic poets, and the dramatic poet Lycophron the comic poets, but singularly little is known about these editions.

Somewhat later the great poet Callimachus (*c.* 305–*c.* 240 BC) compiled the *Pinakes* ("Tablets"), a vast catalogue raisonné of the chief authors, with biographical and bibliographical information. Callimachus is said to have written a book opposing the chief Peripatetic critic of the time, Praxiphanes, and is widely held to have criticized Peripatetic literary theory; but the scantiness of the evidence for this enjoins great caution.

Rather later the great geographer and mathematician Eratosthenes (*c.* 276–*c.* 194 BC), the third librarian, laid the foundations of a systematic chronography; more of his work would be known had it not been largely superseded in popular use by the 2nd-century chronicles of Apollodorus of Athens, which were a learned compilation but left out the important scientific and mathematical part.

Zenodotus' editions of Homer and Hesiod were improved upon by the fourth librarian, Aristophanes of Byzantium (*c.* 257–180 BC), who also edited the lyric poets, setting out their verses according to a systematic metrical theory; edited Aristophanes, Menander, and perhaps other comic poets; edited Sophocles and at least part of Euripides; and compiled useful summaries of the plots of plays with details of their productions. His *Lexeis* ("Readings") was the most important of the numerous lexicographical works produced at this time, which included lexicons of particular authors and dialects; he also wrote some of the many treatises about literature that were now appearing.

Aristarchus of Samothrace (*c.* 217–145 BC), the sixth librarian, wrote not only monographs about poetry but also important commentaries on Homer, Pindar, and much of tragedy and comedy. Aristarchus was one of the many learned men who left Alexandria in consequence of the disastrous persecution of learning by Ptolemy VIII, from which that city's standing as a great centre of learning never quite recovered. (The great library survived a fire set in Alexandria in 47 BC by Julius Caesar, whose army supported Cleopatra in a civil war. It was finally destroyed in AD 391 by the patriarch Theophilus of Alexandria.)

During the 3rd century BC the Stoics, particularly Chrysippus (*c.* 280–*c.* 206 BC), made important contributions to the study of grammar, linked with the development of Stoic logic. Early in that century the Stoic Crates of Mallus emigrated to the court of King Eumenes II of Pergamum, which the Attalid dynasty had begun to make into a literary centre comparable with, though hardly equal to, Alexandria. Crates probably wrote commentaries on the *Iliad* and the *Odyssey,* characterized by the allegorical interpretation, faith in the accuracy of Homer's geography, and grammatical rigour typical of the Stoic school. Under Stoic influence the Pergamenes tended to stress the element of anomaly in grammar, while the Alexandrians stressed the element of analogy; that is, the Alexandrians insisted on the natural, inherent orderliness of grammar, while the Pergamenes approached the subject as empiricists, being content to organize observations of actual usage into a body of knowledge. But the details of the alleged controversy over this matter are obscure and known largely from suspiciously late sources. If the extant grammar ascribed to Dionysius Thrax, a pupil of Aristarchus active about 120 BC, is genuine, then the Alexandrian school of grammar was by that time already considerably influenced by the Stoics.

During the 1st century BC, by which time Rome was beginning to be the chief centre of Greek scholarship, Philoxenus wrote on Greek dialects, among which he included Latin; he was the first scholar to be aware of the existence of monosyllabic roots. Under Augustus, Tryphon studied the language of prose and made the first study of syntax, the first vocabulary of the written language, and a classification of the so-called figures of speech. About the same time Didymus, known as Chalcenterus ("Brazen-Gutted"), incorporated into huge variorum editions much of the precious material contained in the many commentaries on literature compiled during the Hellenistic Age. This vastly productive scholar was lacking in critical judgment, but it is on his work that the later less extensive commentaries that in part survive depended. Under Tiberius, Theon studied the Hellenistic poets, as well as Pindar.

The Attic
Revival

The 1st century AD saw the beginning of the "Attic Revival," the movement to imitate the language and style of the classical Athenian writers, which lasted far into the Byzantine period with disastrous effects that have not even yet died away. This resulted in the production of many lexica and manuals meant to help people to write correct Attic, such as the works of Phrynichus, Moeris, and Pollux, all probably dating from the 2nd century AD. At that time much learned work was still being done, but it was becoming increasingly mechanical and repetitive. More and more of the chief writers survived only in selections; texts were being produced, often with commentaries, but these derived mainly from the stores of learning accumulated in the past. However, under Hadrian, Apollonius Dyscolus produced a treatment of syntax that acquired great authority, and his son Herodianus produced the standard treatise on accentuation; they were the last known producers of important original work on grammar.

The rise
of Christianity

Christianity proved less hostile to pagan culture than might have been expected. From the 2nd century on, Church Fathers such as Justin, Clement of Alexandria, and Origen used an impressive knowledge of pagan literature to debate with pagan philosophers on equal terms. Prominent on the pagan side was the Neoplatonist Porphyry (*c.* 234–*c.* 305). Besides his published attacks on Christianity, he wrote commentaries on Plato, Aristotle, Theophrastus, and Plotinus. Even after the triumph of Christianity in 313 under Constantine the Great, pagan and Christian scholars often attended one another's lectures. The pagan Libanius of Antioch, the most celebrated rhetor of the time and author of the surviving hypotheses of the orations of Demosthenes, taught Theodore of Mopsuestia, St. John Chrysostom, and probably also St. Basil and Gregory of Nazianzus. Basil (*c.* 329–379) wrote a treatise on the value of pagan literature in which he recommends at least a passing acquaintance with the pagan classics, but he and the other leading Christian authors of his time possessed a good deal more than this. Theodore (*c.* 350–428/429), bishop of Mopsuestia and leader of the school of Antioch, applied what could be called pagan methods of criticism to the Bible by using his knowledge of history and language to illuminate passages of Scripture. Members of the Christian school of Gaza in the 5th and 6th centuries even wrote dialogues modeled on those of Plato. The school's leading member, Procopius, invented the *catena* ("chain"), a commentary on a book of the Bible consisting of a compilation of excerpts from earlier commentaries—something obviously suggested by the variorum editions of classical authors. Notes based on the learned commentaries of the Hellenistic Age now came to be written into the margins of manuscripts; to these scholia is owed most of what is known of ancient scholarship.

The Neoplatonists of the 5th and 6th centuries produced commentaries on Plato, Aristotle, and other philosophers, thus preserving many priceless fragments of earlier philosophical texts now lost. Grammatical work also continued: Proclus wrote a commentary on Hesiod's *Works and Days;* Hesychius of Alexandria compiled a Greek lexicon that preserved vocabulary from the Homeric age up to his own time; and Orus contributed to the work on Greek orthography. Education even received some government support; the 4th-century rhetor Themistius described a plan for the creation of a government scriptorium to ensure the survival of important writers, and some 50 years later, in 425, Emperor Theodosius II is said to have set up a university at Constantinople.

The age of Justinian I (527–565) produced the antiquarian works of Johannes Lydus and the geographical gazetteer of Stephanus of Byzantium. The historians of that era, Procopius and Agathias, wrote in the classical tradition of historiography, publishing chronicles of warfare that weighed the influences on historical events of fate and divine retribution. But in 529 Justinian issued an edict closing the schools of pagan philosophy; some philosophical activity continued after that, but the edict marked an era of Christian intolerance of pagan scholarship. During the 7th century the Arab conquests cut off Syria, Palestine, and Egypt from Greek civilization. The Arab threat forced the Byzantine Empire to submit to the rule of vigorous but not well-educated emperors, some of whom were religious fundamentalists opposed to the use of images, or icons, which was a central feature of worship in the Eastern Church. The resulting Iconoclastic Controversy was a major factor in the creation of a dark age of Byzantine culture that lasted from about the middle of the 7th until the beginning of the 9th century.

The first Byzantine renaissance. The dark age was not completely dark. It saw, for example, the extensive but exceedingly uninspired work of the grammarians Georgius Choeroboscus, active during the second half of the 8th century, and Theognostus, early in the 9th century, as well as the letters of the deacon Ignatius with their surprising wealth of literary allusions. Also, certain developments that occurred at this time were important for the future. In about 800, paper was acquired from the Arabs, who are said to have learned how to make it from Chinese prisoners taken in a battle at Samarkand. It came into general use only very gradually; the Byzantines continued to import it from the Arabs instead of making their own, but since it was less expensive than papyrus, its effect was bound to be important. The Italians acquired it from the Byzantines, and by the 13th century they had developed a flourishing paper industry. From about the same time must date the invention of a new cursive script, the Byzantine minuscule, which was in its early forms the most elegant that the Greeks ever invented. The earliest surviving specimen, the Uspenskij Gospel, dates from 835, but this displays such accomplished writing that the new script probably originated some 50 years earlier. The invention greatly facilitated the rapid production of books. The Stoudion monastery in Constantinople, which flourished under its great abbot St. Theodore (759–826), was once thought to have introduced the new script—and indeed the monastery had a flourishing scriptorium—but this conjecture is by no means certain. During the 9th and 10th centuries the works of many classical authors were transferred from manuscripts in the old uncial writing to the new minuscule, and the surviving books of this period show that script in its most perfect form. Later the elegance of minuscule was spoiled by the admixture of uncial letters and the increasing use of ligatures.

Byzantine minuscule script

The first important scholar of the first Byzantine renaissance was Leo the Philosopher (*c.* 790–*c.* 869), a notable teacher in Constantinople who numbered among his pupils St. Cyril, one of the apostles of the Slavs; Leo had considerable knowledge of Greek culture, particularly of science and mathematics. But the dominant figure in the revival of the 9th century was the patriarch Photius (*c.* 820–891?), who not only compiled a notable Greek lexicon but also produced the *Myriobiblon,* or *Bibliotheca,* a vast collection of summaries and evaluations of various ancient books, mainly historical. Photius also compiled a learned miscellany called the *Amphilochia* and an interesting collection of letters. Arethas (born *c.* 850), archbishop of Caesarea Cappadociae, owned a remarkable private library, from which eight priceless books, commissioned from the finest calligraphers of the time, survive; Euclid, Plato, Aristotle, Lucian, and Aristides are among them. Other valuable classical manuscripts still extant formed part of his collection.

During the 10th century education was encouraged by the learned emperor Constantine VII Porphyrogenitus (905–959), who apart from producing his own series of historical works preserved several histories by others and planned a vast 53-section encyclopaedia of human activities that was probably never completed. The 10th century also saw the production of a large encyclopaedia cum dictionary, formerly thought to have been the work of one Suidas, but now known to have been called the *Suda,* from a Byzantine Greek word for fortress. Platonism was actively studied by the chief intellectual figure of the 11th century, Michael Psellus (1018–*c.* 1078). His numerous writings show a wide acquaintance with classical culture, though also a very imperfect sympathy with some of its elements. His pupil, John Italus, was anathematized by the ecclesiastical authorities for allowing Platonism to contaminate his Christianity. But Platonic studies continued, and Isaac Sebastocrator, a brother or son of the emperor Alexius I Comnenus, wrote three essays based on Proclus. Early in the 12th century Alexius' daughter, Anna Comnena, was the centre of a circle of Aristotelian scholars, including Michael of Ephesus and Eustratius, who together produced a commentary on the *Ethics.* Gregory of Corinth, active during the same period, wrote works on syntax and style and also one of the few ancient treatments of the Greek dialects that have come down to the present. John Tzetzes wrote some 60 books on Greek literature that are learned but uncritical, and Eustathius of Thessalonica wrote vast commentaries on the *Iliad* and *Odyssey* that incorporate much earlier learning.

Losses during the Fourth Crusade

This epoch of Byzantine learning was rudely put to an end when the knights of the Fourth Crusade, under Venetian leadership, sacked Constantinople in 1204. It may well be argued that this event was an even greater disaster for learning than the Turkish capture of the city in 1453, for which the crusaders paved the way. The sack of the city destroyed a quantity of Greek literature that is difficult to estimate; certainly included among the lost works were the *Aitia* and *Hekale* of Callimachus, which were known to Michael Choniates, archbishop of Athens at the time of the Crusade.

The second Byzantine renaissance. Between 1204, when the imperial capital was moved to Nicaea, and 1261, when Constantinople was recovered by the Palaeologus dynasty, classical studies continued under the difficult conditions outlined in the autobiography of Nicephorus Blemmydes, the leading intellectual of the time. The emperor Theodore II Lascaris (reigned 1254–58) did much to assist cultural life during this time. The period sometimes called the Palaeologan Renaissance saw a revival of classical studies that, under the circumstances, must be called remarkable. Maximus Planudes (1260–*c.* 1310) made many compilations, including a new anthology of epigrams, and even translated into Greek such Latin texts as parts of Ovid, Augustine, and Boethius. He also had some knowledge of Hellenistic poetry and even Arab astronomy and mathematics. From about 1300 the texts of the Greek dramatists were studied critically by Manuel Moschopoulos, Thomas Magister, and finally Demetrius Triclinius. Triclinius had read the metrical handbook of the 2nd-century scholar Hephaestion and understood the simpler metres, and he was also aware of the principle of strophic responsion. He was therefore able to make a number of emendations worthy of serious notice. Theodore Metochites (*c.* 1260–1332), one of the leading intellectuals and public men of his time, commented on Aristotle and wrote a miscellany that contains interesting reflections upon classical authors, especially orators and historians.

Greek in the West. During the 3rd and 4th centuries the knowledge of Greek in the West died out with shocking suddenness; Augustine had only a rudimentary knowledge of the Greek language, and translators such as Jerome (*c.* 347–419/420) and Rufinus (*c.* 345–410/411) were scarce indeed. The few Greek studies were undertaken for the sake of theology or philosophy, and translation of secular authors was rare; Calcidius' (Chalcidius') 4th-century version of the *Timaeus* was for eight centuries the only Latin translation of a Platonic dialogue, Boethius' plan for a series of translations of Plato and Aristotle being interrupted by his execution. Sicily remained Byzantine until the Arab conquest of the 9th century, and Calabria, Lucania, and Apulia (Puglia) until the Norman conquest of the 11th century. The Normans and later the Hohenstaufen rulers favoured Greek studies. In the 12th century Greek, too, benefited from the intellectual revival; Henricus Aristippus, archdeacon of Catania, translated Plato's *Meno* and *Phaedo,* and the admiral Eugenius collaborated in a Latin version of the *Almagest,* an encyclopaedia compiled by the astronomer Ptolemy of Alexandria in the 2nd century AD. Also during the 12th century two Italian scholars, James of Venice and Burgundio of Pisa, traveled to Constantinople in search of theological and philosophical learning; Burgundio brought back literary as well as theo-

Aristotelian revival

logical manuscripts, though he was probably incapable of reading them. The Aristotelian revival of the 13th century led to the production of many translations of Aristotle by William of Moerbeke in Rome, and in England Aristotle was read in the original by Robert Grosseteste and Roger Bacon. During the 14th century contact between Rome and Constantinople was continued; Petrarch (see below *Latin scholarship*) acquired a Byzantine manuscript of Homer, though he never made the effort to enable himself to read it, and later in the century another such manuscript was in the hands of the humanists of Padua. In about 1397 the Byzantine scholar Manuel Chrysoloras went to Italy to teach Greek in Florence. At the Council of Ferrara-Florence in 1438–45 the union of the churches was agreed upon, but it was later repudiated. George Gemistus Plethon (*c.* 1355–1450/52), the famous Neoplatonist of Mistra, was present at that council; with him was his pupil John Bessarion of Trebizond (1403–72), who continued to support church union as an individual, so that when the repudiation took place he converted to the Western church. He stayed behind in Italy, became a cardinal, and made an important gift of books to Venice. Early in the 15th century Italians such as Francesco Filelfo and Giovanni Aurispa were bringing back Greek manuscripts from Constantinople in large quantities, so that well before the capture of Constantinople by the Turks in 1453 many Greek books had found their way to the West.

Latin scholarship. *Republic and early empire.* From the beginning, Roman scholarship imitated Greek: Hellenistic techniques were applied to the treatment of Latin texts, and Latin grammar adopted Greek categories and terminology. Learned Greeks such as Tyrannion, Alexander Polyhistor, and Parthenius were brought to Rome as prisoners in the Mithradatic Wars. Even before that, as early as about 100 BC, the Roman knight Lucius Aelius Stilo Praeconinus had been teaching and writing about Latin grammar. Marcus Terentius Varro (116–27 BC) by his vast learning and prodigious output influenced almost every branch of scholarship; of his 25 books about the Latin language, books V to X survive in nearly complete form. In scholarship as in other matters the early imperial period was one of great achievement. It was the age of commentators such as Gaius Julius Hyginus, who was in charge of the Palatine Library in Rome founded by Augustus; of editors such as Marcus Valerius Probus (*c.* AD 20–105), who made critical editions of Plautus, Terence, Lucretius, Virgil, and Horace; of grammarians such as Verrius Flaccus, the author of a vast work on the meaning of words; of the elder Pliny (AD 23/24–79), whose encyclopaedic *Historia naturalis* (*Natural History*) was a major sourcebook during the Middle Ages; of Gaius Suetonius Tranquillus (*c.* AD 69–after 122), who wrote the lives of poets and grammarians as well as of emperors; and of Aulus Gellius, whose miscellany called *Noctes Atticae* preserved much ancient learning.

Later empire. The barbarian invasions of the 3rd century marked the beginning of a testing time for Latin as well as for Greek scholarship, and the scholars of the 4th and 5th centuries—such as Aelius Donatus, the grammarian and teacher of rhetoric; Servius, the learned commentator on Virgil; Priscian, the Greek author of the most famous Latin grammar of antiquity; and Macrobius, who during the first half of the 5th century wrote the learned miscellany called *Saturnalia*—were epitomizers and compilers living on inherited capital. In the Western Empire the knowledge of Greek was practically extinct, and the earlier literature of Rome itself was threatened with extinction. The classics were still the staple of such education as there was, but the dominance of rhetoric favoured only certain authors; classical poets such as Virgil, Horace, Ovid, and Terence were protected by critics and commentators, but among earlier authors Ennius and Lucilius disappeared and Plautus narrowly escaped. The book, in the form of the vellum or parchment codex, was superseding the papyrus roll, and authors who were not recopied were doomed to oblivion.

The dominance of rhetoric

The early Middle Ages. The period during which the Merovingian dynasty founded by Clovis (*c.* 466–511) was in power was a dark age for learning, but there was no complete breach with the past. Under the influence of the church the barbarian invaders wished to base their civilization on the Latin model, and since it was the language of the church, Latin continued to be the language of literature. Although interest in antiquity for its own sake had little part in the late imperial and early medieval ideal, under the protection of the church learning survived in the medieval schools, and classical texts provided a grounding in grammar, a training in logical thought, and a philosophical premise for theology. Flavius Cassiodorus, a retired statesman who founded a monastery at Vivarium, in southern Italy, sometime after AD 540, encouraged his monks to copy pagan as well as Christian authors, a practice that spread later to other monasteries, particularly those of the Benedictine order. About 563 the Irish missionary Columba (*c.* 521–597) founded a church and monastery on the island of Iona in the Inner Hebrides of Scotland, and soon afterward Irish missionaries converted the whole of Scotland and established monasteries in the north of England. Later Irish missions led by Columban (*c.* 543–615) founded Luxovium (Luxeuil) in the Vosges Mountains of Gaul (590), Bobbio on the Trebbia (*c.* 612–614), and St. Gall in Switzerland; Corbie near Amiens was founded from Luxovium a century later, and these monasteries played a leading role in the preservation of ancient literature. In England, Aldhelm (*c.* 639–709) and Bede (672/673–735) were men of considerable learning. At this time learning was also alive in Visigothic Spain, as is shown by the vast encyclopaedia of Isidore of Seville (*c.* 560–636), which was of great importance for the remainder of the Middle Ages. During the late 7th and 8th centuries the successors of Columban converted first the Frisians and then much of Germany; they founded the important monasteries of Fulda (744), Lorsch, in Hesse (764), and Hersfeld (*c.* 770), while Reichenau, on Lake Constance (724), and nearby Murbach (727) were founded by a refugee from Visigothic Spain.

Christianization of the British Isles

The Carolingian Renaissance. Pepin III the Short (reigned 751–768) began ecclesiastical reforms that Charlemagne continued, and these led to revived interest in classical literature. Charlemagne appointed as head of the cathedral school at Aachen the distinguished scholar and poet Alcuin of York, who had a powerful influence on education in the empire. Many ancient texts were now copied into the new Carolingian minuscule, and the palace library allowed its books to be copied for other libraries, so that learning was rapidly diffused. Latin poetry of some merit was composed at and about the imperial court, and Einhard's life of Charlemagne (probably written *c.* 830–833) is modeled on the biographies of Suetonius. Learned work was resumed, and the historian Paul the Deacon (Paulus Diaconus) abridged the abridgement of the lexicon of Verrius Flaccus that had been made by Festus during the 2nd century AD. The nearest approach in the Middle Ages to a humanistic scholar was Servatus Lupus, abbot of Ferrières (*c.* 805–862), who collected, copied, and excerpted ancient manuscripts on a large scale. Despite the splitting up of the Carolingian Empire in 843 and the troubles resulting from the barbarian attacks on Europe of the 9th and 10th centuries, the educational apparatus created by the so-called Carolingian Renaissance provided enough momentum to keep the classical tradition going until a new impulse arrived to carry it on to fresh developments.

Carolingian minuscule

12th-century renaissance

The later Middle Ages. A renewed period of intellectual activity in the ancient Benedictine foundation of Monte Cassino heralded the renaissance of the 12th century. Dante Alighieri (1265–1321) was familiar not only with Virgil but also with Lucan, Statius, and Ovid, and *The Divine Comedy*'s picture of the cosmos is deeply indebted to Aristotle's *On the Heavens.* William of Malmesbury (died *c.* 1143) and John of Salisbury (1115/20–1180) were considerable Latin scholars. During the 13th century a group of scholars in Padua around Lovato Lovati (1241–1309) and Albertino Mussato (1261–1329) were active humanists. Lovato read Lucretius and Catullus, studied Seneca's tragedies in the famous *Codex Etruscus,* and found and read some of the lost books of Livy. Both men wrote Latin poetry, Mussato composing a Senecan tragedy, the *Ecerinis,* designed to open the eyes of the Paduans to the danger presented by Cangrande della Scala, the tyrant of Verona, by describing the tyrannical conduct of their own former despot, Ezzelino III.

THE REVIVAL OF LEARNING

Renaissance humanism. The humanist movement was consolidated by the generation of Petrarch (Francesco Petrarca; 1304–74). Petrarch actively looked for manuscripts, building up what was for his day a remarkable library, and taught himself to write an elegant classicizing Latin very different from what had been customary during the Middle Ages. Like Politian later, he was a great poet in Italian; but he valued far more than his vernacular poetry his Latin epic *Africa,* a skillful imitation of the Roman poets. Like almost everyone before Politian, Petrarch knew little or no Greek (on the manuscript of Homer that he possessed, see above, *Greek in the West*). Giovanni Boccaccio (1313–75) also looked actively for ancient manuscripts and actively forwarded the aims of humanism.

The age of Petrarch

The revival of classical learning that Petrarch and Boccaccio promoted was only one aspect of the complex phenomenon of the Renaissance. In origin the movement was utilitarian, seeking to exploit classical antiquity in the service of modern man; the early Italian humanists were not scholars so much as *litterati* and educators, and it is a mistake to think that they were pagans. The earlier idea that the invention of printing was an effective agent in the revival is also erroneous; for by 1470, when the first editions of the Latin classics were quickly coming off the presses, the Renaissance was already well past its early stages. Thus, although Greek teachers and Greek manuscripts had long before begun to enter Italy, the advanced study of Greek, apart from the activities of an isolated genius like Politian, made little headway before the 16th century. The early humanists saw that the manuscripts they discovered contained many corruptions and enjoyed trying to emend them, but many of their conjectures were frivolous, and they often omitted to mark them as conjectures, a practice that irritated later scholars.

Petrarch's successor as the leader of the humanist movement was Coluccio Salutati (1331–1406), chancellor of Florence, who acquired many manuscripts and built up a splendid library; it was he who invited Chrysoloras to Florence. A later chancellor, Leonardo Bruni (*c.* 1370–1444), translated into Latin Plutarch, Xenophon, six dialogues of Plato, and Aristotle's *Ethics* and *Politics.* Poggio Bracciolini (1380–1459) was the most active and the most successful hunter of manuscripts, traveling to France, Germany, and even England in pursuit of them. The same period saw the beginning of the study of the ancient monuments of Italy and the collection of coins and inscriptions as well as works of art by scholars such as Flavio Biondo (Flavius Blondus; 1392–1463) and later Pomponius Laetus (1428–97). Cyriacus of Ancona (1391–1452) broke new ground by traveling to the countries of the Turkish Empire, where he drew monuments and copied inscriptions, thus providing the only record of many objects that were later lost.

Beginnings of modern scholarship. What may be called professional standards of scholarship are seen first in the work of Lorenzo Valla (1407–57) and Politian (Angelo Poliziano; 1454–94). Valla in his *Elegantiae* demonstrated the technique of pure and elegant classical Latin, free of medieval awkwardness; when Pope Nicholas V ordered the chief Greek prose writers to be translated into Latin, Valla was responsible for Thucydides. He also translated part of the *Iliad* into Latin prose. In his philosophical works, which include treatises on pleasure and on free will, he was the first modern to throw light on Epicurus. Gifted with the historical sense of the true critic, Valla perceived the spuriousness of several famous documents: a treatise forged in the name of Dionysius the Areopagite, the New Testament convert of St. Paul, by a later writer now called Pseudo-Dionysius; a collection of letters supposedly exchanged by St. Paul and the 1st-century-AD Roman philosopher Seneca; and the so-called Donation of Constantine, by which the emperor Constantine the Great was alleged to have granted to the papacy spiritual and temporal dominion over Rome and the West.

Politian

Politian, like Petrarch a great poet in the vernacular, began studying Greek at the age of 10 and attained a better knowledge of it than any modern to that date; in his collection of notes called *Miscellanea,* the second volume of which was unfortunately lost and was published only in 1972, he threw light on a variety of ancient writers, including even Greek poets of the Hellenistic Age.

By 1500 most of the chief Latin authors were in print. In that year Aldus Manutius (1449–1515) founded in Venice his "Neacademia" (or Aldine Academy), dedicated to, among other things, the issuing of large and relatively cheap editions of ancient authors. Working in conjunction with the learned Cretan Marcus Musurus (1470–1517), he brought out in 21 years 27 editiones principes (first editions) of Greek authors, including five in the year 1502 alone. During the century that followed, the book evolved from what was essentially an expensive facsimile of a medieval manuscript into a working tool for scholars. Other printers, such as the Giunta family in Florence, followed Aldus' example, and Zacharias Callierges in Rome brought out the first printed texts of Pindar, Callimachus, and the Homeric scholia. Aldus' son Paulus Manutius (1512–74) carried on his father's business and did much for the texts of Cicero. Petrus Victorius (1499–1585) was the leading Italian scholar of his time, editing Aeschylus and Euripides and writing commentaries on Aristotle's *Rhetoric, Poetics, Politics,* and *Nichomachean Ethics,* as well as editing other Greek texts and doing important work on Cicero; he concentrated on producing careful editions of the best manuscripts available, in a reaction against the excessive emendation of earlier scholars. Francesco Robortello (1516–67) also did important work on Aeschylus and Aristotle's *Poetics.* Fulvius Ursinus (1529–1600) built up the Farnese library in Rome, edited the Greek lyric poets, and made important contributions to numismatics and iconography. Carolus Sigonius (1523–84) and Pirro Ligorio (*c.* 1510–83) were active in the field of history and antiquities, Ligorio producing much genuine material besides his notorious forgeries. But after the 16th century, the atmosphere of the Counter-Reformation was not favourable to disinterested inquiry, and Italian scholarship declined. The Jesuits in their educational activities made use of the forms of humanism while abolishing its content.

Advancement of the book

The Renaissance outside Italy. In Spain the Renaissance had made a promising beginning; Antonio of Nebrija (1444–1522) anticipated Erasmus in showing that the Greek language had been pronounced by the ancients differently from the modern Greeks, and later Antonio Agostino, archbishop of Tarragona (1517–86), did important work on ancient law and numismatics. But the Spanish Renaissance was frozen by the Counter-Reformation.

During the late 15th and early 16th centuries the new learning began to establish itself north of the Alps. William Grocyn, who had studied in Italy, was probably the first man to teach Greek in an English university; he was friendly with John Colet and Thomas More, both of England, and later with the Dutch humanist Desiderius Erasmus. Thomas Linacre, later an eminent physician, studied Greek in Italy under Politian; on his return to England he gave lectures at which More was present.

The Northern Renaissance

Erasmus (*c.* 1466–1536), the first editor of the New Testament, was more concerned with biblical and patristic studies than with the Greek and Latin classics for their own sake. Yet his *philosophia Christi,* an attempt to mediate between ancient wisdom and Christian faith, was closely linked with classical scholarship, and he found time to produce numerous editions and translations of Greek and Latin authors, besides making such contributions to scholarship as his famous collection of proverbs, the *Adagia.* The *Utopia* of his English friend Thomas More was profoundly influenced by Platonism. Erasmus' pupil Beatus Rhenanus was one of a group of German scholars who brought out important editions of Latin texts. Philipp Melanchthon (1497–1560) actively promoted scholarship in Germany; his associate Joachim Camerarius (1500–74) did much for Plautus, as did Hieronymus Wolf (1516–80) for the Attic orators.

Erasmus formed a close connection with the great printer of Basel, Johannes Froben. Amerbach, Cratander, and Hervagius were other notable printers of that city, active in the production of critical editions of ancient texts.

Obliged as they were to concede primacy in Latin studies to the Italians, the French during the 16th century took the lead in Greek, although Denis Lambin (Lambinus; 1516–72) did valuable work on Cicero, Lucretius, and Horace. Guillaume Budé (Budaeus; 1467–1540) laid the foundations in Greek studies, and Jean Dorat (Auratus; 1508–88) and Adrien de Tournebu (Turnebus; 1512–65), pioneers in the study of Greek poetry, inspired the contemporary poets Ronsard and du Bellay, the leaders of the Pléiade group, with admiration for Greek literature. The great printer Robert Estienne (Stephanus; 1503–59) produced the first critical edition of the Greek New Testament (1551), reprinting Erasmus' text but adding variants from 15 manuscripts. Estienne's son, Henri, published many editiones principes of Greek authors and a Greek *Thesaurus* (1572) not superseded until the early 19th century.

Scaliger and Casaubon

Two French scholars—Joseph Justus Scaliger (1540–1609) and Isaac Casaubon (1559–1614)—deserve particular mention. Like Erasmus, Scaliger saw that classical learning should be a unity. His diversity was that of the explorer, not the dilettante; each edition opened up a new path: that of Festus (1575) to Old Latin, that of Manilius' *Astronomica* (1579) to ancient astronomy, for example. He assisted Janus Gruterus (1560–1627) by compiling the indexes to his famous *Inscriptiones antiquae totius orbis Romani* and encouraged the collection of the fragments of classical literature. But his greatest achievement was to bring order into the chaos of ancient chronology in his *De emendatione temporum* (1583) and *Thesaurus temporum* (1606).

Casaubon, too, perceived that antiquity must be studied as a whole and also (and this too Erasmus understood) that the study must begin from Greek. Through his series of detailed commentaries on difficult and prolific authors (Strabo, Athenaeus, Polybius), he was instrumental in turning scholarship—hitherto an art—into a science.

Henri Estienne, Scaliger, and Casaubon were all Huguenots, and all died in exile—Estienne in Lyon, Scaliger in Leiden, and Casaubon in London. Another eminent Huguenot scholar of the time, Marcus Antonius Muretus (Marc-Antoine de Muret; 1526–85), the most elegant writer of Ciceronian Latin since Cicero, who defended the practice of emendation against the cautious Victorius, left France when accused of homosexuality, became a Catholic, and enjoyed great success in Rome.

Development of paleography

Scholarship in the 17th century. After the conversion of Henri IV to Roman Catholicism French scholarship declined, as Italian scholarship declined during the age of the Counter-Reformation. But the action of the Jesuits in challenging the authenticity on which the privileges of the Benedictines depended caused the latter to turn to the study of paleography in order to defend themselves, thus occasioning the chief contribution of France to classical studies during the 17th century. Jean Mabillon (1632–1707) established Latin paleography as a modern science, and another inmate of the monastery of St. Germain-des-Prés, Bernard de Montfaucon (1655–1741), did the same for Greek paleography. This kind of work was continued by the great antiquarians of the following century, notably L.A. Muratori (1672–1750) and Scipione Maffei (1675–1755).

As scholarship declined in France (where the series of Delphin Classics supervised by Pierre-Daniel Huet from 1670 to 1680 marks the summit of strictly classical achievement), so it rose and flourished in the Netherlands. Christophe Plantin had founded his great press in Antwerp in 1550 and the Elzevirs theirs in Leiden in 1580 and later in Amsterdam. Scaliger ended his days in the newly founded State University of Leiden. Justus Lipsius (1547–1606) produced important editions of Tacitus and Seneca, at the same time promoting a new Christian Stoicism. The great jurist Hugo Grotius (1583–1645), in many ways the true successor of Erasmus, did brilliant work in classical studies as well as in many other fields. Nicolas Heinsius (1620–81) produced editions, based on extensive study of manuscripts, that earned him the title "saviour of the Latin poets." His counterpart in prose, John Gronovius (1611–71), produced editions of Livy, Seneca, Pliny, and others. The letters of Heinsius and Gronovius testify to as ample a conception of classical studies as that of Scaliger and Casaubon.

Impact of the Thirty Years' War

The Thirty Years' War (1618–48) had a disastrous effect on scholarship. Among the classicists in Holland and Germany, stagnation set in; unwieldy, uncritical variorum editions became the fashion, and the collection of "antiquities," divorced from linguistic study and from critical scholarship, degenerated into the mere piling up of information.

The 18th century: the age of Bentley. Since the late 16th century little had been heard of English scholarship; once the study of Greek had been established by Linacre, Grocyn, Sir John Cheke, and their contemporaries, the English preoccupation with education had set in. John Selden is the most notable of few exceptions, and he was a jurist and antiquary, not an academic, though his *De Diis Syris* (1617) laid the foundations of Eastern scholarship. A new era began with the *Epistola ad Joannum Millium* (1691) of Richard Bentley (1662–1742). This collection of brilliant miscellaneous observations, prompted by the editio princeps of the 6th-century Byzantine chronicle of John Malalas, displayed already the comprehensive learning and rare power of divination that were to enable Bentley to lay the foundations of the critical scholarship of the coming age. Although his achievements in textual criticism were singularly brilliant, Bentley must not be thought of as a mere editor of texts but as the creator of a critical method that was to be applied with powerful effect in every department of antiquity. This is in evidence above all in his *Dissertation upon the Epistles of Phalaris* (expanded edition, 1699), the first important work of classical scholarship written in a modern language. His editions of Horace (1711), Terence (1726), and Manilius (1739) were all of masterly quality. He did remarkable work in collecting fragments of Menander and Callimachus, and although he never completed his proposed editions of Homer and the New Testament, the preparatory work he did toward them had a revolutionary effect in both fields of study.

After Bentley's death the only part of his inheritance taken up by his countrymen was his work in textual criticism. The work of his English contemporaries in this field, who include such important scholars as Jeremiah Markland (1693–1776), Thomas Tyrwhitt (1730–86), Benjamin Heath (1704–66), and Samuel Musgrave (1732–80), was carried further by the next generation. Richard Porson (1759–1808), Peter Elmsley (1773–1825), and P.P. Dobree (1782–1825) all concentrated upon Attic drama, Porson showing a particularly fine feeling for Greek.

Edward Gibbon

In 1786 Sir William Jones (1746–94) began the study of Sanskrit that was to lead to the establishment of the new discipline of comparative philology. Edward Gibbon (1737–94), essentially self-educated despite his early residence at Magdalen College, Oxford, made with *The History of the Decline and Fall of the Roman Empire* (1776–88) the greatest single contribution to the study of ancient history in the whole 18th century. The *Essay on the Original Genius of Homer* by Robert Wood (*c.* 1717–71), printed privately in 1767 and published posthumously in 1775, not only marked a new stage in Homeric studies but also assisted the movement toward exploration of ancient sites in Greece. Exploration was powerfully promoted by the publications in London of the Society of Dilettanti, especially the drawings in *The Antiquities of Athens* (four volumes, 1762–1808), by James Stuart and Nicholas Revett.

Meanwhile in the Netherlands, where Bentley's greatness had at once been recognized, a distinguished series of scholars—Tiberius Hemsterhuys (1685–1766), L.K. Valckenaer (1715–85), the German emigrant David Ruhnken (1723–98), and, later, Daniel Wyttenbach (1746–1820)—continued to do valuable work on Greek texts, including the difficult but rewarding remains of ancient lexicography. Bibliographical works and dictionaries were now improved; Johann Albert Fabricius (1668–1736) put the bibliography of Greek and then Latin literature on a new footing, and Egidio Forcellini in Padua superseded the

Latin thesaurus of Robert Estienne. The study of ancient coins was greatly advanced by the work of the Swiss-born scholar Ezechiel Spanheim (1629–1710) and the Austrian scholar J.H. Eckhel (1737–98).

In archaeology the 18th century saw the beginning of the excavation of Herculaneum and Pompeii and of exploration of the remains of the Etruscan civilization. Historical source criticism (*Quellenkritik*) began in the work of the German historian Barthold Niebuhr (1776–1831). But technical progress was not enough. A new spirit was needed to arouse classical studies to take their place in the modern world, and it came from Germany.

MODERN CLASSICAL SCHOLARSHIP

The new German humanism. The "new humanism" that transformed German intellectual life in the late 18th century was a complex phenomenon, acting through scholarship, education, philosophy, and literature. Educationally the University of Göttingen played a leading part: there J.M. Gesner (1691–1761) and C.G. Heyne (1729–1812) introduced a new approach—an attempt to enter into the spirit of the past as displayed in its artistic monuments as well as in its literature. J.J. Winckelmann (1717–68) was the first to mark out the successive periods into which the history of Greek art falls. He was also the first to isolate and describe the essentially Hellenic element in Greek art and to relate the development of art in antiquity to other aspects of culture. He demonstrated that a large number of vases then known as Etruscan because they had been found in Etruscan cemeteries were in fact Greek, although the original error was to be perpetuated by Josiah Wedgwood, who named his pottery works "Etruria" in 1769. Winckelmann's influence ranged over the literary as well as the academic world, powerfully affecting such figures as Lessing, Herder, and Goethe, who named the memorial essay that he published in 1805 *Winckelmann und sein Jahrhundert* ("Winckelmann and His Century"). Goethe (1749–1832) made a systematic effort to know and understand Greek art and literature, particularly Greek sculpture and the poetry of Homer, and to utilize them for his own purposes; it may be doubted whether anyone since ancient times had understood the Greeks so well. He was a friend of Wilhelm von Humboldt (1767–1835), the great statesman and pioneer of the study of language who founded Frederick William University (later the University of Berlin), which rapidly became the leading university of Europe.

The achievement of Goethe

Goethe was also in touch with F.A. Wolf (1759–1824), a Göttingen pupil of Heyne. Wolf defined the "science of antiquity" (*Altertumswissenschaft*) and mapped out its constituent provinces and principles. Influenced by Herder, with his special interest in the early literatures of various peoples and their special characteristics, Wolf in his *Prolegomena ad Homerum* (1795) raised such questions about Homer as to give rise to a debate that has continued ever since. Goethe was at first carried away by Wolf's theory that the *Iliad* and *Odyssey* were composed orally by a number of authors and that the artistic unity of the poems was a later imposition, but he eventually returned to his belief in an individual Homer, as scholars have done increasingly in the 20th century.

Another great classical scholar in close touch with Goethe was Gottfried Hermann (1772–1848), who continued the tradition of 18th-century rationalism, applying to the study of ancient poetry a critical method based on a strict Kantian logic. Hermann did much to advance the study of Homer, Pindar, late epic poetry, and Greek metre. By his editions of Sophocles, Aeschylus, and part of Euripides he effected a great and permanent improvement in the texts of these poets.

Hermann had many distinguished pupils, including C.A. Lobeck (1781–1860), a grammarian of great learning and acuteness, who in his famous book *Aglaophamus* (1829) refuted the seductive but dubious theory of the Heidelberg professor G.F. Creuzer that the mythology of Homer and Hesiod contained symbolic elements of an ancient Oriental revelation from which it was ultimately derived. August Meineke (1790–1870) did important work on Hellenistic poetry and produced an excellent edition of the fragments of Greek comedy. August Immanuel Bekker (1785–1871), a pupil of Wolf, took advantage of the accumulation in Paris of many previously inaccessible manuscripts from various countries following the Napoleonic conquests to make a valuable contribution to the texts of many prose authors. Wilhelm Dindorf (1802–83) edited many texts, including the scholia on Rome and on Demosthenes. With his brother Ludwig and K.B. Hase, he revised the great Greek *Thesaurus* of Henri Estienne. Hermann's son-in-law Moritz Haupt (1808–74) did important work on Latin poetry. H.L. Ahrens (1809–81) wrote on the Greek dialects and on the bucolic poets. August Nauck (1822–92), who taught in St. Petersburg, made a notable contribution to the establishment of the texts of Greek tragedy.

The school of Hermann with its strong emphasis on linguistic study came occasionally into conflict with the representatives of a newer trend in the approach to antiquity. In Berlin August Boeckh (1785–1867) did important work on Greek poetry, particularly Pindar, but also established on a firm footing the study of Greek private and public economy and the systematic collection of Greek inscriptions. K.O. Müller (1797–1840), the author of an important history of Greek literature, which first appeared in English, was a pioneer of the study of Greek, Roman, and Etruscan origins and of the mythology of the different parts of Greece, which he believed could shed much light on early history. He was a strong upholder of the importance of art and archaeology in the study of antiquity, as was F.G. Welcker (1784–1868), who applied deep knowledge of Greek art and religion to the interpretation of literature and did much to shape the wider conception of the study of antiquity that was now coming to maturity.

Comparative linguistics

The comparative study of Indo-European languages that was initiated by Franz Bopp (1791–1867), one of the famous scholars who gave the University of Berlin its enviable reputation, profoundly influenced the study of the ancient as well as other languages. One field in which this was seen was the study of early Latin, which was now placed on a new basis by Friedrich Ritschl (1806–76), who applied knowledge gained from the study of inscriptions to the elucidation of early Latin texts. There followed much important work on early Latin, such as that of Johannes Vahlen (1830–1911) on Ennius and that of Otto Ribbeck (1827–98) on Roman tragedy.

The rise of textual criticism. Knowledge of ancient literature must always rest on the standards of editing and criticism of Greek and Latin texts that have come down in a corrupt and sometimes mutilated state. Early in the 19th century great advances were made in this field of classical studies. Angelo Cardinal Mai (1782–1854) published hitherto unknown Greek and Latin texts, including much of Cicero's *De republica,* from newly discovered palimpsests. A.I. Bekker, as well as editing many unknown Greek texts in the Paris collections (see above), was able, by use of newly discovered earlier and better manuscripts, to produce better editions of the classical authors than those then current. But the formulation of a technique of systematic recension (*i.e.,* analysis and evaluation of a manuscript tradition) was gradual, with its roots in the 18th century. Such New Testament scholars as J.A. Bengel (1687–1752) had established the principle that the witnesses to a text must be classified and their testimony evaluated according to their textual genealogy. For a time, the perceived barrier between "sacred" and "profane" texts limited the influence of such work on the analysis of "pagan" sources. During the first half of the 19th century the combined efforts of several scholars, notably Ritschl, Jakob Bernays (1824–81), and the Danish Latinist J.N. Madvig (1804–86), evolved the critical method usually associated with the name of Karl Lachmann (1793–1851) because it is most strikingly exemplified in his edition of Lucretius (1850).

The respect felt for Lachmann by such men as his friend and pupil Moritz Haupt turned into something like critical orthodoxy, however: the new techniques were rigorously applied by less gifted scholars, so that in this department of scholarship some work came to be distinguished by a blind confidence in the so-called scientific method as needing little intelligence in its handling. Madvig had realized the importance, restressed in the mid-20th century, of allowing for inequalities and anomalies in an author's style; but these warnings were lost on those who, exuberantly confident in their own powers, proceeded to wholesale athetesis, or rejection of works as spurious, based on inconsistencies within a text. It was a similarly rigid insistence on analogical methods of criticism that marred the achievements of even such a great critic as the Dutch C.G. Cobet (1813–89) and so set a bad example to lesser scholars.

Developments in the study of ancient history and philosophy. Corresponding progress was made in the field of ancient history. Berthold Niebuhr, the pioneer in historical source criticism, applied a rational skepticism to ancient legends and traditions; he also promoted the collection of Latin inscriptions. J.G. Droysen (1808–84) wrote notable histories of Alexander the Great and of the Hellenistic Age; in fact, the very concept of a Hellenistic Age was his invention. Theodor Mommsen (1817–1903), starting as a professor of Roman law, made vast contributions to almost every branch of Roman studies, but particularly to the history of law and government and of the administration of the Roman provinces. He took a central part in the systematic collection of Latin inscriptions and was familiar with virtually every text and document relevant to Roman history. Mommsen's method of studying an entire civilization had influence on historical studies in general far beyond the limits of ancient history; perhaps the most distinguished of his pupils was the great sociologist Max Weber (1864–1920). In the second half of the century, Eduard Meyer (1855–1930), equipped with wide knowledge of Oriental as well as Greek and Latin sources, wrote an important history of the ancient world. Notable histories of Greece were brought out by Georg Busolt (1850–1920) and Karl Julius Beloch (1854–1929).

Greek philosophy

At the beginning of the 19th century Friedrich Schleiermacher (1768–1834), another of the scholars who gave the University of Berlin its special lustre, revitalized the study of Plato. Eduard Zeller (1814–1908) wrote a history of ancient philosophy that has been several times revised and is still useful. Later Hermann Diels (1848–1922) collected the fragments of pre-Socratic philosophers and of the so-called doxographers who preserved much of the evidence for our knowledge of ancient philosophy. The texts relevant to Epicureanism were edited by Hermann Usener (1834–1905), who employed the new methodology of comparative religion to throw much light on the religion of Greece, not disdaining the study of popular culture and of folklore as well; his work was continued by a line of pupils, and he had an important influence on the great art historian Aby Warburg (1866–1929). Late in the century Erwin Rohde (1845–98) wrote *Psyche,* an important study of Greek beliefs about the soul.

Nations other than Germany made more modest contributions to scholarship, most of them being more concerned with teaching than with research. (For the Italian contribution to papyrology, see below.) The literary scholarship of the French, though elegant and polished, was superficial in comparison with that of the Germans. It is significant that the leading Hellenist of France was the German-born Henri Weil (1818–1909). Great figures were exceptional; among them were J.A. Letronne (1787–1848), an archaeologist and epigraphist, and Paul-Émile Littré (1801–81), the famous lexicographer and positivist philosopher whose remarkable translation of Hippocrates emanated merely as a side interest of his philosophical vocation.

In England a notable contribution to ancient history and to the study of Plato and Aristotle came from the banker George Grote (1794–1871), who saw antiquity from the viewpoint of modern liberalism and utilitarianism. Later in the century the ancient universities produced a few distinguished scholars: H.A.J. Munro (1819–85) did important work on Catullus and Lucretius; Sir Richard Jebb (1841–1905) wrote a good commentary on the newly discovered Bacchylides and also one on Sophocles, which despite some technical deficiencies is still useful because of the author's rare feeling for Greek; and Ingram Bywater (1840–1914) contributed significantly to the study of Aristotle's *Ethics* and *Poetics* and acquired rare knowledge of the history of scholarship.

In the United States the leading figure was B.L. Gildersleeve (1831–1924), who insisted that American classical scholars should aim at the highest European standards. Germany was rightly taken as a model, and valuable work was done, especially in grammar, syntax, and linguistics, by such scholars as W.W. Goodwin (1831–1912), J.W. White (1849–1917), and H.W. Smyth (1857–1937). But too often it was not the admirable qualities of the best German scholars but the dryness, pedantry, and verbosity of the worst that were reproduced. This led to a reaction that went too far in the opposite direction and so did considerable damage. In archaeology, however, the vast resources of America were applied with ever-increasing effectiveness.

Developments in archaeology and art history. The foundation of the Instituto di Correspondenza Archeologica in Rome in 1829 provided an international centre for archaeological studies in Italy, which now progressed rapidly. Eduard Gerhard (1795–1867) founded the study of Greek vase painting as a scientific discipline; his report on the numerous Greek vases excavated from the Etruscan necropolis of Vulci (1831) was epoch-making. In Bonn, Welcker built up the first large collection of plaster casts of Greek sculpture. Another pioneer of the study of Greek art was his colleague Otto Jahn (1813–69), also an excellent Latinist. After the establishment of the Greek kingdom in 1830 the various European nations set up schools in Athens as they had done in Rome, and excavations on a large scale took place not only in Greece but all over the eastern Mediterranean world.

In archaeology the great impetus came from an amateur, Heinrich Schliemann (1822–90), whom no one can deprive of the credit for having guessed that remarkable finds might be made at Troy, Mycenae, and Tiryns, for having deliberately made a fortune so that he might do so, and for having discovered and promoted the great archaeologist Wilhelm Dörpfeld (1853–1940). In 1900 the ancient city of Knossos on Crete was excavated by the English archaeologist Sir Arthur Evans (1851–1941), which enabled the study of Mycenaean civilization to be supplemented by that of Minoan. The French excavated the two great Apollonian shrines at Delos and Delphi.

Papyri had been found in large numbers in the Epicurean library at Herculaneum discovered during the 18th century, and from 1878, when a roll turned up in Egypt, sporadic finds were made. From about the 1870s systematic excavation led to a steady stream of discoveries, mostly from al-Fayyūm, where the sunshine acts upon the soil in such a way as to preserve papyrus. The Italian Amedeo Peyron (1785–1870) was a pioneer in the new discipline of papyrology, as was Domenico Comparetti (1835–1927), the author of a famous book about the fortune of Virgil's works during the Middle Ages. The eminent Italian legal scholar and paleographer Girolamo Vitelli (1849–1935) became an expert papyrologist and had great personal influence. In Germany important papyri were published under the supervision of Wilhelm Schubart with the help of Wilamowitz-Moellendorff. In 1891 the *Constitution of Athens* by Aristotle and the poems of Herodas were published from a papyrus in the British Museum, and in 1897 they were followed by the poems of Bacchylides from the 5th century BC. In 1898 the Oxford scholars B.P. Grenfell and A.S. Hunt brought out the first volume of the series, still not concluded, that contains the texts of the papyri found by them at Oxyrhynchus. Documentary papyri supply useful evidence for law and government in Roman Egypt, and literary papyri supply a priceless supplement to the knowledge of Greek (and occasionally Latin) literature.

The rise of papyrology

The 19th century saw the beginning of many great enterprises, both individual and collective, that have equipped scholars with invaluable tools: collections of fragments, inscriptions, and works of art; and improved dictionaries, special lexica, handbooks, encyclopaedias, and catalogs of manuscripts. The invention of photography made it possible to produce facsimiles of manuscripts and documents and to distribute better likenesses of monuments and works of art. Many of these projects were sponsored by the various national academies, which were now linked by the Association des Académies, the driving force of which was Mommsen.

The rise of professionalism. Associated with Germany was the movement toward what may be called professionalism during the second half of the 19th century. Though Wolf's example in founding a classical periodical in the vernacular had been followed elsewhere (*e.g.,* the English *Classical Journal,* 1810–29), journals written primarily by professional scholars for professional scholars only began to proliferate after about 1850. Coupled with this proliferation were the increased importance of universities, seminars, and academies (with their published proceedings) and the growing habit of early publication of, for instance, the Ph.D. dissertation, the academic "program," and the technical monograph.

Professional journals

Specialization was accompanied by a rise in technical standards of argument and presentation and a tendency toward the use of learned jargon—a phenomenon particularly noticeable in classical studies because of the contrast with earlier scholarly literature. An allied change was the replacement of Latin by the vernacular as a medium of scholarly intercourse and publication (with traditional exceptions, such as the preface and apparatus of a critical text). Thus, after about 1850 a classical scholar who wished to keep abreast of developments in his subject had to be able to read at the least English, French, German, Italian, and, in some cases, Russian. These changes had more immediate results in continental Europe and the United States; in England their effects were delayed in part by the insularity that characterized English scholarship after Bentley, in part by the concentration of the older universities on teaching, and a consequent distrust by tutors of a strong professoriate and of "pure research."

Late 19th-century developments in German scholarship. Germany made so vast a contribution to 19th-century classical scholarship that it would be impossible to name all of the eminent scholars of the period. But from a time rather earlier than the establishment of the German Empire (1871), signs of decline might be observed; the new methods had begun to harden into orthodoxy, mechanically applied by a mass of inferior practitioners. There was a strong tendency toward excessive emendation and deletion, and the overconscientious accumulation of details led to much dullness. From this situation German scholarship was to make a remarkable, though not complete, recovery, thanks to the generation of Ulrich von Wilamowitz-Moellendorff (1848–1931), who broke down the barriers that had grown up between the divisions of his subject, making important contributions to them all. He was the author of the first commentary on a Greek poem in which the entire apparatus of modern scholarship, encompassing not only literary knowledge but also that of history, art, archaeology, linguistics, and religion, was brought to bear on the elucidation of the work in question; this was his commentary on the *Herakles* of Euripides (1st edition, with a remarkable introduction to Attic tragedy, 1889; 2nd edition, 1895). Wilamowitz-Moellendorff produced many more texts and commentaries, besides important work on Greek history, religion, metre, and the history of scholarship. As a professor in Greifswald, Göttingen, and finally Berlin, he exercised a powerful influence.

Wilamowitz-Moellendorff

At the same time Eduard Schwartz (1851–1940) did much not only for the study of Greek history and literature but also for the history of the Christian Church; Georg Kaibel (1850–1901) advanced the study of Greek drama and of verse inscriptions; and Carl Robert (1850–1922) combined archaeological with literary expertise in remarkable fashion. Friedrich Leo (1851–1914) contributed significantly to Plautine studies and began a history of Latin literature of high quality. Jacob Wackernagel (1853–1938) of Basel and Wilhelm Schulze (1863–1935) used their mastery of comparative linguistics to throw light on Greek and Latin texts. Richard Reitzenstein (1861–1931) was eminent not only in the field of Greek literature and lexicography but also in that of ancient religion. Ludwig Traube (1861–1907) did important work in Latin paleography.

Classical scholarship in the 20th century. World War I dealt a heavy blow to classical studies, as to all humane letters, and the numbers of those studying Greek and Latin were noticeably affected; but scholars showed courage and energy in adapting themselves to new conditions. Wilamowitz-Moellendorff continued to be active, and his last decade saw more abundant and more important publications than any other of his career. His pupils produced much important detailed work: Felix Jacoby (1876–1959) began and carried far a learned edition of the fragments of the Greek historians; Paul Maas (1880–1964) showed rare expertise in Greek metrics, textual criticism, and paleography; Eduard Fraenkel (1888–1970) did valuable work on Plautus' relation to his Greek originals and later devoted to Aeschylus' *Agamemnon* one of the most learned of all commentaries; and Rudolf Pfeiffer (1889–1979) wrote a masterly commentary on Callimachus and an important history of classical scholarship.

Reacting against the classicism of the age of Goethe, scholars of the late 19th century saw the study of antiquity mainly from a historical standpoint: they accumulated masses of detail, which sometimes led to dryness, and tended to think exclusively in terms of concrete fact. Discontent arose with the recognition that an excessive preoccupation with the details of their development can harm the understanding of works of literature and thought. Attempts were made to revive classical scholarship by rescuing it from the domination of historical study. Werner Jaeger (1888–1961), an Aristotelian scholar who succeeded Wilamowitz-Moellendorff in his Berlin chair, attempted, without much success, to achieve this by institutional means. More was accomplished by Karl Reinhardt (1886–1958), who, though a devoted pupil of Wilamowitz-Moellendorff, had been in contact from his youth with the ideas of Nietzsche and of the circle around the poet Stefan George. Combining deep learning with refined sensibility, Reinhardt did important work on pre-Socratic philosophy and on Poseidonius and later on Sophocles, Aeschylus, and Homer.

Even before the start of World War II, National Socialist persecution had gravely damaged scholarship in Germany, the main centre of classical studies. The United States and, to an even greater extent, England benefited from the efflux of scholars from the Continent. Jaeger and two other eminent pupils of Wilamowitz-Moellendorff, Paul Friedländer (1882–1968) and Hermann Fränkel (1888–1977), spent the rest of their lives in the United States. So did the Russian M.I. Rostovtzev (1870–1952), who made a vast contribution to the study of the social and economic history of the ancient world. Thaddeus Zielinski (1859–1944), the Polish scholar who did important work on Ciceronian clausulae (clauses) and other topics, was murdered by the Nazis. Eduard Norden (1868–1941), who studied the formal prose of the ancients and did important work on ancient religion and on Latin literature, died in Switzerland. Jacoby, Maas, Fraenkel, and Pfeiffer, as well as the eminent archaeologist Paul Jacobsthal (1880–1957), settled in England, where Fraenkel in particular taught most effectively, creating links between English and continental scholarship. Pfeiffer, like Kurt von Fritz (1900–85), who spent the war years in America, returned to Germany.

Impact of Nazism

In Italy the school founded by Vitelli continued under the leadership of Giorgio Pasquali (1887–1952), a pupil of Schwartz and Leo, and Gaetano de Sanctis (1870–1957) did important work on ancient history. In Sweden Einar Löfstedt (1880–1955) and his school threw much light on Vulgar Latin and indirectly on Latin in general, and M.P. Nilsson (1874–1967) wrote a learned history of Greek religion.

In France classical studies to some degree slumbered under the conservative establishment, but Antoine Meillet (1866–1936) and others advanced the study of linguistics, and Louis Gernet (1882–1962) founded an important school of scholars who applied the techniques of modern sociology and anthropology to the study of antiquity.

In England A.E. Housman (1859–1936) continued with great distinction the tradition of exclusively textual study, editing Juvenal, Lucan, and most notably Manilius. J.D. Denniston (1887–1949) made a valuable study of the Greek particles. Edgar Lobel (1887–1981) from 1927 edited the literary papyri from Oxyrhynchus with unrivaled expertise. Sir Denys Page (1908–78) edited many Greek poetical texts with great success. Gilbert Murray (1866–1957) was not only a literary scholar but, like Jane Harrison (1850–1928), a pioneer in the use of anthropological and sociological methods in the study of antiquity. F.M. Cornford (1874–1943) shared this interest but went on to contribute significantly to the study of Plato and the pre-Socratics. E.R. Dodds, starting with Neoplatonism, applied psychological as well as anthropological knowledge to the study of early Greek thought, also writing excellent commentaries on Euripides' *Bacchae* and Plato's *Gorgias.* Sir John Beazley (1885–1970), with deep learning and refined sensibility, put the whole study of Greek vase painting on a new basis by applying the method of the 19th-century Italian art critic Giovanni Morelli to the identification of individual painters.

The Homeric Question

The way in which research may (and indeed must) transcend the conventional limits of individual disciplines is exemplified during this period in the history of the Homeric Question: the efforts of scholars in such diverse fields as linguistics, archaeology, Hittite studies, folklore, and comparative oral literature have materially advanced understanding of the poems. The problem was transformed by the proof of an American scholar, Milman Parry (1902–35), that the poems are typical of a poetic tradition that has passed through a long phase of oral transmission.

Linear B script

Excavation continued, despite many political and financial difficulties, and a steady stream of discoveries came from Greece, Italy, and other Mediterranean lands. Perhaps the most exciting new find after World War II was the discovery by the Greek archaeologist Spyridon Marinatos of a Minoan town, with fine and well-preserved frescoes, on the island of Thera. Although large-scale excavations in search of papyri have been discontinued for many years, new papyri have not ceased to be discovered. Since World War II the authors who have benefited most have been Callimachus, Menander, and Stesichorus. In 1952 Michael Ventris showed that the language of the so-called Linear B syllabic script on clay tablets found at Mycenae and other places is Greek, thus throwing light on a far earlier stage of the language than had previously been known.

The history of classical scholarship has continued to be one of activity and progress. The publication of new inscriptions and of new papyri and other manuscripts has yielded important new material, and, considering the limited resources available, the task of presenting the texts of literary works and documents in up-to-date editions has been carried out with considerable success. Lately the Hellenistic and Imperial periods have received greater emphasis and have been given greater credit for their achievements.

But such are the threats presented by social change and utilitarian pressures that heroic efforts will be needed if progress is to continue. In Europe at the beginning of the 20th century many schools gave a good grounding in the ancient languages. This is now no longer the case and, as a result, the years when the memory is at its best for learning new languages are wasted. In the United States, vast reserves not only of money but also of talent and enthusiasm make a large contribution to classical studies, but progress is impeded not only by the failure of the schools to teach the ancient languages but also by the materialism and utilitarianism that are gaining ground both there and in Europe.

BIBLIOGRAPHY. JOHN EDWIN SANDYS, *A History of Classical Scholarship,* 3 vol. (1903–08, reissued 1967), while not a critical study, is useful for factual information. RUDOLF PFEIFFER, *History of Classical Scholarship from the Beginnings to the End of the Hellenistic Age* (1968), is a masterly critical survey, and *History of Classical Scholarship from 1300 to 1850* (1976), contains much valuable material but is uneven and lacks adequate treatment of the important 19th-century period. The best brief survey is U. VON WILAMOWITZ-MOELLENDORFF, *History of Classical Scholarship* (1982; originally published in German, 1921). A history of classical scholarship in antiquity is found in JAMES E.G. ZETZEL, *Latin Textual Criticism in Antiquity* (1981, reprinted 1984). For a discussion of the transmission of Greek and Latin literature, see L.D. REYNOLDS and N.G. WILSON, *Scribes and Scholars,* 2nd rev. ed. (1974); and L.D. REYNOLDS (ed.), *Texts and Transmission: A Survey of the Latin Classics* (1983). N.G. WILSON, *Scholars of Byzantium* (1983), chronicles the history of Byzantine scholarship. ROBERTO WEISS, *Medieval and Humanist Greek* (1977), is a collection of essays detailing the use of the Greek language in the Latin Middle Ages. Weiss also covers a later age in *The Renaissance Discovery of Classical Antiquity* (1969, reissued 1973). ANTHONY GRAFTON, *Joseph Scaliger: A Study in the History of Classical Scholarship,* vol. 1, *Textual Criticism and Exegesis* (1983), presents a biography of the 16th-century French classicist. For English scholarship, see C.O. BRINK, *English Classical Scholarship: Historical Reflections on Bentley, Porson and Housman* (1985); and M.L. CLARKE, *Greek Studies in England, 1700–1830* (1945). On the history of Greek vase painting, see R.M. COOK, *Greek Painted Pottery,* 2nd ed. (1972). On the history of papyrology, see E.G. TURNER, *Greek Papyri: An Introduction* (1968, reissued 1980). E.J. KENNEY, *The Classical Text: Aspects of Editing in the Age of the Printed Book* (1974); and SEBASTIANO TIMPANARO, *La genesi del metodo del Lachmann,* new rev. ed. (1981), treat the development of textual criticism. For a discussion of classical influences in the 19th and 20th centuries, see HUGH LLOYD-JONES, *Blood for the Ghosts* (1983), and *Classical Survivals: The Classics in the Modern World* (1982). In general see the collections of essays by ARNALDO MOMIGLIANO, many in English: vol. 1 appeared as *Contributo alla storia degli studi classici* (1955, reprinted 1979), and the most recent addition appeared as vol. 7, *Settimo contributo alla storia degli studi classici e del mondo antico* (1984).

(H.L.-J.)

The History of Science

On the simplest level, science is knowledge of the world of nature. There are many regularities in nature that mankind has had to recognize for survival since the emergence of *Homo sapiens* as a species. The Sun and the Moon periodically repeat their movements. Some motions, like the daily "motion" of the Sun, are simple to observe; others, like the annual "motion" of the Sun, are far more difficult. Both motions correlate with important terrestrial events. Day and night provide the basic rhythm of human existence; the seasons determine the migration of animals upon which humans depended for millennia for survival. With the invention of agriculture, the seasons became even more crucial, for failure to recognize the proper time for planting could lead to starvation. Science defined simply as knowledge of natural processes is universal among mankind, and it has existed since the dawn of human existence.

The mere recognition of regularities does not exhaust the full meaning of science, however. In the first place, regularities may be simply constructs of the human mind. Humans leap to conclusions; the mind cannot tolerate chaos, so it constructs regularities even when none objectively exists. Thus, for example, one of the astronomical "laws" of the Middle Ages was that the appearance of comets presaged a great upheaval, as the Norman Conquest of Britain followed the comet of 1066. True regularities must be established by detached examination of data. Science, therefore, must employ a certain degree of skepticism to prevent premature generalization.

Regularities, even when expressed mathematically as laws of nature, are not fully satisfactory to everyone. Some insist that genuine understanding demands explanations of the causes of the laws, but it is in the realm of causation that there is the greatest disagreement. Modern quantum mechanics, for example, has given up the quest for causation and today rests only on mathematical description. Modern biology, on the other hand, thrives on causal chains that permit the understanding of physiological and evolutionary processes in terms of the physical activities of entities such as molecules, cells, and organisms. But even if causation and explanation are admitted as necessary, there is little agreement on the kinds of causes that are permissible, or possible, in science. If the history of science is to make any sense whatsoever, it is necessary to deal with the past on its own terms, and the fact is that for most of the history of science natural philosophers appealed to causes that would be summarily rejected by modern scientists. Spiritual and divine forces were accepted as both real and necessary until the end of the 18th century and, in areas such as biology, deep into the 19th century as well.

Certain conventions governed the appeal to God or the gods or to spirits. Gods and spirits, it was held, could not be completely arbitrary in their actions; otherwise the proper response would be propitiation, not rational investigation. But since the deity or deities were themselves rational, or bound by rational principles, it was possible for humans to uncover the rational order of the world. Faith in the ultimate rationality of the creator or governor of the world could actually stimulate original scientific work. Kepler's laws, Newton's absolute space, and Einstein's rejection of the probabilistic nature of quantum mechanics were all based on theological, not scientific, assumptions. For sensitive interpreters of phenomena, the ultimate intelligibility of nature has seemed to demand some rational guiding spirit. A notable expression of this idea is Einstein's statement that the wonder is not that mankind comprehends the world, but that the world is comprehensible.

Science, then, is to be considered in this article as knowledge of natural regularities that is subjected to some degree of skeptical rigour and explained by rational causes. One final caution is necessary. Nature is known only through the senses, of which sight, touch, and hearing are the dominant ones, and the human notion of reality is skewed toward the objects of these senses. The invention of such instruments as the telescope, the microscope, and the Geiger counter has brought an ever-increasing range of phenomena within the scope of the senses. Thus, scientific knowledge of the world is only partial, and the progress of science follows the ability of humans to make phenomena perceivable.

This article provides a broad survey of the development of science as a way of studying and understanding the world, from the primitive stage of noting important regularities in nature to the epochal revolution in our notion of what constitutes reality that has occurred in 20th-century physics. More detailed treatments of the histories of specific sciences, including developments of the later 20th century, may be found in the articles BIOLOGICAL SCIENCES; EARTH SCIENCES; and PHYSICAL SCIENCES. See also the references in the *Propædia,* Part Ten, Division III.

The article is divided into the following sections:

Science as natural philosophy

PRECRITICAL SCIENCE

Science, as it has been defined above, made its appearance before writing. It is necessary, therefore, to infer from archaeological remains what was the content of that science. From cave paintings and from apparently regular scratches on bone and reindeer horn, it is known that prehistoric humans were close observers of nature who carefully tracked the seasons and times of the year. About 2500 BC there was a sudden burst of activity that seems to have had clear scientific importance. Great Britain and northwestern Europe contain large stone structures from that era, the most famous of which is Stonehenge on the Salisbury Plain in England, that are remarkable from a scientific point of view. Not only do they reveal technical and social skills of a high order—it was no mean feat to move such enormous blocks of stone considerable distances and place them in position—but the basic conception of Stonehenge and the other megalithic structures also seems to combine religious and astronomical purposes. Their layouts suggest a degree of mathematical sophistication that was first suspected only in the mid-20th century. Stonehenge is a circle, but some of the other megalithic structures are egg-shaped and, apparently, constructed on mathematical principles that require at least practical knowledge of the Pythagorean theorem that the square of the hypotenuse of a right triangle is equal to the sum of the squares of the other two sides. This theorem, or at least the Pythagorean numbers that can be generated by it, seems to have been known throughout Asia, the Middle East, and Neolithic Europe two millennia before the birth of Pythagoras.

This combination of religion and astronomy was fundamental to the early history of science. It is found in Mesopotamia, Egypt, China (although to a much lesser extent than elsewhere), Central America, and India. The spectacle of the heavens, with the clearly discernible order and regularity of most heavenly bodies highlighted by extraordinary events such as comets and novae and the peculiar motions of the planets, obviously was an irresistible intellectual puzzle to early mankind. In its search for order and regularity, the human mind could do no better than to seize upon the heavens as the paradigm of certain knowledge. Astronomy was to remain the queen of the sciences (welded solidly to theology) for the next 4,000 years.

Science, in its mature form, developed only in the West. But it is instructive to survey the protoscience that appeared in other areas, especially in light of the fact that until quite recently this knowledge was often, as in China, far superior to Western science.

China. As has already been noted, astronomy seems everywhere to have been the first science to emerge. Its intimate relation to religion gave it a ritual dimension that then stimulated the growth of mathematics. Chinese savants, for example, early devised a calendar and methods of plotting the positions of stellar constellations. Since changes in the heavens presaged important changes on the Earth (for the Chinese considered the universe to be a vast organism in which all elements were connected), astronomy and astrology were incorporated into the system of government from the very dawn of the Chinese state in the 2nd millennium BC. As the Chinese bureaucracy developed, an accurate calendar became absolutely necessary to the maintenance of order. The result was a system of astronomical observations and records unparalleled elsewhere, thanks to which there are, today, star catalogs and observations of eclipses and novae that go back for millennia.

The Chinese calendar

In other sciences, too, the overriding emphasis was on practicality, for the Chinese, almost alone among ancient peoples, did not fill the cosmos with gods and demons whose arbitrary wills determined events. Order was inherent and, therefore, expected. It was for man to detect and describe this order and to profit from it. Chemistry (or, rather, alchemy), medicine, geology, geography, and technology were all encouraged by the state and flourished. Practical knowledge of a high order permitted the Chinese to deal with practical problems for centuries on a level not attained in the West until the Renaissance.

India. Far less is known about science in India, largely because few scholars have investigated it. It is known that astronomy was studied for calendrical purposes to set the times for both practical and religious tasks. Primary emphasis was placed on solar and lunar motions, the fixed stars serving only as a background against which these luminaries moved. Indian mathematics, on the other hand, seems to have been quite advanced, with particular sophistication in geometrical and algebraic techniques. This latter branch was undoubtedly stimulated by the flexibility of the Indian system of numeration that later was to come into the West as the Hindu-Arabic numerals. Indian thought, however, was primarily philosophical and otherworldly and was concerned more with escaping this world than with understanding it.

America. Quite independently of China, India, and the other civilizations of Europe and Asia, the Maya of Central America, building upon older cultures, created a complex society in which astronomy and astrology played important roles. Determination of the calendar, again, had both practical and religious significance. Solar and lunar eclipses were important, as was the position of the bright planet Venus. No sophisticated mathematics are known to have been associated with this astronomy, but the Mayan calendar was both ingenious and the result of careful observation.

The Middle East. In the cradles of Western civilization in Egypt and Mesopotamia, there were two rather different situations. In Egypt, as in China, there was an assumption of cosmic order guaranteed by a host of benevolent gods. But unlike China, whose rugged geography often produced disastrous floods, earthquakes, and violent storms that destroyed crops, Egypt was surpassingly placid and delightful. Life was, in fact, so pleasant that the major concern of most Egyptians was over leaving it. Egyptians found it difficult to believe that all ended with death; enormous intellectual and physical labour, therefore, was devoted to preserving life after death. Both Egyptian theology and the pyramids are testaments to this preoccupation. Science did not flourish in this atmosphere. All of the important questions were answered by religion, so the Egyptians did not concern themselves overmuch with speculations about the universe. The stars and the planets had astrological significance in that the major heavenly bodies were assumed to "rule" the land when they were in the ascendant (from the succession of these "rules" came the seven-day week, after the five planets and the Sun and the Moon), but astronomy was largely limited to the calendrical calculations necessary to predict the annual life-giving flood of the Nile. None of this required much mathematics, and there was, consequently, little of any importance.

Mesopotamia was more like China. The life of the land depended upon the two great rivers, the Tigris and the Euphrates, as that of China depended upon the Huang Ho (Yellow River) and the Yangtze. The land was harsh and made habitable only by extensive damming and irrigation works. Storms, insects, floods, and invaders made life insecure. To create a stable society required both great technological skill, for the creation of hydraulic works, and the ability to hold off the forces of disruption. These latter were early identified with powerful and arbitrary gods who dominated Mesopotamian theology. The cities of the plain were centred on temples run by a priestly caste whose functions included the planning of major public works, like canals, dams, and irrigation systems, the allocation of the resources of the city to its members, and the averting of a divine wrath that could wipe everything out.

Public works in Mesopotamia

Mathematics and astronomy thrived under these conditions. The number system, probably drawn from the system of weights and coinage, was based on 60 (it was in ancient Mesopotamia that the system of degrees, minutes, and seconds developed) and was adapted to a practical arithmetic. The heavens were the abode of the gods, and because heavenly phenomena were thought to presage terrestrial disasters, they were carefully observed and recorded. Out of these practices grew, first, a highly developed mathematics that went far beyond the requirements of daily business, and then, some centuries later, a descriptive astronomy that was the most sophisticated of the ancient world until the Greeks took it over and perfected it.

Nothing is known of the motives of these early mathematicians for carrying their studies beyond the calculations of volumes of dirt to be removed from canals and the provisions necessary for work parties. It may have been simply intellectual play—the role of playfulness in the history of science should not be underestimated—that led them onward to abstract algebra. There are texts from about 1700 BC that are remarkable for their mathematical suppleness. Babylonian mathematicians knew the

Pythagorean relationship well and used it constantly. They could solve simple quadratic equations and could even solve problems in compound interest involving exponents. From about a millennium later there are texts that utilize these skills to provide a very elaborate mathematical description of astronomical phenomena.

Although China and Mesopotamia provide examples of exact observation and precise description of nature, what is missing is explanation in the scientific mode. The Chinese assumed a cosmic order that was vaguely founded on the balance of opposite forces (yin–yang) and the harmony of the five elements (water, wood, metal, fire, and earth). Why this harmony obtained was not discussed. Similarly, the Egyptians found the world harmonious because the gods willed it so. For Babylonians and other Mesopotamian cultures, order existed only so long as all-powerful and capricious gods supported it. In all these societies, humans could describe nature and use it, but to understand it was the function of religion and magic, not reason. It was the Greeks who first sought to go beyond description and to arrive at reasonable explanations of natural phenomena that did not involve the arbitrary will of the gods. Gods might still play a role, as indeed they did for centuries to come, but even the gods were subject to rational laws.

GREEK SCIENCE

The birth of natural philosophy. There seems to be no good reason why the Hellenes, clustered in isolated city-states in a relatively poor and backward land, should have struck out into intellectual regions that were only dimly perceived, if at all, by the splendid civilizations of the Yangtze, the Tigris and Euphrates, and the Nile valleys. There were many differences between ancient Greece and the other civilizations, but perhaps the most significant was religion. What is striking about Greek religion, in contrast to the religions of Mesopotamia and Egypt, is its puerility. Both of the great river civilizations evolved complex theologies that served to answer most, if not all, of the large questions about mankind's place and destiny. Greek religion did not. It was, in fact, little more than a collection of folk tales, more appropriate to the campfire than to the temple. Perhaps this was the result of the collapse of an earlier Greek civilization, now called Mycenaean, toward the end of the 2nd millennium BC, when a dark age descended upon Greece that lasted for three centuries. All that was preserved were stories of gods and men, passed along by poets, that dimly reflected Mycenaean values and events. Such were the great poems of Homer, the *Iliad* and the *Odyssey,* in which heroes and gods mingled freely with one another. Indeed, they mingled too freely, for the gods appear in these tales as little more than immortal adolescents whose tricks and feats, when compared with the concerns of a Marduk or Jehovah, are infantile. There really was no Greek theology in the sense that theology provides a coherent and profound explanation of the workings of both the cosmos and the human heart. Hence, there were no easy answers to inquiring Greek minds. The result was that ample room was left for a more penetrating and ultimately more satisfying mode of inquiry. Thus were philosophy and its oldest offspring, science, born.

The absence of theology

The first natural philosopher, according to Hellenic tradition, was Thales of Miletus, who flourished in the 6th century BC. We know of him only through later accounts, for nothing he wrote has survived. He is supposed to have predicted a solar eclipse in 585 BC and to have invented the formal study of geometry in his demonstration of the bisecting of a circle by its diameter. Most importantly, he tried to explain all observed natural phenomena in terms of the changes of a single substance, water, which can be seen to exist in solid, liquid, and gaseous states. What for Thales guaranteed the regularity and rationality of the world was the innate divinity in all things that directed them to their divinely appointed ends. From these ideas there emerged two characteristics of classical Greek science. The first was the view of the universe as an ordered structure (the Greek *kósmos* means "order"). The second was the conviction that this order was not that of a mechanical contrivance but that of an organism; all parts of the universe had purposes in the overall scheme of things, and objects moved naturally toward the ends they were fated to serve. This motion toward ends is called teleology and, with but few exceptions, it permeated Greek as well as much later science.

Thales inadvertently made one other fundamental contribution to the development of natural science. By naming a specific substance as the basic element of all matter, Thales opened himself to criticism, which was not long in coming. His own disciple, Anaximander, was quick to argue that water could not be the basic substance. His argument was simple: water, if it is anything, is essentially wet; nothing can be its own contradiction. Hence, if Thales were correct, the opposite of wet could not exist in a substance, and that would preclude all of the dry things that are observed in the world. Therefore, Thales was wrong. Here was the birth of the critical tradition that is fundamental to the advance of science.

Thales' conjectures set off an intellectual explosion, most of which was devoted to increasingly refined criticisms of his doctrine of fundamental matter. Various single substances were proposed and then rejected, ultimately in favour of a multiplicity of elements that could account for such opposite qualities as wet and dry, hot and cold. Two centuries after Thales, most natural philosophers accepted a doctrine of four elements: earth (cold and dry), fire (hot and dry), water (cold and wet), and air (hot and wet). All bodies were made from these four.

Pythagoras

The presence of the elements only guaranteed the presence of their qualities in various proportions. What was not accounted for was the form these elements took, which served to differentiate natural objects from one another. The problem of form was first attacked systematically by the philosopher and cult leader Pythagoras in the 6th century BC. Legend has it that Pythagoras became convinced of the primacy of number when he realized that the musical notes produced by a monochord were in simple ratio to the length of the string. Qualities (tones) were reduced to quantities (numbers in integral ratios). Thus was born mathematical physics, for this discovery provided the essential bridge between the world of physical experience and that of numerical relationships. Number provided the answer to the question of the origin of forms and qualities.

Aristotle and Archimedes. Hellenic science was built upon the foundations laid by Thales and Pythagoras. It reached its zenith in the works of Aristotle and Archimedes. Aristotle represents the first tradition, that of qualitative forms and teleology. He was, himself, a biologist whose observations of marine organisms were unsurpassed until the 19th century. Biology is essentially teleological—the parts of a living organism are understood in terms of what they do in and for the organism—and Aristotle's biological works provided the framework for the science until the time of Charles Darwin. In physics, teleology is not so obvious, and Aristotle had to impose it on the cosmos. From Plato, his teacher, he inherited the theological proposition that the heavenly bodies (stars and planets) are literally divine and, as such, perfect. They could, therefore, move only in perfect, eternal, unchanging motion, which, by Plato's definition, meant perfect circles. The Earth, being obviously not divine, and inert, was at the centre. From the Earth to the sphere of the Moon, all things constantly changed, generating new forms and then decaying back into formlessness. Above the Moon the cosmos consisted of contiguous and concentric crystalline spheres moving on axes set at angles to one another (this accounted for the peculiar motions of the planets) and deriving their motion either from a fifth element that moved naturally in circles or from heavenly souls resident in the celestial bodies. The ultimate cause of all motion was a prime, or unmoved, mover (God) that stood outside the cosmos.

Aristotle was able to make a great deal of sense of observed nature by asking of any object or process: what is the material involved, what is its form and how did it get that form, and, most important of all, what is its purpose? What should be noted is that, for Aristotle, all activity that occurred spontaneously was natural. Hence, the proper means of investigation was observation. Experiment, that is, altering natural conditions in order to throw light on the hidden properties and activities of objects, was unnatural and could not, therefore, be expected to reveal the essence of things. Experiment was thus not essential to Greek science.

The problem of purpose did not arise in the areas in which Archimedes made his most important contributions. He was, first of all, a brilliant mathematician whose work on conic sections and on the area of the circle prepared the way for the later invention of the calculus. It was in mathematical physics, however, that he made his greatest contributions to science. His mathematical demonstration of the law of the lever was as exact as a Euclidean proof in geometry. Similarly, his work on hydrostatics introduced and developed the method whereby physical characteristics, in this case specific gravity, which Archimedes discovered, are given mathematical shape and then manipulated by mathematical methods to yield mathematical conclusions that can be translated back into physical terms.

In one major area the Aristotelian and the Archimedean approaches were forced into a rather inconvenient marriage. Astronomy was the dominant physical science throughout antiquity, but it had never been successfully reduced to a coherent system. The Platonic-Aristotelian astral religion required that planetary orbits be circles. But, particularly after the conquests of Alexander the Great had made the observations and mathematical methods of the Babylonians available to the Greeks, astronomers found it impossible to reconcile theory and observation. Astronomy then split into two parts: one was physical and accepted Aristotelian theory in accounting for heavenly motion; the other ignored causation and concentrated solely on the creation of a mathematical model that could be used for computing planetary positions. Ptolemy, in the 2nd century AD, carried the latter tradition to its highest point in antiquity in his *Hē mathēmatikē syntaxis* ("The Mathematical Collection," better known under its Greek-Arabic title, *Almagest*).

Medicine. The Greeks not only made substantial progress in understanding the cosmos but also went far beyond their predecessors in their knowledge of the human body. Pre-Greek medicine had been almost entirely confined to religion and ritual. Disease was considered the result of divine disfavour and human sin, to be dealt with by spells, prayers, and other propitiatory measures. In the 5th century BC a revolutionary change came about that is associated with the name of Hippocrates. It was Hippocrates and his school who, influenced by the rise of natural philosophy, first insisted that disease was a natural, not a supernatural, phenomenon. Even maladies as striking as epilepsy, whose seizures appeared to be divinely caused, were held to originate in natural causes within the body.

Hippocrates

The height of medical science in antiquity was reached late in the Hellenistic period. Much work was done at the museum of Alexandria, a research institute set up under Greek influence in Egypt in the 3rd century BC to sponsor learning in general. The heart and the vascular system were investigated, as were the nerves and the brain. The organs of the thoracic cavity were described, and attempts were made to discover their functions. It was on these researches, and on his own dissections of apes and pigs, that the last great physician of antiquity, Galen of Pergamum, based his physiology. It was, essentially, a tripartite system in which so-called spirits—natural, vital, and animal—passed respectively through the veins, the arteries, and the nerves to vitalize the body as a whole. Galen's attempts to correlate therapeutics with his physiology were not successful, and so medical practice remained eclectic and a matter of the physician's choice. Usually the optimal choice was that propounded by the Hippocratics, who relied primarily on simple, clean living and the ability of the body to heal itself.

Science in Rome and Christianity. The apogee of Greek science in the works of Archimedes and Euclid coincided with the rise of Roman power in the Mediterranean. The Romans were deeply impressed by Greek art, literature, philosophy, and science, and after their conquest of Greece many Greek intellectuals served as household slaves tutoring noble Roman children. The Romans were a practical people, however, and, while they contemplated the Greek intellectual achievement with awe, they also could not help but ask what good it had done the Greeks. Roman common sense was what kept Rome great; science and philosophy were either ignored or relegated to rather low status. Even such a Hellenophile as the statesman and orator Cicero used Greek thought more to buttress the old Roman ways than as a source of new ideas and viewpoints.

The spirit of independent research was quite foreign to the Roman mind, so scientific innovation ground to a halt. The scientific legacy of Greece was condensed and corrupted into Roman encyclopaedias whose major function was entertainment rather than enlightenment. Typical of this spirit was the 1st-century-AD aristocrat Pliny the Elder, whose *Natural History* was a multivolume collection of myths, odd tales of wondrous creatures, magic, and some science, all mixed together uncritically for the titillation of other aristocrats. Aristotle would have been embarrassed by it.

At its height Rome incorporated a host of peoples with different customs, languages, and religions within its empire. One religious sect that proved more significant than the rest was Christianity. Jesus and his kingdom were not of this world, but his disciples and their followers were. This world could not be ignored, even though concern with worldly things could be dangerous to the soul. So the early Christians approached the worldly wisdom of their time with ambivalence: on the one hand, the rhetoric and the arguments of ancient philosophy were snares and delusions that might mislead the simple and the unwary; on the other hand, the sophisticated and the educated of the empire could not be converted unless the Christian message was presented in the terms and rhetoric of the philosophical schools. Before they knew it, the early Christians were enmeshed in metaphysical arguments, some of which involved physics. What, for example, was the nature of Jesus, in purely physical terms? How was it possible that anybody could have two different essential natures, as was claimed for Jesus? Such questions revealed how important knowledge of the arguments of Greek thinkers on the nature of substance could be to those engaged in founding a new theology.

Science and theology

Ancient learning, then, did not die with the fall of Rome and the occupation of the Western Empire by tribes of Germanic barbarians. To be sure, the lamp of learning burned very feebly, but it did not go out. Monks in monasteries faithfully copied out classics of ancient thought and early Christianity and preserved them for posterity. Monasteries continued to teach the elements of ancient learning, for little beyond the elementary survived in the Latin West. In the East, the Byzantine Empire remained strong, and there the ancient traditions continued. There was little original work done in the millennium following the fall of Rome, but the ancient texts were preserved along with knowledge of the ancient Greek language. This was to be a precious reservoir of learning for the Latin West in later centuries.

SCIENCE IN ISLĀM

The torch of ancient learning passed first to one of the invading groups that helped bring down the Eastern Empire. In the 7th century the Arabs, inspired by their new religion, burst out of the Arabian peninsula and laid the foundations of an Islāmic empire that eventually rivalled that of ancient Rome. To the Arabs, ancient science was a precious treasure. The Qur'ān, the sacred book of Islām, particularly praised medicine as an art close to God. Astronomy and astrology were believed to be one way of glimpsing what God willed for mankind. Contact with Hindu mathematics and the requirements of astronomy stimulated the study of numbers and of geometry. The writings of the Hellenes were, therefore, eagerly sought and translated, and thus much of the science of antiquity passed into Islāmic culture. Greek medicine, Greek astronomy and astrology, and Greek mathematics, together with the great philosophical works of Plato and, particularly, Aristotle, were assimilated in Islām by the end of the 9th century. Nor did the Arabs stop with assimilation.

They criticized and they innovated. Islāmic astronomy and astrology were aided by the construction of great astronomical observatories that provided accurate observations against which the Ptolemaic predictions could be checked. Numbers fascinated Islāmic thinkers, and this fascination served as the motivation for the creation of algebra (from Arabic *al-jabr*) and the study of algebraic functions.

MEDIEVAL EUROPEAN SCIENCE

Medieval Christendom confronted Islām chiefly in military crusades, in Spain and the Holy Land, and in theology. From this confrontation came the restoration of ancient learning to the West. The Reconquista in Spain gradually pushed the Moors south from the Pyrenees, and among the treasures left behind were Arabic translations of Greek works of science and philosophy. In 1085 the city of Toledo, with one of the finest libraries in Islām, fell to the Christians. Among the occupiers were Christian monks who quickly began the process of translating ancient works into Latin. By the end of the 12th century much of the ancient heritage was again available to the Latin West.

The medieval world was caricatured by thinkers of the 18th-century Enlightenment as a period of darkness, superstition, and hostility to science and learning. On the contrary, it was one of great technological vitality. The advances that were made may appear today as trifling, but that is because they were so fundamental. They included the horseshoe and the horse collar, without which horsepower cannot be efficiently exploited. The invention of the crank, the brace and bit, the wheelbarrow, and the flying buttress made possible the great Gothic cathedrals. Improvements in the gear trains of waterwheels and the development of windmills harnessed these sources of power with great efficiency. Mechanical ingenuity, building on experience with mills and power wheels, culminated in the 14th century in the mechanical clock, which not only set a new standard of chronometrical accuracy but also provided philosophers with a new metaphor for nature itself.

The mechanical clock

An equal amount of energy was devoted to achieving a scientific understanding of nature, but it is essential to understand to what use medieval thinkers put this kind of knowledge. As the fertility of the technology shows, medieval Europeans had no deep prejudices against utilitarian knowledge. But the areas in which scientific knowledge could find useful expression were few. Instead, science was viewed chiefly as a means of understanding God's creation and, thereby, the Godhead itself. The best example of this attitude is found in the medieval study of optics. Light, as Genesis makes clear, was among the first creations of God. The 12th–13th-century cleric-scholar Robert Grosseteste saw in light the first creative impulse. As light spread it created both space and matter, and, in its reflection from the outermost circle of the cosmos, it gradually solidified into the heavenly spheres. To understand the laws of the propagation of light was to understand, in some slight way, the nature of the creation.

In the course of studying light, particular problems were isolated and attacked. What, for example, is the rainbow? It is impossible to get close enough to a rainbow to see clearly what is going on, for as the observer moves, so too does the rainbow. It does seem to depend upon the presence of raindrops, so medieval investigators sought to bring the rainbow down from the skies into their studies. Insight into the nature of the rainbow could be achieved by simulating the conditions under which rainbows occur. For raindrops the investigators substituted hollow glass balls filled with water, so that the rainbow could be studied at leisure. Valid conclusions about rainbows could then be drawn by assuming the validity of the analogy between raindrops and water-filled globes. This involved the implicit assumptions that nature was simple (*i.e.,* governed by a few general laws) and that similar effects had similar causes. Such a nature was what could be expected of a rational, benevolent deity; hence, the assumption could be persuasively adopted.

Medieval philosophers were not content, as the above example shows, to repeat what the ancients had said. They subjected ancient texts to close critical scrutiny. Usually the intensity of the criticism was directly proportional to the theological significance of the problem involved. Such was the case with motion. Medieval philosophers examined all aspects of motion with great care, for the nature of motion had important theological implications. Thomas Aquinas used Aristotle's dictum, that everything that moves is moved by something else, to show that God must exist, for otherwise the existence of any motion would imply an infinite regression of prior causal motions.

It should be clear that there was no conscious conflict between science and religion in the Middle Ages. As Aquinas pointed out, God was the author of both the book of Scripture and the book of nature. The guide to nature was reason, the faculty that was the image of God in which mankind was made. Scripture was direct revelation, although it needed interpretation, for there were passages that were obscure or difficult. The two books, having the same author, could not contradict each other. For the short term, science and revelation marched hand in hand. Aquinas carefully wove knowledge of nature into his theology, as in his proof from motion of the existence of God. But if his scientific concepts of motion should ever be challenged, there would necessarily be a theological challenge as well. By working science into the very fabric of his theology, he virtually guaranteed that someday there would be conflict. Theologians would side with theology and scientists with science, to create a breach that neither particularly desired.

The glory of medieval science was its integration of science, philosophy, and theology into a magnificent and comprehensible whole. It can be best contemplated in the greatest of all medieval poems, Dante's *Divine Comedy.* Here was an essentially Aristotelian cosmos, finite and easily understood, over which God, his Son, and his saints reigned. Humanity and the Earth occupied the centre, as befitted their centrality in God's plan. The nine circles of hell were populated by humans whose exercise of their free will had led to their damnation. Purgatory contained lesser sinners still capable of salvation. The heavenly spheres were populated by the saved and the saintly. The natural hierarchy gave way to the spiritual hierarchy as one ascended toward the throne of God. Such a hierarchy was reflected in the social and political institutions of medieval Europe, and God, the supreme monarch, ruled his creation with justice and love. All fit together in a grand cosmic scheme, one not to be abandoned lightly.

Natural and spiritual hierarchy

The rise of modern science

THE AUTHORITY OF PHENOMENA

Even as Dante was writing his great work, deep forces were threatening the unitary cosmos he celebrated. The pace of technological innovation began to quicken. Particularly in Italy, the political demands of the time gave new importance to technology, and a new profession emerged, that of civil and military engineer. These people faced practical problems that demanded practical solutions. Leonardo da Vinci is certainly the most famous of them, though he was much more as well. A painter of genius, he closely studied human anatomy in order to give verisimilitude to his paintings. As a sculptor he mastered the difficult techniques of casting metal. As a producer-director of the form of Renaissance dramatic production called the masque, he devised complicated machinery to create special effects. But it was as a military engineer that he observed the path of a mortar bomb being lobbed over a city wall and insisted that the projectile did not follow two straight lines—a slanted ascent followed by a vertical drop—as Aristotle had said it must. Leonardo and his colleagues needed to know nature truly; no amount of book learning could substitute for actual experience, nor could books impose their authority upon phenomena. What Aristotle and his commentators asserted as philosophical necessity often did not gibe with what could be seen with one's own eyes. The hold of ancient philosophy was too strong to be broken lightly, but a healthy skepticism began to emerge.

The first really serious blow to the traditional acceptance of ancient authorities was the discovery of the New World at the end of the 15th century. Ptolemy, the great astronomer and geographer, had insisted that only the three continents of Europe, Africa, and Asia could exist, and Christian scholars from St. Augustine on had accepted it, for otherwise men would have to walk upside down at the antipodes. But Ptolemy, St. Augustine, and a host of other authorities were wrong. The dramatic expansion of the known world also served to stimulate the study of mathematics, for wealth and fame awaited those who could turn navigation into a real and trustworthy science.

In large part the Renaissance was a time of feverish intellectual activity devoted to the complete recovery of the ancient heritage. To the Aristotelian texts that had been the foundation of medieval thought were added translations of Plato, with his vision of mathematical harmonies, of Galen, with his experiments in physiology and anatomy, and, perhaps most important of all, of Archimedes, who showed how theoretical physics could be done outside the traditional philosophical framework. The results were subversive.

The search for antiquity turned up a peculiar bundle of manuscripts that added a decisive impulse to the direction in which Renaissance science was moving. These manuscripts were taken to have been written by or to report almost at first hand the activities of the legendary priest, prophet, and sage Hermes Trismegistos. Hermes was supposedly a contemporary of Moses, and the Hermetic writings contained an alternative story of creation that gave man a far more prominent role than the traditional account. God had made man fully in his image: a creator, not just a rational animal. Man could imitate God by creating. To do so, he must learn nature's secrets, and this could be done only by forcing nature to yield them through the tortures of fire, distillation, and other alchemical manipulations. The reward for success would be eternal life and youth, as well as freedom from want and disease. It was a heady vision, and it gave rise to the notion that, through science and technology, man could bend nature to his wishes. This is essentially the modern view of science, and it should be emphasized that it occurs only in Western civilization. It is probably this attitude that permitted the West to surpass the East, after centuries of inferiority, in the exploitation of the physical world.

Hermes Trismegistos

The Hermetic tradition also had more specific effects. Inspired, as is now known, by late Platonist mysticism, the Hermetic writers had rhapsodized on enlightenment and on the source of light, the Sun. Marsilio Ficino, the 15th-century Florentine translator of both Plato and the Hermetic writings, composed a treatise on the Sun that came close to idolatry. A young Polish student visiting Italy at the turn of the 16th century was touched by this current. Back in Poland, he began to work on the problems posed by the Ptolemaic astronomical system. With the blessing of the church, which he served formally as a canon, Nicolaus Copernicus set out to modernize the astronomical apparatus by which the church made such important calculations as the proper dates for Easter and other festivals.

THE SCIENTIFIC REVOLUTION

Copernicus. In 1543, as he lay on his deathbed, Copernicus finished reading the proofs of his great work; he died just as it was published. His *De revolutionibus orbium coelestium* (*On the Revolutions of the Celestial Spheres*) was the opening shot in a revolution whose consequences were greater than those of any other intellectual event in the history of mankind. The scientific revolution radically altered the conditions of thought and of material existence in which the human race lives, and its effects are not yet exhausted.

All this was caused by Copernicus' daring in placing the Sun, not the Earth, at the centre of the cosmos. Copernicus actually cited Hermes Trismegistos to justify this idea, and his language was thoroughly Platonic. But he meant his work as a serious work in astronomy, not philosophy, so he set out to justify it observationally and mathematically. The results were impressive. At one stroke, Copernicus reduced a complexity verging on chaos to elegant simplicity. The apparent back-and-forth movements of the planets, which required prodigious ingenuity to accommodate within the Ptolemaic system, could be accounted for just in terms of the Earth's own orbital motion added to or subtracted from the motions of the planets. Variation in planetary brightness was also explained by this combination of motions. The fact that Mercury and Venus were never found opposite the Sun in the sky Copernicus explained by placing their orbits closer to the Sun than that of the Earth. Indeed, Copernicus was able to place the planets in order of their distances from the Sun by considering their speeds and thus to construct a system of the planets, something that had eluded Ptolemy. This system had a simplicity, coherence, and aesthetic charm that made it irresistible to those who felt that God was the supreme artist. His was not a rigorous argument, but aesthetic considerations are not to be ignored in the history of science.

Copernicus did not solve all of the difficulties of the Ptolemaic system. He had to keep some of the cumbrous apparatus of epicycles and other geometrical adjustments, as well as a few Aristotelian crystalline spheres. The result was neater, but not so striking that it commanded immediate universal assent. Moreover, there were some implications that caused considerable concern: Why should the crystalline orb containing the Earth circle the Sun? And how was it possible for the Earth itself to revolve on its axis once in 24 hours without hurling all objects, including humans, off its surface? No known physics could answer these questions, and the provision of such answers was to be the central concern of the scientific revolution.

More was at stake than physics and astronomy, for one of the implications of the Copernican system struck at the very foundations of contemporary society. If the Earth revolved around the Sun, then the apparent positions of the fixed stars should shift as the Earth moves in its orbit. Copernicus and his contemporaries could detect no such shift (called stellar parallax), and there were only two interpretations possible to explain this failure. Either the Earth was at the centre, in which case no parallax was to be expected, or the stars were so far away that the parallax was too small to be detected. Copernicus chose the latter and thereby had to accept an enormous cosmos consisting mostly of empty space. God, it had been assumed, did nothing in vain, so for what purposes might he have created a universe in which the Earth and mankind were lost in immense space? To accept Copernicus was to give up the Dantean cosmos. The Aristotelian hierarchy of social place, political position, and theological gradation would vanish, to be replaced by the flatness and plainness of Euclidean space. It was a grim prospect and not one that recommended itself to most 16th-century intellectuals, and so Copernicus' grand idea remained on the periphery of astronomical thought. All astronomers were aware of it, some measured their own views against it, but only a small handful eagerly accepted it.

In the century and a half following Copernicus, two easily discernible scientific movements developed. The first was critical, the second, innovative and synthetic. They worked together to bring the old cosmos into disrepute and, ultimately, to replace it with a new one. Although they existed side by side, their effects can more easily be seen if they are treated separately.

Tycho, Kepler, and Galileo. The critical tradition began with Copernicus. It led directly to the work of Tycho Brahe, who measured stellar and planetary positions more accurately than had anyone before him. But measurement alone could not decide between Copernicus and Ptolemy, and Tycho insisted that the Earth was motionless. Copernicus did persuade Tycho to move the centre of revolution of all other planets to the Sun. To do so, he had to abandon the Aristotelian crystalline spheres that otherwise would collide with one another. Tycho also cast doubt upon the Aristotelian doctrine of heavenly perfection, for when, in the 1570s, a comet and a new star appeared, Tycho showed that they were both above the sphere of the Moon. Perhaps the most serious critical blows struck were those delivered by Galileo after the invention of the telescope. In quick succession, he announced that there were mountains on the Moon, satellites circling Jupiter, and spots upon the Sun. Moreover, the Milky Way was

Invention of the telescope

composed of countless stars whose existence no one had suspected until Galileo saw them. Here was criticism that struck at the very roots of Aristotle's system of the world.

At the same time Galileo was searching the heavens with his telescope, in Germany Johannes Kepler was searching them with his mind. Tycho's precise observations permitted Kepler to discover that Mars (and, by analogy, all the other planets) did not revolve in a circle at all, but in an ellipse, with the Sun at one focus. Ellipses tied all the planets together in grand Copernican harmony. The Keplerian cosmos was most un-Aristotelian, but Kepler hid his discoveries by burying them in almost impenetrable Latin prose in a series of works that did not circulate widely.

What Galileo and Kepler could not provide, although they tried, was an alternative to Aristotle that made equal sense. If the Earth revolves on its axis, then why do objects not fly off it? And why do objects dropped from towers not fall to the west as the Earth rotates to the east beneath them? And how is it possible for the Earth, suspended in empty space, to go around the Sun—whether in circles or ellipses—without anything pushing it? The answers were long in coming.

Galileo attacked the problems of the Earth's rotation and its revolution by logical analysis. Bodies do not fly off the Earth because they are not really revolving rapidly, even though their speed is high. In revolutions per minute, any body on the Earth is going very slowly and, therefore, has little tendency to fly off. Bodies fall to the base of towers from which they are dropped because they share with the tower the rotation of the Earth. Hence, bodies already in motion preserve that motion when another motion is added. So, Galileo deduced, a ball dropped from the top of a mast of a moving ship would fall at the base of the mast. If the ball were allowed to move on a frictionless horizontal plane, it would continue to move forever. Hence, Galileo concluded, the planets, once set in circular motion, continue to move in circles forever. Therefore, Copernican orbits exist. Galileo never acknowledged Kepler's ellipses; to do so would have meant abandoning his solution to the Copernican problem.

Kepler realized that there was a real problem with planetary motion. He sought to solve it by appealing to the one force that appeared to be cosmic in nature, namely magnetism. The Earth had been shown to be a giant magnet by William Gilbert in 1600, and Kepler seized upon this fact. A magnetic force, Kepler argued, emanated from the Sun and pushed the planets around in their orbits, but he was never able to quantify this rather vague and unsatisfactory idea.

Descartes

By the end of the first quarter of the 17th century Aristotelianism was rapidly dying, but there was no satisfactory system to take its place. The result was a mood of skepticism and unease, for, as one observer put it, "The new philosophy calls all in doubt." It was this void that accounted largely for the success of a rather crude system proposed by René Descartes. Matter and motion were taken by Descartes to explain everything by means of mechanical models of natural processes, even though he warned that such models were not the way nature probably worked. They provided merely "likely stories," which seemed better than no explanation at all.

Armed with matter and motion, Descartes attacked the basic Copernican problems. Bodies once in motion, Descartes argued, remain in motion in a straight line unless and until they are deflected from this line by the impact of another body. All changes of motion are the result of such impacts. Hence, the ball falls at the foot of the mast because, unless struck by another body, it continues to move with the ship. Planets move around the Sun because they are swept around by whirlpools of a subtle matter filling all space. Similar models could be constructed to account for all phenomena; the Aristotelian system could be replaced by the Cartesian. There was one major problem, however, and it sufficed to bring down Cartesianism. Cartesian matter and motion had no purpose, nor did Descartes's philosophy seem to need the active participation of a deity. The Cartesian cosmos, as Voltaire later put it, was like a watch that had been wound up at the creation and continues ticking to eternity.

Newton. The 17th century was a time of intense religious feeling, and nowhere was that feeling more intense than in Great Britain. There a devout young man, Isaac Newton, was finally to discover the way to a new synthesis in which truth was revealed and God was preserved.

Newton was both an experimental and a mathematical genius, and it was this combination that enabled him to establish both the Copernican system and a new mechanics. His method was simplicity itself: "from the phenomena of motions to investigate the forces of nature, and then from these forces to demonstrate the other phenomena." Newton's genius guided him in the selection of phenomena to be investigated, and his creation of a new and fundamental mathematical tool—the calculus (simultaneously invented by Gottfried Leibniz)—permitted him to submit the forces he inferred to calculation. The result was *Philosophiae Naturalis Principia Mathematica* (*Mathematical Principles of Natural Philosophy,* usually called simply the *Principia*), which appeared in 1687. Here was a new physics that applied equally well to terrestrial and celestial bodies. Copernicus, Kepler, and Galileo were all justified by Newton's analysis of forces. Descartes was utterly routed.

The role of the divine

Newton's three laws of motion and his principle of universal gravitation sufficed to regulate the new cosmos, but only, Newton believed, with the help of God. Gravity, he more than once hinted, was direct divine action, as were all forces for order and vitality. Absolute space, for Newton, was essential, because space was the "sensorium of God," and the divine abode must necessarily be the ultimate coordinate system. Finally, Newton's analysis of the mutual perturbations of the planets caused by their individual gravitational fields predicted the natural collapse of the solar system unless God acted to set things right again.

The diffusion of scientific method. The publication of the *Principia* marks the culmination of the movement begun by Copernicus and, as such, has always stood as the symbol of the scientific revolution. There were, however, similar attempts to criticize, systematize, and organize natural knowledge that did not lead to such dramatic results. In the same year as Copernicus' great volume, there appeared an equally important book on anatomy: Andreas Vesalius' *De humani corporis fabrica* ("On the Fabric of the Human Body," called the *De fabrica*), a critical examination of Galen's anatomy in which Vesalius drew on his own studies to correct many of Galen's errors. Vesalius, like Newton a century later, emphasized the phenomena, *i.e.,* the accurate description of natural facts. Vesalius' work touched off a flurry of anatomical work in Italy and elsewhere that culminated in the discovery of the circulation of the blood by William Harvey, whose *De Motu Cordis et Sanguinis in Animalibus* (*On the Motion of the Heart and Blood in Animals*) was published in 1628. This was the *Principia* of physiology that established anatomy and physiology as sciences in their own right. Harvey showed that organic phenomena could be studied experimentally and that some organic processes could be reduced to mechanical systems. The heart and the vascular system could be considered as a pump and a system of pipes and could be understood without recourse to spirits or other forces immune to analysis.

In other sciences the attempt to systematize and criticize was not so successful. In chemistry, for example, the work of the medieval and early modern alchemists had yielded important new substances and processes, such as the mineral acids and distillation, but had obscured theory in almost impenetrable mystical argot. Robert Boyle in England tried to clear away some of the intellectual underbrush by insisting upon clear descriptions, reproducibility of experiments, and mechanical conceptions of chemical processes. Chemistry, however, was not yet ripe for revolution.

In many areas there was little hope of reducing phenomena to comprehensibility, simply because of the sheer number of facts to be accounted for. New instruments like the microscope and the telescope vastly multiplied the worlds with which man had to reckon. The voyages of discovery brought back a flood of new botanical and zoological specimens that overwhelmed ancient classificatory schemes. The best that could be done was to describe new things accurately and hope that someday they could all be fitted together in a coherent way.

The growing flood of information put heavy strains upon old institutions and practices. It was no longer sufficient to publish scientific results in an expensive book that few could buy; information had to be spread widely and rapidly. Nor could the isolated genius, like Newton, comprehend a world in which new information was being produced faster than any single person could assimilate it. Natural philosophers had to be sure of their data, and to that end they required independent and critical confirmation of their discoveries. New means were created to accomplish these ends. Scientific societies sprang up, beginning in Italy in the early years of the 17th century and culminating in the two great national scientific societies that mark the zenith of the scientific revolution: the Royal Society of London for the Promotion of Natural Knowledge, created by royal charter in 1662, and the Académie des Sciences of Paris, formed in 1666. In these societies and others like them all over the world, natural philosophers could gather to examine, discuss, and criticize new discoveries and old theories. To provide a firm basis for these discussions, societies began to publish scientific papers. The Royal Society's *Philosophical Transactions,* which began as a private venture of its secretary, was the first such professional scientific journal. It was soon copied by the French academy's *Mémoires,* which won equal importance and prestige. The old practice of hiding new discoveries in private jargon, obscure language, or even anagrams gradually gave way to the ideal of universal comprehensibility. New canons of reporting were devised so that experiments and discoveries could be reproduced by others. This required new precision in language and a willingness to share experimental or observational methods. The failure of others to reproduce results cast serious doubts upon the original reports. Thus were created the tools for a massive assault on nature's secrets.

The rise of scientific societies

Even with the scientific revolution accomplished, much remained to be done. Again, it was Newton who showed the way. For the macroscopic world, the *Principia* sufficed. Newton's three laws of motion and the principle of universal gravitation were all that was necessary to analyze the mechanical relations of ordinary bodies, and the calculus provided the essential mathematical tools. For the microscopic world, Newton provided two methods. Where simple laws of action had already been determined from observation, as the relation of volume and pressure of a gas (Boyle's law, $pv = k$), Newton assumed forces between particles that permitted him to derive the law. He then used these forces to predict other phenomena, in this case the speed of sound in air, that could be measured against the prediction. Conformity of observation to prediction was taken as evidence for the essential truth of the theory. Second, Newton's method made possible the discovery of laws of macroscopic action that could be accounted for by microscopic forces. Here the seminal work was not the *Principia* but Newton's masterpiece of experimental physics, the *Opticks,* published in 1704, in which he showed how to examine a subject experimentally and discover the laws concealed therein. Newton showed how judicious use of hypotheses could open the way to further experimental investigation until a coherent theory was achieved. The *Opticks* was to serve as the model in the 18th and early 19th centuries for the investigation of heat, light, electricity, magnetism, and chemical atoms.

THE CLASSIC AGE OF SCIENCE

Mechanics. Just as the *Principia* preceded the *Opticks,* so, too, did mechanics maintain its priority among the sciences in the 18th century, in the process becoming transformed from a branch of physics into a branch of mathematics. Many physical problems were reduced to mathematical ones that proved amenable to solution by increasingly sophisticated analytical methods. The Swiss Leonhard Euler was one of the most fertile and prolific workers in mathematics and mathematical physics. His development of the calculus of variations provided a powerful tool for dealing with highly complex problems. In France, Jean Le Rond d'Alembert and Joseph-Louis Lagrange succeeded in completely mathematizing mechanics, reducing it to an axiomatic system requiring only mathematical manipulation.

The test of Newtonian mechanics was its congruence with physical reality. At the beginning of the 18th century it was put to a rigorous test. Cartesians insisted that the Earth, because it was squeezed at the Equator by the etherial vortex causing gravity, should be somewhat pointed at the poles, a shape rather like that of an American football; Newtonians, arguing that centrifugal force was greatest at the Equator, calculated an oblate sphere that was flattened at the poles and bulged at the Equator. The Newtonians were proved correct after careful measurements of a degree of the meridian were made on expeditions to Lapland and to Peru. The final touch to the Newtonian edifice was provided by Pierre-Simon, marquis de Laplace, whose masterly *Traité de mécanique céleste* (1798–1827; *Celestial Mechanics*) systematized everything that had been done in celestial mechanics under Newton's inspiration. Laplace went beyond Newton by showing that the perturbations of the planetary orbits caused by the interactions of planetary gravitation are in fact periodic and that the solar system is, therefore, stable, requiring no divine intervention.

Chemistry. Although Newton was unable to bring to chemistry the kind of clarification he brought to physics, the *Opticks* did provide a method for the study of chemical phenomena. One of the major advances in chemistry in the 18th century was the discovery of the role of air, and of gases generally, in chemical reactions. This role had been dimly glimpsed in the 17th century, but it was not fully seen until the classic experiments of Joseph Black on *magnesia alba* (basic magnesium carbonate) in the 1750s. By extensive and careful use of the chemical balance, Black showed that an air with specific properties could combine with solid substances like quicklime and could be recovered from them. This discovery served to focus attention on the properties of "air," which was soon found to be a generic, not a specific, name. Chemists discovered a host of specific gases and investigated their various properties: some were flammable, others put out flames; some killed animals, others made them lively. Clearly, gases had a great deal to do with chemistry.

Lavoisier

The Newton of chemistry was Antoine-Laurent Lavoisier. In a series of careful balance experiments Lavoisier untangled combustion reactions to show that, in contradiction to established theory, which held that a body gave off the principle of inflammation (called phlogiston) when it burned, combustion actually involves the combination of bodies with a gas that Lavoisier named oxygen. The chemical revolution was as much a revolution in method as in conception. Gravimetric methods made possible precise analysis, and this, Lavoisier insisted, was the central concern of the new chemistry. Only when bodies were analyzed as to their constituent substances was it possible to classify them and their attributes logically and consistently.

The imponderable fluids. The Newtonian method of inferring laws from close observation of phenomena and then deducing forces from these laws was applied with great success to phenomena in which no ponderable matter figured. Light, heat, electricity, and magnetism were all entities that were not capable of being weighed, *i.e.,* imponderable. In the *Opticks,* Newton had assumed that particles of different sizes could account for the different refrangibility of the various colours of light. Clearly, forces of some sort must be associated with these particles if such phenomena as diffraction and refraction are to be accounted for. During the 18th century heat, electricity, and magnetism were similarly conceived as consisting of particles with which were associated forces of attraction or repulsion. In the 1780s, Charles-Augustin de Coulomb was able to measure electrical and magnetic forces, using a delicate torsion balance of his own invention, and to show that these forces follow the general form of Newtonian universal attraction. Only light and heat failed to disclose such general force laws, thereby resisting reduction to Newtonian mechanics.

Science and the Industrial Revolution. It has long been a commonsensical notion that the rise of modern science and the Industrial Revolution were closely connected. It is difficult to show any direct effect of scientific discoveries upon the rise of the textile or even the metallurgical industry in Great Britain, the home of the Industrial Revolution, but there certainly was a similarity in attitude to be found in science and nascent industry. Close observation and careful generalization leading to practical utilization were characteristic of both industrialists and experimentalists alike in the 18th century. One point of direct contact is known, namely James Watt's interest in the efficiency of the Newcomen steam engine, an interest that grew from his work as a scientific-instrument maker and that led to his development of the separate condenser that made the steam engine an effective industrial power source. But in general the Industrial Revolution proceeded without much direct scientific help. Yet the potential influence of science was to prove of fundamental importance.

What science offered in the 18th century was the hope that careful observation and experimentation might improve industrial production significantly. In some areas, it did. The potter Josiah Wedgwood built his successful business on the basis of careful study of clays and glazes and by the invention of instruments like the pyrometer with which to gauge and control the processes he employed. It was not, however, until the second half of the 19th century that science was able to provide truly significant help to industry. It was then that the science of metallurgy permitted the tailoring of alloy steels to industrial specifications, that the science of chemistry permitted the creation of new substances, like the aniline dyes, of fundamental industrial importance, and that electricity and magnetism were harnessed in the electric dynamo and motor. Until that period science probably profited more from industry than the other way around. It was the steam engine that posed the problems that led, by way of a search for a theory of steam power, to the creation of thermodynamics. Most importantly, as industry required ever more complicated and intricate machinery, the machine tool industry developed to provide it and, in the process, made possible the construction of ever more delicate and refined instruments for science. As science turned from the everyday world to the worlds of atoms and molecules, electric currents and magnetic fields, microbes and viruses, and nebulae and galaxies, instruments increasingly provided the sole contact with phenomena. A large refracting telescope driven by intricate clockwork to observe nebulae was as much a product of 19th-century heavy industry as were the steam locomotive and the steamship.

Scientific and technical schools

The Industrial Revolution had one further important effect on the development of modern science. The prospect of applying science to the problems of industry served to stimulate public support for science. The first great scientific school of the modern world, the École Polytechnique in Paris, was founded in 1794 to put the results of science in the service of France. The founding of scores more technical schools in the 19th and 20th centuries encouraged the widespread diffusion of scientific knowledge and provided further opportunity for scientific advance. Governments, in varying degrees and at different rates, began supporting science even more directly, by making financial grants to scientists, by founding research institutes, and by bestowing honours and official posts on great scientists. By the end of the 19th century the natural philosopher following his private interests had given way to the professional scientist with a public role.

The Romantic revolt. Perhaps inevitably, the triumph of Newtonian mechanics elicited a reaction, one that had important implications for the further development of science. Its origins are many and complex, and it is possible here to focus on only one, that associated with the German philosopher Immanuel Kant. Kant challenged the Newtonian confidence that the scientist can deal directly with subsensible entities such as atoms, the corpuscles of light, or electricity. Instead, Kant insisted, all that the human mind can know is forces. This epistemological axiom freed Kantians from having to conceive of forces as embodied in specific and immutable particles. It also placed new emphasis on the space between particles; indeed, if one eliminated the particles entirely, there remained only space containing forces. From these two considerations were to come powerful arguments, first, for the transformations and conservation of forces and, second, for field theory as a representation of reality. What makes this point of view Romantic is that the idea of a network of forces in space tied the cosmos into a unity in which all forces were related to all others, so that the universe took on the aspect of a cosmic organism. The whole was greater than the sum of all its parts, and the way to truth was contemplation of the whole, not analysis.

What Romantics, or nature philosophers, as they called themselves, could see that was hidden from their Newtonian colleagues was demonstrated by Hans Christian Ørsted. He found it impossible to believe that there was no connection between the forces of nature. Chemical affinity, electricity, heat, magnetism, and light must, he argued, simply be different manifestations of the basic forces of attraction and repulsion. In 1820 he showed that electricity and magnetism were related, for the passage of an electrical current through a wire affected a nearby magnetic needle. This fundamental discovery was explored and exploited by Michael Faraday, who spent his whole scientific life converting one force into another. By concentrating on the patterns of forces produced by electric currents and magnets, Faraday laid the foundations for field theory, in which the energy of a system was held to be spread throughout the system and not localized in real or hypothetical particles.

The rudiments of field theory

The transformations of force necessarily raised the question of the conservation of force. Is anything lost when electrical energy is turned into magnetic energy, or into heat or light or chemical affinity or mechanical power? Faraday, again, provided one of the early answers in his two laws of electrolysis, based on experimental observations that quite specific amounts of electrical "force" decomposed quite specific amounts of chemical substances. This work was followed by that of James Prescott Joule, Robert Mayer, and Hermann von Helmholtz, each of whom arrived at a generalization of basic importance to all science, the principle of the conservation of energy.

The nature philosophers were primarily experimentalists who produced their transformations of forces by clever experimental manipulation. The exploration of the nature of elemental forces benefitted as well from the rapid development of mathematics. In the 19th century the study of heat was transformed into the science of thermodynamics, based firmly on mathematical analysis; the Newtonian corpuscular theory of light was replaced by Augustin-Jean Fresnel's mathematically sophisticated undulatory theory; and the phenomena of electricity and magnetism were distilled into succinct mathematical form by William Thomson (Lord Kelvin) and James Clerk Maxwell. By the end of the century, thanks to the principle of the conservation of energy and the second law of thermodynamics, the physical world appeared to be completely comprehensible in terms of complex but precise mathematical forms describing various mechanical transformations in some underlying ether.

The submicroscopic world of material atoms became similarly comprehensible in the 19th century. Beginning with John Dalton's fundamental assumption that atomic species differ from one another solely in their weights, chemists were able to identify an increasing number of elements and to establish the laws describing their interactions. Order was established by arranging elements according to their atomic weights and their reactions. The result was the periodic table, devised by Dmitry Mendeleyev, which implied that some kind of subatomic structure underlay elemental qualities. That structure could give rise to qualities, thus fulfilling the prophecy of the 17th-century mechanical philosophers, was shown in the 1870s by Joseph-Achille Le Bel and Jacobus van't Hoff, whose studies of organic chemicals showed the correlation between the arrangement of atoms or groups of atoms in space and specific chemical and physical properties.

The founding of modern biology. The study of living matter lagged far behind physics and chemistry, largely because organisms are so much more complex than inanimate bodies or forces. Harvey had shown that living matter could be studied experimentally, but his achievement stood alone for two centuries. For the time being, most students of living nature had to be content to classify living forms as best they could and to attempt to isolate and study aspects of living systems.

As has been seen, an avalanche of new specimens in both botany and zoology put severe pressure on taxonomy. A giant step forward was taken in the 18th century by the Swedish naturalist Carl von Linné, known by his Latinized name, Linnaeus, who introduced a rational, if somewhat artificial, system of binomial nomenclature. The very artificiality of Linnaeus' system, focussing as it did on only a few key structures, encouraged criticism and attempts at more natural systems. The attention thus called to the organism as a whole reinforced a growing intuition that species are linked in some kind of genetic relationship, an idea first made scientifically explicit by Jean-Baptiste, chevalier de Lamarck.

Darwin

Problems encountered in cataloging the vast collection of invertebrates at the Museum of Natural History in Paris led Lamarck to suggest that species change through time. This idea was not so revolutionary as it is usually painted, for although it did upset some Christians who read the book of Genesis literally, naturalists who noted the shading of natural forms one into another had been toying with the notion for some time. Lamarck's system failed to gain general assent largely because it relied upon an antiquated chemistry for its causal agents and appeared to imply a conscious drive to perfection on the part of organisms. It was also opposed by one of the most powerful paleontologists and comparative anatomists of the day, Georges Cuvier, who happened to take Genesis quite literally. In spite of Cuvier's opposition, however, the idea remained alive and was finally elevated to scientific status by the labours of Charles Darwin. Darwin not only amassed a wealth of data supporting the notion of transformation of species, but he also was able to suggest a mechanism by which such evolution could occur without recourse to other than purely natural causes. The mechanism was natural selection, according to which minute variations in offspring were either favoured or eliminated in the competition for survival, and it permitted the idea of evolution to be perceived with great clarity. Nature shuffled and sorted its own productions, through processes governed purely by chance, so that those organisms that survived were better adapted to a constantly changing environment.

Darwin's *On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life,* published in 1859, brought order to the world of organisms. A similar unification at the microscopic level had been brought about by the cell theory announced by Theodor Schwann and Matthias Schleiden in 1838, whereby cells were held to be the basic units of all living tissues. Improvements in the microscope during the 19th century made it possible gradually to lay bare the basic structures of cells, and rapid progress in biochemistry permitted the intimate probing of cellular physiology. By the end of the century the general feeling was that physics and chemistry sufficed to describe all vital functions and that living matter, subject to the same laws as inanimate matter, would soon yield up its secrets. This reductionist view was triumphantly illustrated in the work of Jacques Loeb, who showed that so-called instincts in lower animals are nothing more than physicochemical reactions, which he labelled tropisms.

The most dramatic revolution in 19th-century biology was the one created by the germ theory of disease, championed by Louis Pasteur in France and Robert Koch in Germany. Through their investigations, bacteria were shown to be the specific causes of many diseases. By means of immunological methods first devised by Pasteur, some of mankind's chief maladies were brought under control.

THE 20TH-CENTURY REVOLUTION

By the end of the 19th century, the dream of the mastery of nature for the benefit of mankind, first expressed in all its richness by Sir Francis Bacon, seemed on the verge of realization. Science was moving ahead on all fronts, reducing ignorance and producing new tools for the amelioration of the human condition. A comprehensible, rational view of the world was gradually emerging from laboratories and universities. One savant went so far as to express pity for those who would follow him and his colleagues, for they, he thought, would have nothing more to do than to measure things to the next decimal place.

Einstein

But this sunny confidence did not last long. One annoying problem was that the radiation emitted by atoms proved increasingly difficult to reduce to known mechanical principles. More importantly, physics found itself relying more and more upon the hypothetical properties of a substance, the ether, that stubbornly eluded detection. Within a span of 10 short years, roughly 1895–1905, these and related problems came to a head and wrecked the mechanistic system the 19th century had so laboriously built. The discovery of X rays and radioactivity revealed an unexpected new complexity in the structure of atoms. Max Planck's solution to the problem of thermal radiation introduced a discontinuity into the concept of energy that was inexplicable in terms of classical thermodynamics. Most disturbing of all, the enunciation of the special theory of relativity by Albert Einstein in 1905 not only destroyed the ether and all the physics that depended on it but also redefined physics as the study of relations between observers and events, rather than of the events themselves. What was observed, and therefore what happened, was now said to be a function of the observer's location and motion relative to other events. Absolute space was a fiction. The very foundations of physics threatened to crumble.

This modern revolution in physics has not yet been fully assimilated by historians of science. Suffice it to say that scientists managed to come to terms with all of the upsetting results of early 20th-century physics but in ways that made the new physics utterly different from the old. Mechanical models were no longer acceptable, because there were processes (like light) for which no consistent model could be constructed. No longer could physicists speak with confidence of physical reality, but only of the probability of making certain measurements.

All this being said, there is still no doubt that science in the 20th century has worked wonders. The new physics—relativity, quantum mechanics, particle physics—may outrage common sense, but it enables physicists to probe to the very limits of physical reality. Their instruments and mathematics permit modern scientists to manipulate subatomic particles with relative ease, to reconstruct the first moment of creation, and to glimpse dimly the grand structure and ultimate fate of the universe.

The revolution in physics has spilled over into chemistry and biology and led to hitherto undreamed of capabilities for the manipulation of atoms and molecules and of cells and their genetic structures. Chemists perform molecular tailoring today as a matter of course, cutting and shaping molecules at will. Genetic engineering makes possible active human intervention in the evolutionary process and holds out the possibility of tailoring living organisms, including the human organism, to specific tasks. This second scientific revolution may prove to be, for good or ill, the most important event in the history of mankind.

BIBLIOGRAPHY. For the legacy of the founder of the discipline of the history of science, GEORGE SARTON, see his *History of Science,* 2 vol. (1952–59), or his *Introduction to the History of Science,* 3 vol. in 5 (1927–48, reprinted 1975). A number of modern works cover the whole of the history of science. The most extensive is RENÉ TATON, *A General History of the Sciences,* 4 vol. (1963–66; U.S. title, *History of Science,* 4 vol., 1964–66; originally published in French, 1957–64). COLIN A. RONAN, *The Ages of Science* (1966), *Science: Its History and Development Among the World's Cultures* (1982), and *The Atlas of Scientific Discovery* (1983), are well-written surveys. L. PEARCE WILLIAMS and HENRY JOHN STEFFENS, *The History of Science in Western Civilization,* 3 vol. (1978–79), presents a narrative account of the history of Western science and excerpts from many scientific classics. J.D. BERNAL, *Science in History,* new ed., 4 vol. (1969, reissued 1979), is an orthodox Marxist version of the development of science by a distinguished scientist. The Marxist treatment only becomes seriously distorting when Bernal reaches the 19th century and begins to deal with the social sciences. A.E.E. MCKENZIE, *The Major Achievements of Science,* 2 vol. (1960, reissued 1973), is a brief and interesting account, accompanied by selections from primary sources.

The history of ancient mathematics and astronomy before the Greeks is brilliantly detailed by OTTO NEUGEBAUER, *The Exact Sciences in Antiquity*, 2nd ed. (1957); and by BARTEL L. VAN DER WAERDEN, *Science Awakening*, 4th ed., 2 vol. (1975; originally published in Dutch, 1950–65), who treats also the Greek achievement. A general history of Greek science is MARSHALL CLAGETT, *Greek Science in Antiquity* (1955, reissued 1976). The philosophical dimension of Greek science is masterfully presented by WILLIAM K.C. GUTHRIE, *A History of Greek Philosophy*, 6 vol. (1962–81). SAMUEL SAMBURSKY, *The Physical World of the Greeks*, trans. from the Hebrew (1956, reissued 1962), deals primarily with Stoic physics. EDWARD GRANT, *Physical Science in the Middle Ages* (1971, reprinted 1977), is a good survey, complete with an excellent bibliographical essay. More detailed accounts of medieval physics can be found in MARSHALL CLAGETT, *The Science of Mechanics in the Middle Ages* (1959). The influence of Hermeticism on the origins of modern science is treated by FRANCES A. YATES, *Giordano Bruno and the Hermetic Tradition* (1964, reprinted 1979); and D.P. WALKER, *Spiritual and Demonic Magic from Ficino to Campanella* (1958, reissued 1975).

A. RUPERT HALL, *The Scientific Revolution, 1500–1800: The Formation of the Modern Scientific Attitude*, 2nd ed. (1962, reprinted 1967), remains the best survey of that process. The vast literature is surveyed in RICHARD S. WESTFALL, *The Construction of Modern Science: Mechanisms and Mechanics* (1971, reprinted 1977). The same author's *Never at Rest: A Biography of Isaac Newton* (1980, reprinted 1983), is both brilliant and authoritative and should be read by anyone interested in the origins of modern physics. For a general history of science since the scientific revolution, see CHARLES COULSTON GILLISPIE, *The Edge of Objectivity: An Essay in the History of Scientific Ideas* (1960, reprinted 1973). The most useful reference for the history of science in general is CHARLES COULSTON GILLISPIE (ed.), *Dictionary of Scientific Biography*, 16 vol. (1970–80).

(L.P.W.)

The Art of Sculpture

Sculpture is not a fixed term that applies to a permanently circumscribed category of objects or sets of activities. It is, rather, the name of an art that grows and changes and is continually extending the range of its activities and evolving new kinds of objects. The scope of the term is much wider in the second half of the 20th century than it was only two or three decades ago, and in the present fluid state of the visual arts nobody can predict what its future extensions are likely to be.

Certain features, which in previous centuries were considered essential to the art of sculpture, are not present in a great deal of modern sculpture and can no longer form part of its definition. One of the most important of these is representation. Before the 20th century, sculpture was considered a representational art; but its scope has now been extended to include nonrepresentational forms. It has long been accepted that the forms of such functional three-dimensional objects as furniture, pots, and buildings may be expressive and beautiful without being in any way representational; but it is only in the 20th century that nonfunctional, nonrepresentational, three-dimensional works of art have been produced.

Again, before the 20th century, sculpture was considered primarily an art of solid form, or mass. It is true that the negative elements of sculpture—the voids and hollows within and between its solid forms—have always been to some extent an integral part of its design, but their role has been a secondary one. In a great deal of modern sculpture, however, the focus of attention has shifted, and the spatial aspects have become dominant. Spatial sculpture is now a generally accepted branch of the art of sculpture.

It was also taken for granted in the sculpture of the past that its components were of a constant shape and size and did not move. With the recent development of kinetic sculpture, neither the immobility nor immutability of its form can any longer be considered essential to the art of sculpture.

Finally, 20th-century sculpture is not confined to the two traditional forming processes of carving and modelling or to such traditional natural materials as stone, metal, wood, ivory, bone, and clay. Because present-day sculptors use any materials and methods of manufacture that will serve their purposes, the art of sculpture can no longer be identified with any special materials or techniques.

Through all of these changes there is probably only one thing that has remained constant in the art of sculpture, and it is this that emerges as the central and abiding concern of sculptors: the art of sculpture is the branch of the visual arts that is especially concerned with the creation of expressive form in three dimensions.

Sculpture may be either in the round or in relief. A sculpture in the round is a separate, detached object in its own right, leading the same kind of independent existence in space as a human body or a chair. A relief does not have this kind of independence. It projects from and is attached to or is an integral part of something else that serves either as a background against which it is set or a matrix from which it emerges.

The actual three-dimensionality of sculpture in the round limits its scope in certain respects in comparison with the scope of painting. Sculpture cannot conjure up the illusion of space by purely optical means or invest its forms with atmosphere and light as painting can. It does have a kind of reality, a vivid physical presence that is denied to the pictorial arts. The forms of sculpture are tangible as well as visible, and they can appeal strongly and directly to both tactile and visual sensibilities. Blind people, even those who are congenitally blind, can produce and appreciate certain kinds of sculpture. It has, in fact, been argued by the 20th-century art critic Sir Herbert Read that sculpture should be regarded as primarily an art of touch and that the roots of sculptural sensibility can be traced to the pleasure one experiences in fondling things.

All three-dimensional forms are perceived as having an expressive character as well as purely geometric properties. They strike the observer as delicate, aggressive, flowing, taut, relaxed, dynamic, soft, and so on. By exploiting the expressive qualities of form, a sculptor is able to create images in which subject matter and expressiveness of form are mutually reinforcing. Such images go beyond the mere presentation of fact and communicate a wide range of subtle and powerful feelings.

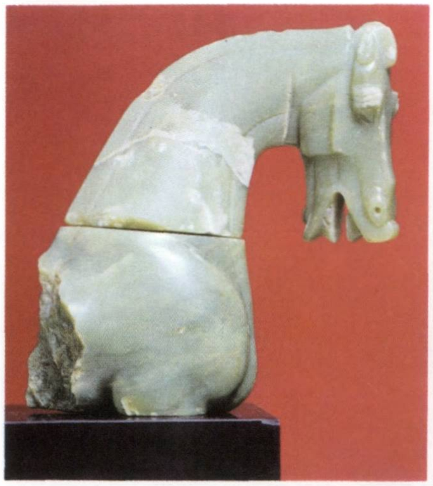
The aesthetic raw material of sculpture is, so to speak, the whole realm of expressive three-dimensional form. A sculpture may draw upon what already exists in the endless variety of natural and man-made form, or it may be an art of pure invention. It has been used to express a vast range of human emotions and feelings from the most tender and delicate to the most violent and ecstatic.

All human beings, intimately involved from birth with the world of three-dimensional form, learn something of its structural and expressive properties and develop emotional responses to them. This combination of understanding and sensitive response, often called a sense of form, can be cultivated and refined. It is to this sense of form that the art of sculpture primarily appeals.

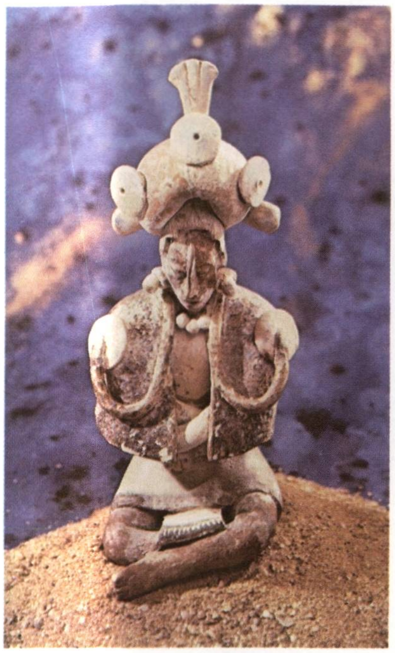
This article deals with the elements and principles of design; the materials, methods, techniques, and forms of sculpture; and its subject matter, imagery, symbolism, and uses. For the history of sculpture in the West, see SCULPTURE, THE HISTORY OF WESTERN. For treatments of sculpture as practiced in non-European cultures, see AFRICAN ARTS; AMERICAN INDIANS: *Visual Arts;* CENTRAL ASIAN ARTS; EAST ASIAN ARTS; EGYPTIAN ARTS AND ARCHITECTURE, ANCIENT; ISLĀMIC ARTS; OCEANIC ARTS; PREHISTORIC PEOPLES AND CULTURES; SOUTH ASIAN ARTS; SOUTHEAST ASIAN ARTS.

The article is divided into the following sections:

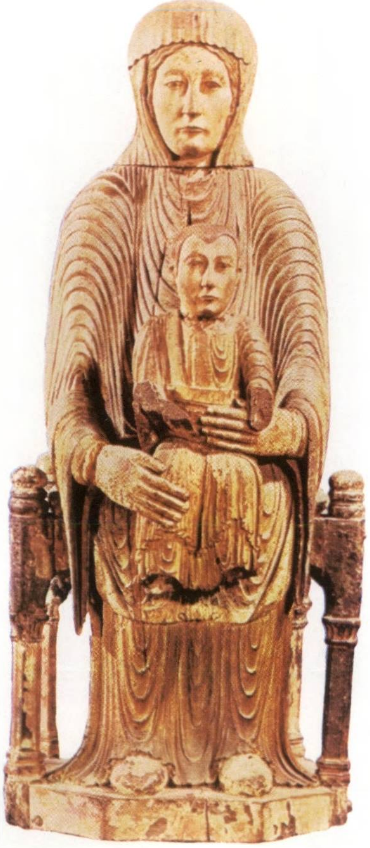
Diverse materials and techniques



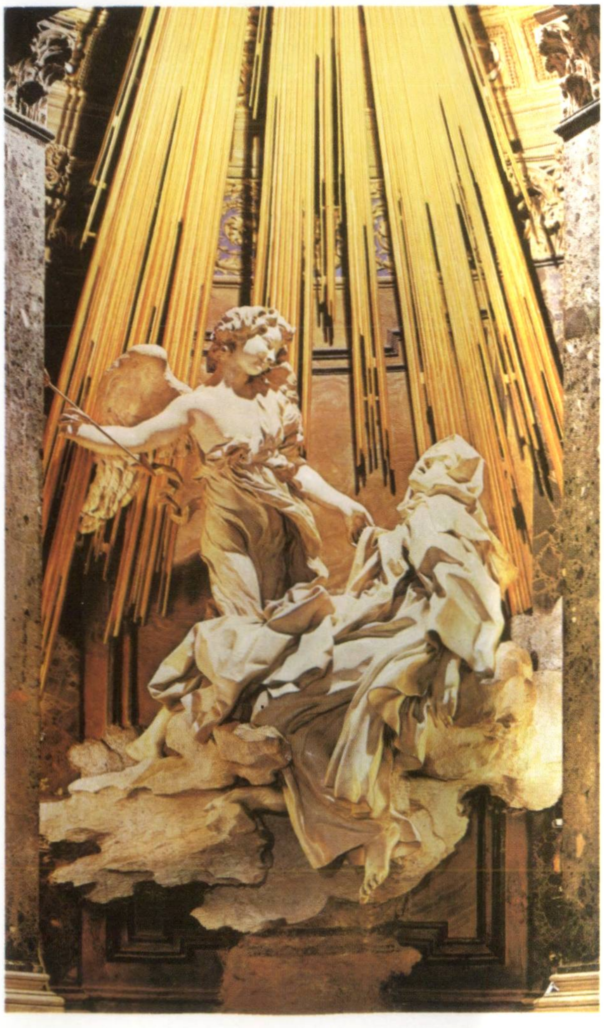
Jade horse head, Chinese, Han dynasty (206 BC–AD 220). In the Victoria and Albert Museum, London. Height 19 cm.

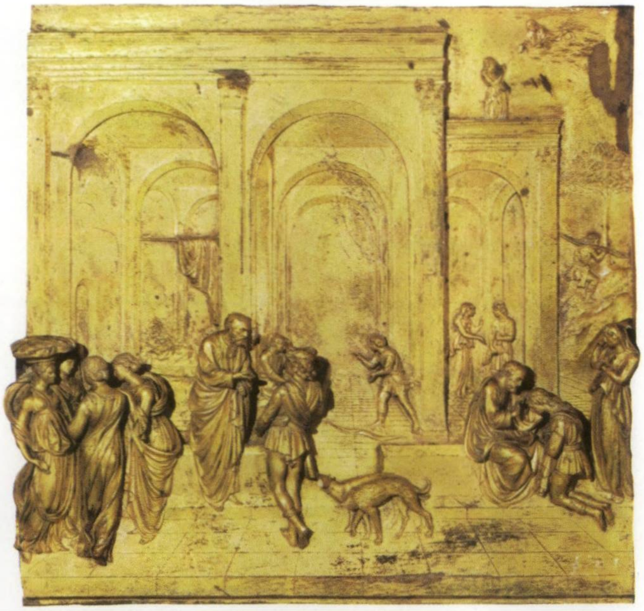


Jaina pottery figurine, late classic Maya style, from the state of Campeche, Mexico. In the collection of Dumbarton Oaks, Washington, D.C. Height 15.5 cm.



"Virgin and Child," polychromed oak statue of the school of Auvergne, France, 12th century. In the Metropolitan Museum of Art, New York. 78.7 × 32.4 cm.



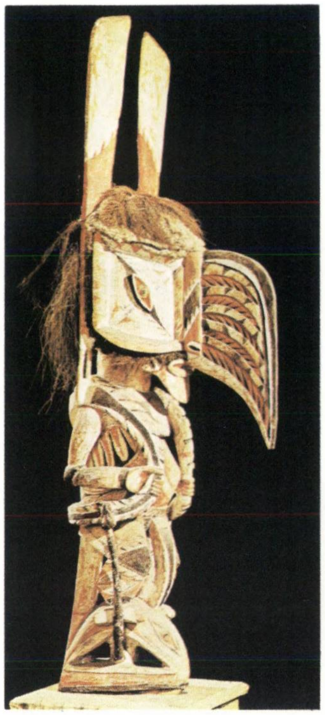


"Isaac, Jacob, and Esau," gilded bronze relief panel from the east doors ("Gates of Paradise") of the baptistery in Florence, by Lorenzo Ghiberti, 1425–52. 79.4 cm square.

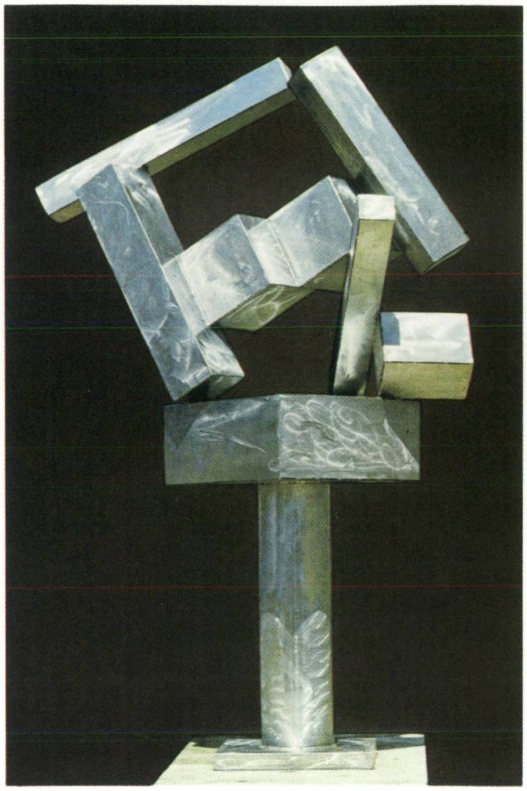
"The Ecstasy of St. Teresa," marble and gilded bronze niche sculpture by Gian Lorenzo Bernini, 1645–52. In the Coronaro Chapel, Sta. Maria della Vittoria, Rome. Lifesize.

Plate 1: By courtesy of (top left) the Victoria and Albert Museum, London, (centre) Dumbarton Oaks, Washington, D.C., (top right) the Metropolitan Museum of Art, New York, gift of J. Pierpont Morgan, 1916; photographs, (bottom left and right) SCALA—Art Resource/EB Inc.

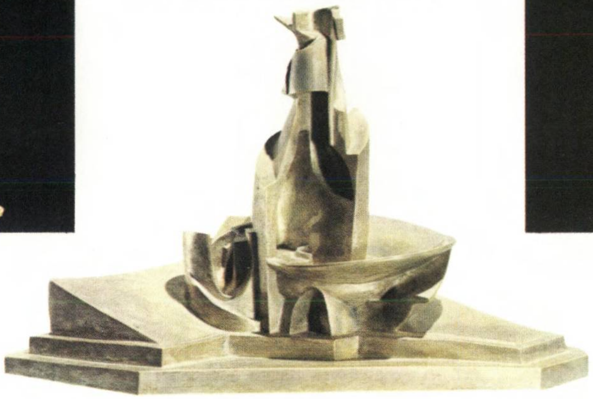
Diverse kinds of representational and nonrepresentational sculpture



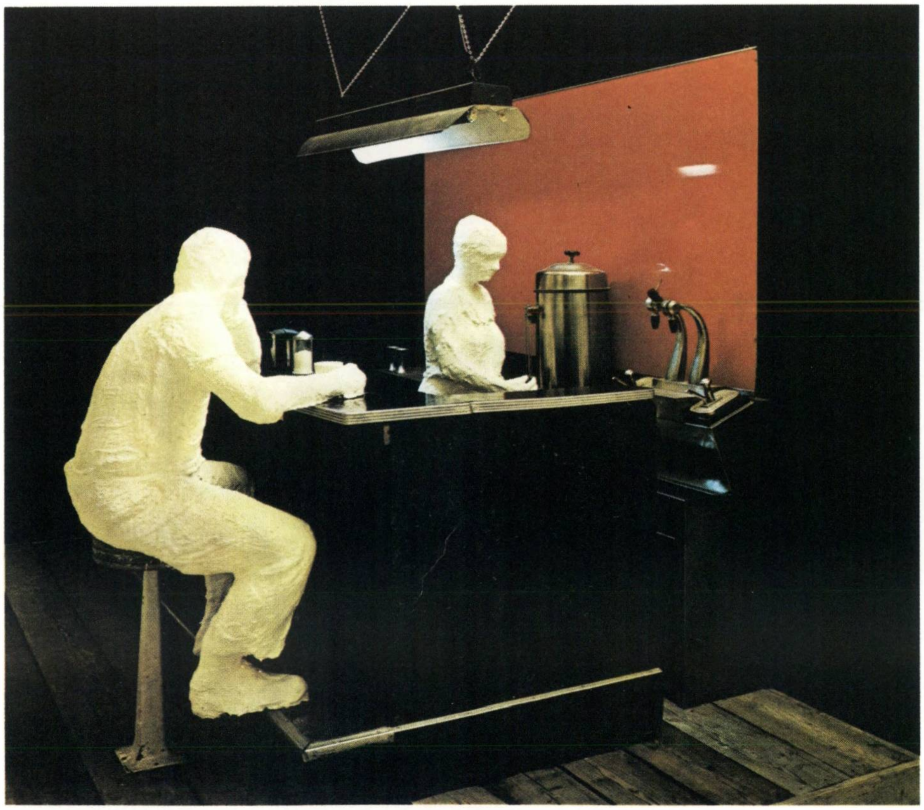
Painted wood male figure standing on a fish, fantastic cult image from Northern New Ireland, Melanesia. In a private collection. Height 1.53 m.



"Cubi XVII," non-objective sculpture by David Smith, stainless steel, 1963. In the Dallas Museum of Fine Arts, Texas. Height 2.74 m.



"Development of a Bottle in Space," nonrepresentational sculpture by Umberto Boccioni, silvered bronze, 1912. In the Museum of Modern Art, New York. 38.1 × 32.7 × 59.7 cm.



"The Diner," representational environmental sculpture by George Segal, mixed media (plaster, wood, chrome, masonite, and formica), 1964–66. In the Walker Art Center, Minneapolis. 2.59 × 2.74 × 2.13 m.

Plate 2: By courtesy of (centre) The Museum of Modern Art, New York, Aristide Maillol Fund, (top right) Dallas Museum of Fine Arts, the Eugene and Margaret McDermott Fund; (bottom) Collection Walker Art Center, Minneapolis; photograph, (top left) N. Heinrich

Elements and principles of sculptural design

Ways in which space enters into the design of sculpture

The two most important elements of sculpture—mass and space—are, of course, separable only in thought. All sculpture is made of a material substance that has mass and exists in three-dimensional space. The mass of sculpture is thus the solid, material, space-occupying bulk that is contained within its surfaces. Space enters into the design of sculpture in three main ways: the material components of the sculpture extend into or move through space; they may enclose or enfold space, thus creating hollows and voids within the sculpture; and they may relate one to another across space. Volume, surface, light and shade, and colour are supporting elements of sculpture.

ELEMENTS OF DESIGN

The amount of importance attached to either mass or space in the design of sculpture varies considerably. In Egyptian sculpture and in most of the sculpture of the 20th-century artist Constantin Brancusi, for example, mass is paramount, and most of the sculptor's thought has been devoted to shaping a lump of solid material. In 20th-century works by Antoine Pevsner or Naum Gabo, on the other hand, mass is reduced to a minimum, consisting only of transparent sheets of plastic or thin metal rods. The solid form of the components themselves is of little importance; their main function is to create movement through space and to enclose space. In works by such 20th-century sculptors as Henry Moore and Barbara Hepworth, the elements of space and mass are treated as more or less equal partners.

It is not possible to see the whole of a fully three-dimensional form at once. The observer can only see the whole of it if he turns it around or goes around it himself. For this reason it is sometimes mistakenly assumed that sculpture must be designed primarily to present a series of satisfactory projective views and that this multiplicity of views constitutes the main difference between sculpture and the pictorial arts, which present only one view of their subject. Such an attitude toward sculpture ignores the fact that it is possible to apprehend solid forms as volumes, to conceive an idea of them in the round from any one aspect. A great deal of sculpture is designed to be apprehended primarily as volume.

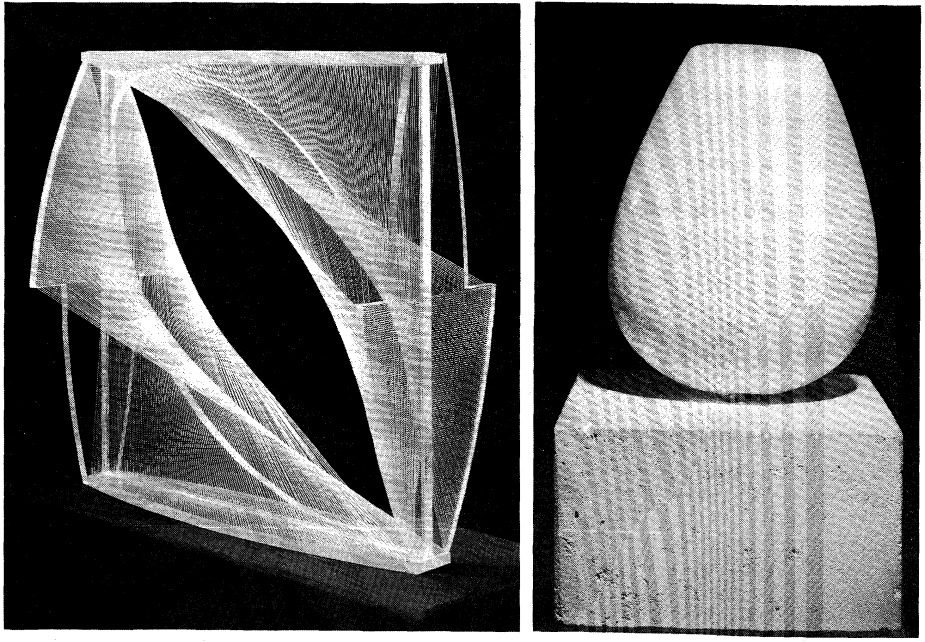
A single volume is the fundamental unit of three-dimensional solid form that can be conceived in the round. Some sculptures consist of only one volume, others are configurations of a number of volumes. The human figure is often treated by sculptors as a configuration of volumes, each of which corresponds to a major part of the body, such as the head, neck, thorax, and thigh.

Holes and cavities in sculpture, which are as carefully shaped as the solid forms and are of equal importance to the overall design, are sometimes referred to as negative volumes.

The surfaces of sculpture are in fact all that one actually sees. It is from their inflections that one makes inferences about the internal structure of the sculpture. A surface has, so to speak, two aspects: it contains and defines the internal structure of the masses of the sculpture, and it is the part of the sculpture that enters into relations with external space.

The expressive character of different kinds of surfaces is of the utmost importance in sculpture. Double-curved convex surfaces suggest fullness, containment, enclosure, the outward pressure of internal forces. In the aesthetics of Indian sculpture such surfaces have a special metaphysical significance. Representing the encroachment of space into the mass of the sculpture, concave surfaces suggest

By courtesy of (left) Miriam Gabo, Middlebury, Connecticut, (right) Philadelphia Museum of Art, the A.E. Gallatin Collection



Mass and space.
(Left) "Linear Construction #1, Variation," Perspex plastic and nylon thread sculpture by Naum Gabo, 1942–43. In the Miriam Gabo Collection, Middlebury, Connecticut. 62 cm. × 62 cm. (Right) "Torso of a Young Girl," onyx on a stone base by Constantin Brancusi, 1922. In the Philadelphia Museum of Art. Onyx height 34.9 cm.; base height 17.1 cm.

the action of external forces and are often indicative of collapse or erosion. Flat surfaces tend to convey a feeling of material hardness and rigidity; they are unbending or unyielding, unaffected by either internal or external pressures. Surfaces that are convex in one curvature and concave in the other can suggest the operation of internal pressures and at the same time a receptivity to the influence of external forces. They are associated with growth, with expansion into space.

The sculptor cannot, like the painter, create his own lighting effects within the work itself. The distribution of light and shade over the forms of his work depends upon the direction and intensity of light from external sources. Nevertheless, to some extent he can determine the kinds of effect this external light will have. If he knows where the work is to be sited, he can adapt it to the kind of light it is likely to receive. The brilliant overhead sunlight of Egypt and India demands a different treatment from the dim interior light of a northern medieval cathedral. Then again, it is possible to create effects of light and shade, or chiaroscuro, by cutting or modelling deep, shadow-catching hollows and prominent, highlighted ridges. Many late Gothic sculptors used light and shade as a powerful expressive feature of their work, aiming at a mysterious obscurity, with forms broken by shadow emerging from a dark background. Greek, Indian, and most Italian Renaissance sculptors shaped the forms of their work to receive light in a way that makes the whole work radiantly clear.

Chiaroscuro effects

The colouring of sculpture may be either natural or applied. In the recent past, sculptors became more aware than ever before of the inherent beauty of sculptural materials. Under the slogan of "truth to materials" many of them worked their materials in ways that exploited their natural properties, including colour and texture. More recently, however, there has been a growing tendency to use bright artificial colouring as an important element in the design of sculpture.

In the ancient world and during the Middle Ages almost all sculpture was artificially coloured, usually in a bold and decorative rather than a naturalistic manner. The sculptured portal of a cathedral, for example, would be coloured and gilded with all the brilliance of a contemporary illuminated manuscript. Combinations of differently coloured materials, such as the ivory and gold of some Greek sculpture, were not unknown before the 17th century; but the early Baroque sculptor Gian Lorenzo Bernini greatly extended the practice by combining variously coloured marbles with white marble and gilt bronze.

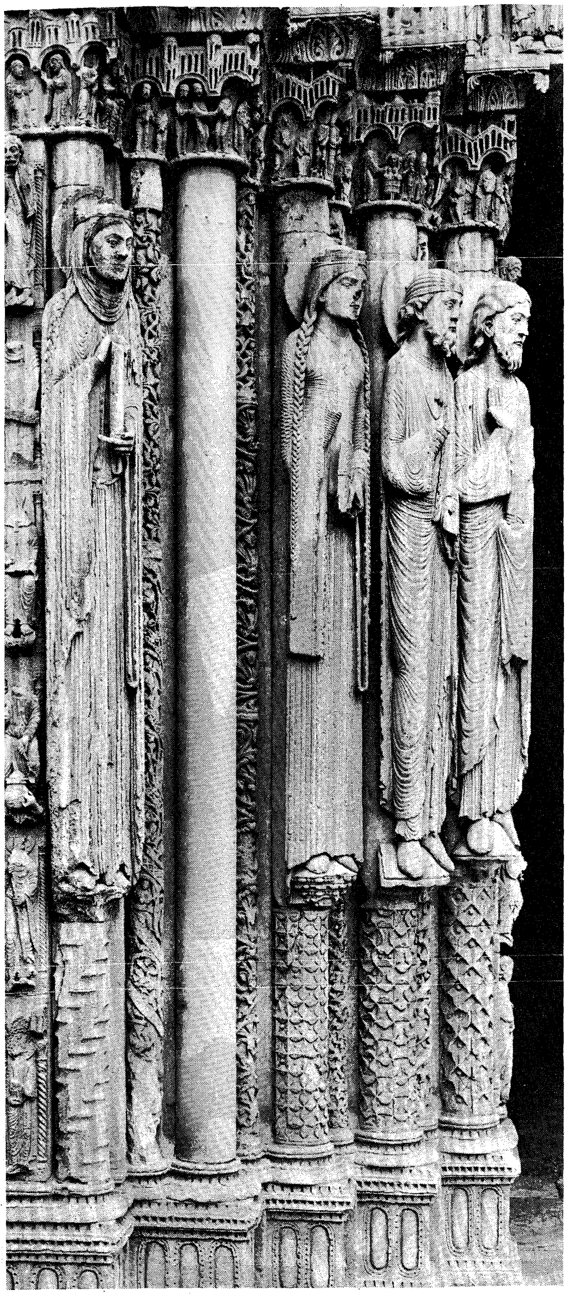
PRINCIPLES OF DESIGN

It is doubtful whether any principles of design are universal in the art of sculpture, for the principles that govern the organization of the elements of sculpture into expressive compositions differ from style to style. In fact, distinctions made among the major styles of sculpture are largely based on a recognition of differences in the principles of design that underlie them. Thus, the art historian Erwin Panofsky was attempting to define a difference of principle in the design of Romanesque and Gothic sculpture when he stated that the forms of Romanesque were conceived as projections from a plane outside themselves, while those of Gothic were conceived as being centred on an axis within themselves. The "principle of axiality" was considered by Panofsky to be "the essential principle of classical statuary," which Gothic had rediscovered.

The principles of sculptural design govern the approaches of sculptors to such fundamental matters as orientation, proportion, scale, articulation, and balance.

For conceiving and describing the orientation of the forms of sculpture in relation to each other, to a spectator, and to their surroundings, some kind of spatial scheme of reference is required. This is provided by a system of axes and planes of reference.

An axis is an imaginary centre line through a symmetrical or near symmetrical volume or group of volumes. Thus, all the main components of the human body have axes of their own, while an upright figure has a single vertical axis running through its entire length. Volumes may rotate or tilt on their axes.



Proportion.
Four figures possibly of the royal family of Judah, stone, 1145–50. Portail Royal of Chartres cathedral, France. Height approximately 2.50 m.
ND photo

Planes of reference are imaginary planes to which the movements, positions, and directions of volumes, axes, and surfaces may be referred. The principal planes of reference are the frontal, the horizontal, and the two profile planes.

The principles that govern the characteristic poses and spatial compositions of upright figures in different styles of sculpture are formulated with reference to axes and the four cardinal planes: for example, the principle of axiality already referred to; the principle of frontality, which governs the design of Archaic sculpture; the characteristic contrapposto (pose in which parts of the body, such as upper and lower, are twisted in opposite directions) of Michelangelo's figures; and in standing Greek sculpture of the Classical period the frequently used balanced "chiastic" pose (stance in which the body weight is taken principally on one leg, thereby creating a contrast of tension and relaxation between the opposite sides of a figure).

Proportional relations exist among linear dimensions,

Use of unnatural proportions

areas, and volumes and masses. All three types of proportion coexist and interact in sculpture, contributing to its expressiveness and beauty. Attitudes toward proportion differ considerably among sculptors. Some sculptors, both abstract and figurative, use mathematical systems of proportion; for example, the refinement and idealization of natural human proportions was a major preoccupation of Greek sculptors. Indian sculptors employed iconometric canons, or systems of carefully related proportions, that determined the proportions of all significant dimensions of the human figure. African and other tribal sculptors base the proportions of their figures on the subjective importance of the parts of the body. Unnatural proportions may be used for expressive purposes or to accommodate a sculpture to its surroundings. The elongation of the figures on the Portail Royal ("Royal Portal") of Chartres cathedral does both: it enhances their otherworldliness and also integrates them with the columnar architecture.

Sometimes it is necessary to adapt the proportions of sculpture to suit its position in relation to a viewer. A figure sited high on a building, for example, is usually made larger in its upper parts in order to counteract the effects of foreshortening. This should be allowed for when a sculpture intended for such a position is exhibited on eye level in a museum.

The scale of sculpture must sometimes be considered in relation to the scale of its surroundings. When it is one element in a larger complex, such as the facade of a building, it must be in scale with the rest. Another important consideration that sculptors must take into account when designing outdoor sculpture is the tendency of sculpture in the open air—particularly when viewed against the sky—to appear less massive than it does in a studio. Because one tends to relate the scale of sculpture to one's own human physical dimensions, the emotional impact of a colossal figure and a small figurine are quite different.

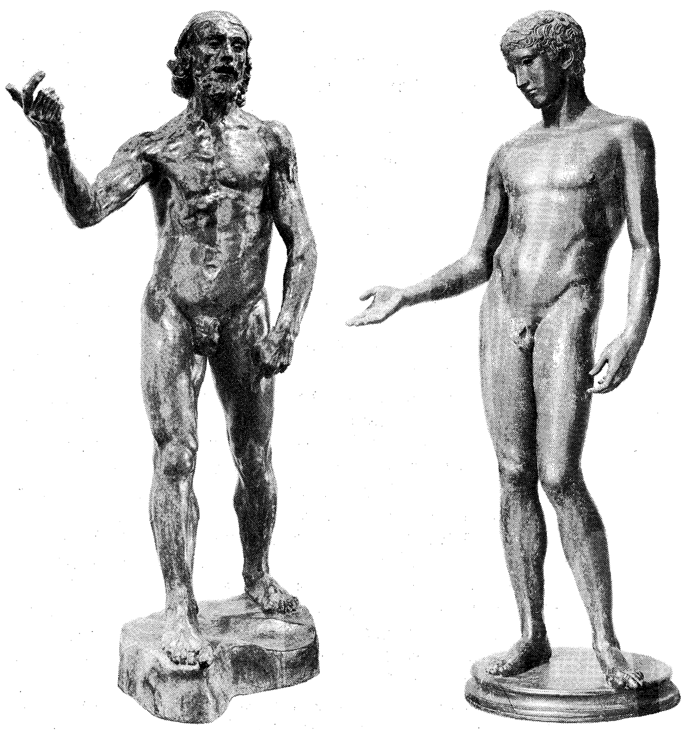
In ancient and medieval sculpture the relative scale of the figures in a composition is often determined by their importance; *e.g.,* slaves are much smaller than kings or nobles. This is sometimes known as hierarchic scale.

The joining of one form to another may be accomplished in a variety of ways. In much of the work of the 19th-century French sculptor Auguste Rodin, there are no clear boundaries, and one form is merged with another in an impressionistic manner to create a continuously flowing surface. In works by the Greek sculptor Praxiteles, the forms are softly and subtly blended by means of smooth, blurred transitions. The volumes of Indian sculpture and the surface anatomy of male figures in the style of the Greek sculptor Polyclitus are sharply defined and clearly articulated. One of the main distinctions between the work of Italian and northern Renaissance sculptors lies in the Italians' preference for compositions made up of clearly articulated, distinct units of form and the tendency of the northern Europeans to subordinate the individual parts to the allover flow of the composition.

The balance, or equilibrium, of freestanding sculpture has three aspects. First, the sculpture must have actual physical stability. This can be achieved by natural balance—that is, by making the sculpture stable enough in itself to stand firmly—which is easy enough to do with a four-legged animal or a reclining figure but not with a standing figure or a tall, thin sculpture, which must be secured to a base. The second aspect of balance is compositional. The interaction of forces and the distribution of weight within a composition may produce a state of either dynamic or static equilibrium. The third aspect of balance applies only to sculpture that represents a living figure. A live human figure balances on two feet by making constant movements and muscular adjustments. Such an effect can be conveyed in sculpture by subtle displacements of form and suggestions of tension and relaxation.

RELATIONSHIPS TO OTHER ARTS

Sculpture has always been closely related to architecture through its role as architectural decoration and also at the level of design. Architecture, like sculpture, is concerned with three-dimensional form; and, although the central problem in the design of buildings is the organization of space rather than mass, there are styles of architecture that are effective largely through the quality and organization of their solid forms. Ancient styles of stone architecture, particularly Egyptian, Greek, and Mexican, tend to treat their components in a sculptural manner. Moreover, most buildings viewed from the outside are compositions of masses. The growth of spatial sculpture is so intimately related to the opening up and lightening of architecture, which the development of modern building technology has made possible, that many 20th-century sculptors can be said to treat their work in an architectural manner.



Articulation.
(Left) Flowing surface exemplified by "St. John the Baptist Preaching," bronze sculpture by Auguste Rodin, 1878. In the Museum of Modern Art, New York City. Height 2 m. (Right) Delineated surface exemplified by "L'Idolino" bronze, Roman copy of a Greek sculpture in the style of Polyclitus, *c.* 440 BC. In the Museo Archeologico, Florence.

(Left) Collection, the Museum of Modern Art, New York, Mrs. Simon Guggenheim Fund; (right) Alinari—Art Resource

Sculpture and the pictorial arts

Some forms of relief sculpture approach very closely the pictorial arts of painting, drawing, engraving, and so on. And sculptures in the round that make use of chiaroscuro and that are conceived primarily as pictorial views rather than as compositions in the round are said to be "painterly"; for example, Bernini's "Ecstasy of St. Teresa" (Sta. Maria della Vittoria, Rome).

The borderlines between sculpture and pottery and the metalworking arts are not clear-cut, and many pottery and metal artifacts have every claim to be considered as sculpture. Today there is a growing affinity between the work of industrial designers and sculptors. Sculptural modelling techniques, and sometimes sculptors themselves, are often involved, for example, in the initial stages of the design of new automobile bodies.

The close relationships that exist between sculpture and the other visual arts are attested by the number of artists who have readily turned from one art to another; for example, Michelangelo, Bernini, Pisanello, Degas, and Picasso.

Materials

Any material that can be shaped in three dimensions can be used sculpturally. Certain materials, by virtue of their structural and aesthetic properties and their availability, have proved especially suitable. The most important of these are stone, wood, metal, clay, ivory, and plaster. There are also a number of materials of secondary importance and many that have only recently come into use.

PRIMARY

Throughout history, stone has been the principal material of monumental sculpture. There are practical reasons for this: many types of stone are highly resistant to the weather and therefore suitable for external use; stone is available in all parts of the world and can be obtained in large blocks; many stones have a fairly homogeneous texture and a uniform hardness that make them suitable for carving; stone has been the chief material used for the monumental architecture with which so much sculpture has been associated.

Stones belonging to all three main categories of rock formation have been used in sculpture. Igneous rocks, which are formed by the cooling of molten masses of mineral as they approach the Earth's surface, include granite, diorite, basalt, and obsidian. These are some of the hardest stones used for sculpture. Sedimentary rocks, which include sandstones and limestones, are formed from accumulated deposits of mineral and organic substances. Sandstones are agglomerations of particles of eroded stone held together by a cementing substance. Limestones are formed chiefly from the calcareous remains of organisms. Alabaster (gypsum), also a sedimentary rock, is a chemical deposit. Many varieties of sandstone and limestone, which vary greatly in quality and suitability for carving, are used for sculpture. Because of their method of formation, many sedimentary rocks have pronounced strata and are rich in fossils.

Metamorphic rocks result from changes brought about in the structure of sedimentary and igneous rocks by extreme pressure or heat. The most well-known metamorphic rocks used in sculpture are the marbles, which are recrystallized limestones. Italian Carrara marble, the best known, was used by Roman and Renaissance sculptors, especially Michelangelo, and is still widely used. The best known varieties used by Greek sculptors, with whom marble was more popular than any other stone, are Pentelic—from which the Parthenon and its sculpture are made—and Parian.

Because stone is extremely heavy and lacks tensile strength, it is easily fractured if carved too thinly and not properly supported. A massive treatment without vulnerable projections, as in Egyptian and pre-Columbian American Indian sculpture, is therefore usually preferred. Some stones, however, can be treated more freely and openly; marble in particular has been treated by some European sculptors with almost the same freedom as bronze, but such displays of virtuosity are achieved by overcoming rather than submitting to the properties of the material itself.

Colours and textures of stone

The colours and textures of stone are among its most delightful properties. Some stones are fine-grained and can be carved with delicate detail and finished with a high polish; others are coarse-grained and demand a broader treatment. Pure white Carrara marble, which has a translucent quality, seems to glow and responds to light in a delicate, subtle manner. (These properties of marble were brilliantly exploited by 15th-century Italian sculptors such as Donatello and Desiderio da Settignano.) The colouring of granite is not uniform but has a salt-and-pepper quality and may glint with mica and quartz crystals. It may be predominantly black or white or a variety of grays, pinks, and reds. Sandstones vary in texture and are often warmly coloured in a range of buffs, pinks, and reds. Limestones vary greatly in colour, and the presence of fossils may add to the interest of their surfaces. A number of stones are richly variegated in colour by the irregular veining that runs through them.

Hardstones, or semiprecious stones, constitute a special group, which includes some of the most beautiful and decorative of all substances. The working of these stones, along with the working of more precious gemstones, is usually considered as part of the glyptic (gem carving or engraving), or lapidary, arts, but many artifacts produced from them can be considered small-scale sculpture. They are often harder to work than steel. First among the hardstones used for sculpture is jade, which was venerated by the ancient Chinese, who worked it, together with other hardstones, with extreme skill. It was also used sculpturally by Mayan and Mexican artists. Other important hardstones are rock crystal, rose quartz, amethyst, agate, and jasper.

The principal material of tribal sculpture in Africa, Oceania, and North America, wood has also been used by every great civilization; it was used extensively during the Middle Ages, for example, especially in Germany and central Europe. Among modern sculptors who have used wood for important works are Ernst Barlach, Ossip Zadkine, and Henry Moore.

Characteristics of wood

Both hardwoods and softwoods are used for sculpture. Some are close-grained, and they cut like cheese; others are open-grained and stringy. The fibrous structure of wood gives it considerable tensile strength, so that it may be carved thinly and with greater freedom than stone. For large or complex open compositions, a number of pieces of wood may be jointed. Wood is used mainly for indoor sculpture, for it is not as tough or durable as stone; changes of humidity and temperature may cause it to split, and it is subject to attack by insects and fungus. The grain of wood is one of its most attractive features, giving variety of pattern and texture to its surfaces. Its colours, too, are subtle and varied. In general, wood has a warmth that stone does not have, but it lacks the massive dignity and weight of stone.

The principal woods for sculpture are oak, mahogany, limewood, walnut, elm, pine, cedar, boxwood, pear, and ebony; but many others are also used. The sizes of wood available are limited by the sizes of trees; North American Indians, for example, could carve gigantic totem poles in pine, but boxwood is available only in small pieces.

In the 20th century, wood was being used by many sculptors as a medium for construction as well as for carving. Laminated timbers, chipboards, and timber in block and plank form can be glued, jointed, screwed, or bolted together, and given a variety of finishes.

Wherever metal technologies have been developed, metals have been used for sculpture. The amount of metal sculpture that has survived from the ancient world does not properly reflect the extent to which it was used, for vast quantities have been plundered and melted down. Countless Far Eastern and Greek metal sculptures have been lost in this way, as has almost all the goldwork of pre-Columbian American Indians.

Advantages of bronze

The metal most used for sculpture is bronze, which is basically an alloy of copper and tin; but gold, silver, aluminum, copper, brass, lead, and iron have also been widely used. Most metals are extremely strong, hard, and durable, with a tensile strength that permits a much greater freedom of design than is possible in either stone or wood. A life-size bronze figure that is firmly attached to a base needs no support other than its own feet and may even be poised on one foot. Considerable attenuation of form is also possible without risk of fracture.

The colour, brilliant lustre, and reflectivity of metal surfaces have been highly valued and made full use of in sculpture although, since the Renaissance, artificial patinas have generally been preferred as finishes for bronze.

Metals can be worked in a variety of ways in order to produce sculpture. They can be cast—that is, melted and poured into molds; squeezed under pressure into dies, as in coin making; or worked directly—for example, by hammering, bending, cutting, welding, and repoussé (hammered or pressed in relief).

Important traditions of bronze sculpture are Greek, Roman, Indian (especially Cōḻa), African (Bini and Yoruba), Italian Renaissance, and Chinese. Gold was used to great effect for small-scale works in pre-Columbian America and medieval Europe. A fairly recent discovery, aluminum has been used a great deal by modern sculptors. Iron has not been used much as a casting material, but in recent years it has become a popular material for direct working by techniques similar to those of the blacksmith. Sheet metal is one of the principal materials used nowadays for constructional sculpture. Stainless steel in sheet form has been used effectively by the American sculptor David Smith.

Clay is one of the most common and easily obtainable of all materials. Used for modelling animal and human figures long before men discovered how to fire pots, it has been one of the sculptor's chief materials ever since.

Clay has four properties that account for its widespread use: when moist, it is one of the most plastic of all substances, easily modelled and capable of registering the most detailed impressions; when partially dried out to a leather-hard state or completely dried, it can be carved and scraped; when mixed with enough water, it becomes a creamy liquid known as slip, which may be poured into molds and allowed to dry; when fired to temperatures of between 700° and 1,400° C (1,300° and 2,600° F), it undergoes irreversible structural changes that make it permanently hard and extremely durable.

Uses of clay

Sculptors use clay as a material for working out ideas; for preliminary models that are subsequently cast in such materials as plaster, metal, and concrete or carved in stone; and for pottery sculpture.

Depending on the nature of the clay body itself and the temperature at which it is fired, a finished pottery product is said to be earthenware, which is opaque, relatively soft, and porous; stoneware, which is hard, nonporous, and more or less vitrified; or porcelain, which is fine-textured, vitrified, and translucent. All three types of pottery are used for sculpture. Sculpture made in low-fired clays, particularly buff and red clays, is known as terra-cotta (baked earth). This term is used inconsistently, however, and is often extended to cover all forms of pottery sculpture.

Unglazed clay bodies can be smooth or coarse in texture and may be coloured white, gray, buff, brown, pink, or red. Pottery sculpture can be decorated with any of the techniques invented by potters and coated with a variety of beautiful glazes.

Paleolithic sculptors produced relief and in-the-round work in unfired clay. The ancient Chinese, particularly during the T'ang (618–907) and Sung (960–1279) dynasties, made superb pottery sculpture, including large-scale human figures. The best known Greek works are the intimate small-scale figures and groups from Tanagra. Mexican and Mayan sculptor-potters produced vigorous, directly modelled figures. During the Renaissance, pottery was used in Italy for major sculptural projects, including the large-scale glazed and coloured sculptures of Luca Della Robbia and his family, which are among the finest works in the medium. One of the most popular uses of the pottery medium has been for the manufacture of figurines—at Staffordshire, Meissen, and Sèvres, for example.

Sources of ivory

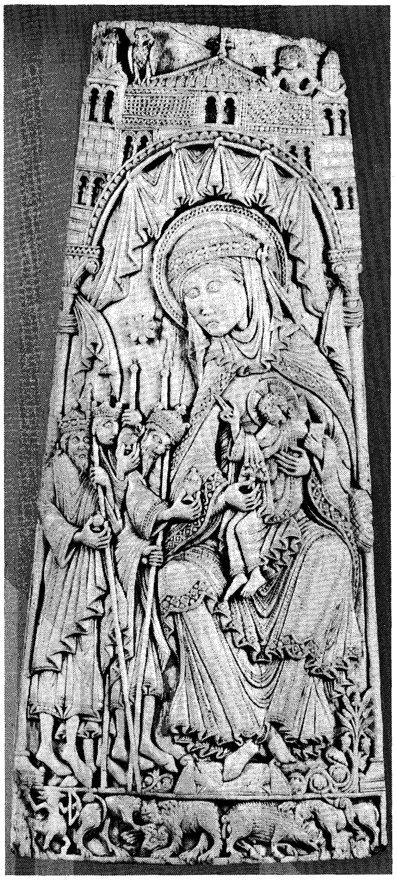
The main source of ivory is elephant tusks; but walrus, hippopotamus, narwhal (an Arctic aquatic animal), and, in Paleolithic times, mammoth tusks also were used for sculpture. Ivory is dense, hard, and difficult to work. Its colour is creamy white, which usually yellows with age; and it will take a high polish. A tusk may be sawed into panels for relief carving or into blocks for carving in the round; or the shape of the tusk itself may be used. The physical properties of the material invite the most delicate, detailed carving, and displays of virtuosity are common.

Ivory was used extensively in antiquity in the Middle and Far East and the Mediterranean. An almost unbroken Christian tradition of ivory carving reaches from Rome and Byzantium to the end of the Middle Ages. Throughout this time, ivory was used mainly in relief, often in conjunction with precious metals, enamels, and precious stones to produce the most splendid effects. Some of its main sculptural uses were for devotional diptychs, portable altars, book covers, retables (raised shelves above altars), caskets, and crucifixes. The Baroque period, too, is rich in ivories, especially in Germany. A fine tradition of ivory carving also existed in Benin, a former kingdom of West Africa.

Related to ivory, horn and bone have been used since Paleolithic times for small-scale sculpture. Reindeer horn and walrus tusks were two of the Eskimo carver's most important materials. One of the finest of all medieval "ivories" is a carving in whalebone, "The Adoration of the Magi" (Victoria and Albert Museum, London).

Plaster of Paris (sulfate of lime) is especially useful for the production of molds, casts, and preliminary models. It was used by Egyptian and Greek sculptors as a casting medium and is today the most versatile material in the sculptor's workshop.

When mixed with water, plaster will in a short time recrystallize, or set—that is, become hard and inert—and its volume will increase slightly. When set, it is relatively fragile and lacking in character and is therefore of limited use for finished work. Plaster can be poured as a liquid, modelled directly when of a suitable consistency, or easily carved after it has set. Other materials can be added to it to retard its setting, to increase its hardness or resistance to heat, to change its colour, or to reinforce it.



Bone carving.
"The Adoration of the Magi," whalebone, English, 11th–12th century. In the Victoria and Albert Museum, London. Height 37 cm.
By courtesy of Victoria and Albert Museum, London

The main sculptural use of plaster in the past was for molding and casting clay models as a stage in the production of cast metal sculpture. Many sculptors today omit the clay-modelling stage and model directly in plaster. As a mold material in the casting of concrete and fibreglass sculpture, plaster is widely used. It has great value as a material for reproducing existing sculpture; many museums, for example, use such casts for study purposes.

SECONDARY

Basically, concrete is a mixture of an aggregate (usually sand and small pieces of stone) bound together by cement. A variety of stones, such as crushed marble, granite chips, and gravel, can be used, each giving a different effect of colour and texture. Commercial cement is gray, white, or black; but it can be coloured by additives. The cement most widely used by sculptors is *ciment fondu,* which is extremely hard and quick setting. A recent invention—at least, in appropriate forms for sculpture—concrete is rapidly replacing stone for certain types of work. Because it is cheap, hard, tough, and durable, it is particularly suitable for large outdoor projects, especially decorative wall surfaces. With proper reinforcement it permits great freedom of design. And by using techniques similar to those of the building industry, sculptors are able to create works in concrete on a gigantic scale.

Advantages of concrete

When synthetic resins, especially polyesters, are reinforced with laminations of glass fibre, the result is a lightweight shell that is extremely strong, hard, and durable. It is usually known simply as fibreglass. After having been successfully used for car bodies, boat hulls, and the like, it has developed recently into an important material for sculpture. Because the material is visually unattractive in itself, it is usually coloured by means of fillers and pigments. It was first used in sculpture in conjunction with powdered metal fillers in order to produce cheap "cold-cast" substitutes for bronze and aluminum, but with the recent tendency to use bright colours in sculpture it is now often coloured either by pigmenting the material itself or by painting.

It is possible to model fibreglass, but more usually it is cast as a laminated shell. Its possibilities for sculpture have not yet been fully exploited.

Various formulas for modelling wax have been used in the past, but these have been generally replaced by synthetic waxes. The main uses of wax in sculpture have been as a preliminary modelling material for metal casting by the lost-wax, or cire perdue, process (see *Methods and techniques,* below) and for making sketches. It is not durable enough for use as a material in its own right, although it has been used for small works, such as wax fruit, that can be kept under a glass dome.

Papier-mâché (pulped paper bonded with glue) has been used for sculpture, especially in the Far East. Mainly used for decorative work, especially masks, it can have considerable strength; the Japanese, for example, made armour from it. Sculpture made of sheet paper is a limited art form used only for ephemeral and usually trivial work.

Numerous other permanent materials—such as shells, amber, and brick—and ephemeral ones—such as feathers, baker's dough, ice and snow, and cake icing—have been used for fashioning three-dimensional images. In view of recent trends in sculpture it is no longer possible to speak of "the materials of sculpture." Modern sculpture has no special materials. Any material, natural or man-made, is likely to be used, including inflated polyethylene, foam rubber, expanded polystyrene, fabrics, and neon tubes; the materials for a mid-20th-century sculpture by Claes Oldenburg, for example, are listed as canvas, cloth, Dacron, metal, foam rubber, and Plexiglas. Real objects, too, may be incorporated in sculpture, as in the mixed-medium compositions of Edward Kienholz; even junk has its devotees, who fashion "junk" sculpture.

Unconventional materials of modern sculpture.
"Giant Hamburger," painted sailcloth stuffed with foam rubber, by Claes Oldenburg, 1962. In the Art Gallery of Ontario. 132 × 213 cm.
By courtesy of the Art Gallery of Ontario

Methods and techniques

Although a sculptor may specialize in, say, stone carving or direct metalwork, the art of sculpture is not identifiable with any particular craft or set of crafts. It presses into its service whatever crafts suit its purposes. Technologies developed for more utilitarian purposes are often easily adapted for sculpture; in fact, useful artifacts and sculptured images have often been produced in the same workshop, sometimes by the same craftsman. The methods and techniques employed in producing a pot, a bronze harness trapping, a decorative stone molding or column, a carved wooden newel post, or even a fibreglass car body are essentially the same as those used in sculpture. For example, the techniques of repoussé, metal casting, blacksmithing, sheet-metal work, and welding, which are used for the production of functional artifacts and decorative metalwork, are also used in metal sculpture; and the preparation, forming, glazing, decoration, and firing of clay are basically the same in both utilitarian pottery and pottery sculpture. The new techniques used by sculptors today are closely related to new techniques applied in building and industrial manufacture.

THE SCULPTOR AS DESIGNER AND AS CRAFTSMAN

The conception of an artifact or a work of art—its form, imaginative content, and expressiveness—is the concern of a designer, and it should be distinguished from the execution of the work in a particular technique and material, which is the task of a craftsman. A sculptor usually functions as both designer and craftsman, but these two aspects of sculpture may be separated.

Expressiveness of direct handling

Certain types of sculpture depend considerably for their aesthetic effect on the way in which their material has been directly manipulated by the artist himself. The direct, expressive handling of clay in a model by Rodin, or the use of the chisel in the stiacciato (very low) reliefs of the 15th-century Florentine sculptor Donatello could no more have been delegated to a craftsman than could the brushwork of Rembrandt. The actual physical process of working materials is for many sculptors an integral part of the art of sculpture, and their response to the working qualities of the material—such as its plasticity, hardness, and texture—is evident in the finished work. Design and craftsmanship are intimately fused in such a work, which is a highly personal expression.

Even when the direct handling of material is not as vital as this to the expressiveness of the work, it still may be impossible to separate the roles of the artist as designer and craftsman. The qualities and interrelationships of forms may be so subtle and complex that they cannot be adequately specified and communicated to a craftsman. Moreover, many aspects of the design may actually be contributed during the process of working. Michelangelo's way of working, for example, enabled him to change his mind about important aspects of composition as the work proceeded.

A complete fusion of design and craftsmanship may not be possible if a project is a large one or if the sculptor is too old or too weak to do all of the work himself. The sheer physical labour of making a large sculpture can be considerable, and sculptors from Phidias in the 5th century BC to Henry Moore in the 20th century have employed pupils and assistants to help with it. Usually the sculptor delegates the time-consuming first stages of the work or some of its less important parts to his assistants and executes the final stages or the most important parts himself.

On occasion, a sculptor may function like an architect or industrial designer. He may do no direct work at all on the finished sculpture, his contribution being to supply exhaustive specifications in the form of drawings and perhaps scale models for a work that is to be entirely fabricated by craftsmen. Obviously, such a procedure excludes the possibility of direct, personal expression through the handling of the materials; thus, works of this kind usually have the same anonymous, impersonal quality as architecture and industrial design. An impersonal approach to sculpture was favoured by many sculptors of the 1960s such as William Tucker, Donald Judd, and William Turnbull. They used the skilled anonymous workmanship of industrial fabrications to make their large-scale, extremely precise, simple sculptural forms that are called "primary structures."

Stages in the production of a major work

General methods. Broadly speaking, the stages in the production of a major work of sculpture conform to the following pattern: the commission; the preparation, submission, and acceptance of the design; the selection and preparation of materials; the forming of materials; surface finishing; installation or presentation.

Almost all of the sculpture of the past and some present-day sculpture originates in a demand made upon the sculptor from outside, usually in the form of a direct commission or through a competition. If the commission is for a portrait or a private sculpture, the client may only require to see examples of the artist's previous work; but if it is a public commission, the sculptor is usually expected to submit drawings and maquettes (small-scale, three-dimensional sketch models) that give an idea of the nature of the finished work and its relation to the site. He may be free to choose his own subject matter or theme, or it may be more or less strictly prescribed. A medieval master sculptor, for example, received the program for a complex scheme of church sculpture from theological advisers, and Renaissance contracts for sculpture were often extremely specified and detailed. Today a great deal of sculpture is not commissioned. It arises out of the sculptor's private concern with form and imagery, and he works primarily to satisfy himself. When the work is finished he may exhibit and attempt to sell it in an art gallery.

Most of the materials used by 20th-century sculptors are readily available in a usable form from builders' or sculptors' suppliers, but certain kinds of sculpture may involve a good deal of preparatory work on the materials. A sculptor may visit a stone quarry in order to select the material for a large project and to have it cut into blocks of the right size and shape. And since stone is costly to transport and best carved when freshly quarried, he may decide to do all of his work at the quarry. Because stone is extremely heavy, the sculptor must have the special equipment required for manoeuvring even small blocks into position for carving. A wood carver requires a supply of well-seasoned timber and may keep a quantity of logs and blocks in store. A modeller needs a good supply of clay of the right kind. For large terra-cottas he may require a specially made-up clay body, or he may work at a brickworks, using the local clay and firing in the brick kilns.

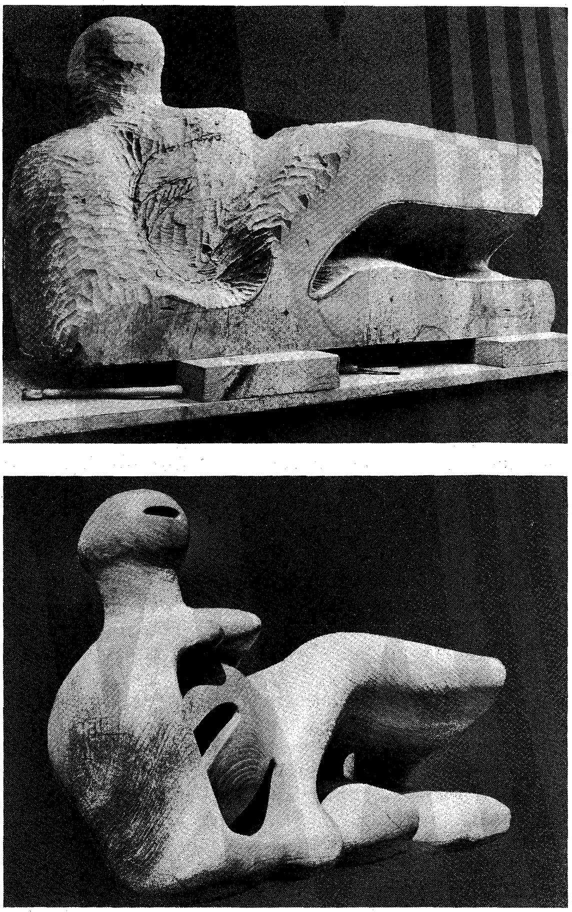
The main part of the sculptor's work, the shaping of the material itself by modelling, carving, or constructional techniques, may be a long and arduous process, perhaps extending over a number of years and requiring assistants. Much of the work, especially architectural decoration, may be carried out at the site, or in situ.

To improve its weathering qualities, to bring out the characteristics of its material to the best advantage, or to make it more decorative or realistic, sculpture is usually given a special surface finish. It may be rubbed down and polished, patinated, metal plated, gilded, painted, inlaid with other materials, and so on.

Installation of sculpture

Finally, the installation of sculpture may be a complex and important part of the work. The positioning and fixing of large architectural sculpture may involve cooperation with builders and engineers; fountains may involve elaborate plumbing; the design and placing of outdoor bases, or plinths, in relation to the site and the spectator may require careful thought. The choice of the materials, shape, and proportions of the base even for a small work requires a considerable amount of care.

Carving. Whatever material is used, the essential features of the direct method of carving are the same; the sculptor starts with a solid mass of material and reduces it systematically to the desired form. After he has blocked out the main masses and planes that define the outer limits of the forms, he works progressively over the whole sculpture, first carving the larger containing forms and planes and then the smaller ones until eventually the surface details are reached. Then he gives the surface whatever finish is required. Even with a preliminary model as a guide, the sculptor's concept constantly evolves and clarifies as the work proceeds; thus, as he adapts his design to the nature of the carving process and the material, his work develops as an organic whole.



Direct carving.
(Top) "Reclining Figure" in progress, elmwood sculpture by Henry Moore, 1945–46. (Bottom) "Reclining Figure." Length 1.90 m.

The process of direct carving imposes a characteristic order on the forms of sculpture. The faces of the original block, slab, or cylinder of material can usually still be sensed, existing around the finished work as a kind of implied spatial envelope limiting the extension of the forms in space and connecting their highest points across space. In a similar way, throughout the whole carving, smaller forms and planes can be seen as contained within implied larger ones. Thus, an ordered sequence of containing forms and planes, from the largest to the smallest, gives unity to the work.

Indirect carving. All of the great sculptural traditions of the past used the direct method of carving, but in Western civilization during the 19th and early 20th centuries it became customary for stone and, to a lesser extent, wood sculpture to be produced by the indirect method. This required the production of a finished clay model that was subsequently cast in plaster and then reproduced in stone or wood in a more or less mechanical way by means of a pointing machine (see *Reproduction and surface-finishing techniques: pointing* below). Usually, the carving was not done by the sculptor himself. At its worst, this procedure results in a carved copy of a design that was conceived in terms of clay modelling. Although indirect carving does not achieve aesthetic qualities that are typical of carved sculpture, it does not necessarily result in bad sculpture. Rodin's marble sculptures, for example, are generally considered great works of art even by those who object to the indirect methods by which they were produced. The indirect method has been steadily losing ground since the revival of direct carving in the early part of the 20th century, and today it is in general disrepute among carvers.

Tools for stone carving

Carving tools and techniques. The tools used for carving differ with the material to be carved. Stone is carved mostly with steel tools that resemble cold chisels. To knock off the corners and angles of a block, a tool called a pitcher is driven into the surface with a heavy iron hammer. The pitcher is a thick, chisel-like tool with a wide bevelled edge that breaks rather than cuts the stone. The heavy point then does the main roughing out, followed by the fine point, which may be used to within a short distance of the final surface. These pointed tools are hammered into the surface at an angle that causes the stone to break off in chips of varying sizes. Claw chisels, which have toothed edges, may then be worked in all directions over the surface, removing the stone in granule form and thus refining the surface forms. Flat chisels are used for finishing the surface carving and for cutting sharp detail. There are many other special tools, including stone gouges, drills, toothed hammers (known as bushhammers or bouchardes), and, often used today, power-driven pneumatic tools, for pounding away the surface of the stone.

Because medieval carvers worked mostly in softer stones and made great use of flat chisels, their work tends to have an edgy, cut quality and to be freely and deeply carved. In contrast is the work done in hard stones by people who lacked metal tools hard enough to cut the stone. Egyptian granite sculpture, for example, was produced mainly by abrasion; that is, by pounding the surface and rubbing it down with abrasive materials. The result is a compact sculpture, not deeply hollowed out, with softened edges and flowing surfaces. It usually has a high degree of tactile appeal.

Although the process of carving is fundamentally the same for wood or stone, the physical structure of wood demands tools of a different type. For the first blocking out of a wood carving a sculptor may use saws and axes, but his principal tools are a wide range of wood-carver's gouges. The sharp, curved edge of a gouge cuts easily through the bundles of fibre and when used properly will not split the wood. Flat chisels are also used, especially for carving sharp details. Wood rasps, or coarse files, and sandpaper can be used to give the surface a smooth finish, or, if preferred, it can be left with a faceted, chiselled appearance. Wood-carving tools have hardwood handles and are struck with round, wooden mallets. African wood sculptors use a variety of adzes rather than gouges and mallets. Ivory is carved with an assortment of saws, knives, rasps, files, chisels, drills, and scrapers.

Modelling. In contrast to the reductive process of carving, modelling is essentially a building-up process in which the sculpture grows organically from the inside. Numerous plastic materials are used for modelling. The main ones are clay, plaster, and wax; but concrete, synthetic resins, plastic wood, stucco, and even molten metal can also be modelled. A design modelled in plastic materials may be intended for reproduction by casting in more permanent and rigid materials, such as metal, plaster, concrete, and fibreglass, or it may itself be made rigid and more permanent through the self-setting properties of its materials (for example, plaster) or by firing.

Clay models

Modelling for casting. The material most widely used for making positive models for casting is clay. A small, compact design or a low relief can be modelled solidly in clay without any internal support; but a large clay model must be formed over a strong armature made of wood and metal. Since the armature may be very elaborate and can only be altered slightly, if at all, once work has started, the modeller must have a fairly clear idea from his drawings and maquettes of the arrangement of the main shapes of the finished model. The underlying main masses of the sculpture are built up firmly over the armature, and then the smaller forms, surface modelling, and details are modelled over them. The modeller's chief tools are his fingers, but for fine work he may use a variety of wooden modelling tools to apply the clay and wire loop tools to cut it away. Reliefs are modelled on a vertical or nearly vertical board. The clay is keyed, or secured, onto the board with galvanized nails or wood laths. The amount of armature required depends on the height of the relief and the weight of clay involved.

To make a cast in metal, a foundry requires from the sculptor a model made of a rigid material, usually plaster. The sculptor can produce this either by modelling in clay and then casting in plaster from the clay model or by modelling directly in plaster. For direct plaster modelling, a strong armature is required because the material is brittle. The main forms may be built up roughly over the armature in expanded wire and then covered in plaster-soaked scrim (a loosely woven sacking). This provides a hollow base for the final modelling, which is done by applying plaster with metal spatulas and by scraping and cutting down with rasps and chisels.

Lost-wax process bronze sculpture.
"Male Portrait Head," bronze from Ife, Nigeria, 12th century. In the Museum of Ife Antiquities. Height 34 cm.
Eliot Elisofon

Fibreglass and concrete sculptures are cast in plaster molds taken from the sculptor's original model. The model is usually clay rather than plaster because if the forms of the sculpture are at all complex it is easier to remove a plaster mold from a soft clay model than from a model in a rigid material, such as plaster.

A great deal of the metal sculpture of the past, including Nigerian, Indian, and many Renaissance bronzes, was produced by the direct lost-wax process, which involves a special modelling technique (see *Reproduction and surface-finishing techniques: casting and molding* below). The design is first modelled in some refractory material to within a fraction of an inch of the final surface, and then the final modelling is done in a layer of wax, using the fingers and also metal tools, which can be heated to make the wax more pliable. Medallions are often produced from wax originals, but because of their small size they do not require a core.

Modelling for pottery sculpture. To withstand the stresses of firing, a large pottery sculpture must be hollow and of an even thickness. There are two main ways of achieving this. In the process of hollow modelling, which is typical of the potter's approach to form, the main forms of the clay model are built up directly as hollow forms with walls of a roughly even thickness. The methods of building are similar to those employed for making hand-built pottery—coiling, pinching, and slabbing. The smaller forms and details are then added, and the finished work is allowed to dry out slowly and thoroughly before firing. The process of solid modelling is more typical of the sculptor's traditional approach to form. The sculpture is modelled in solid clay, sometimes over a carefully considered armature, by the sculptor's usual methods of clay modelling. Then it is cut open and hollowed out, and the armature, if there is one, is removed. The pieces are then rejoined and the work is dried out and fired.

Effects of modelling on the design of sculpture

General characteristics of modelled sculpture. The process of modelling affects the design of sculpture in three important ways. First, the forms of the sculpture tend to be ordered from the inside. There are no external containing forms and planes, as in carved sculpture. The overall design of the work—its main volumes, proportions, and axial arrangement—is determined by the underlying forms; and the smaller forms, surface modelling, and decorative details are all formed around and sustained by this underlying structure. Second, because its extension into space is not limited by the dimensions of a block of material, modelled sculpture tends to be much freer and more expansive in its spatial design than carved sculpture. If the tensile strength of metal is to be exploited in the finished work, there is almost unlimited freedom; designs for brittle materials such as concrete or plaster are more limited. Third, the plasticity of clay and wax encourages a fluent, immediate kind of manipulation, and many sculptors, such as Auguste Rodin, Giacomo Manzù, and Sir Jacob Epstein, like to preserve this record of their direct handling of the medium in their finished work. Their approach contrasts with that of the Benin and Indian bronze sculptors, who refined the surfaces of their work to remove all traces of personal "handwriting."

Constructing and assembling. A constructed or assembled sculpture is made by joining preformed pieces of material. It differs radically in principle from carved and modelled sculpture, both of which are fabricated out of a homogeneous mass of material. Constructed sculpture is made out of such basic preformed components as metal tubes, rods, plates, bars, and sheets; wooden laths, planks, dowels, and blocks; laminated timbers and chipboards; sheets of Perspex, Formica, and glass; fabrics; and wires and threads. These are cut to various sizes and may be either shaped before they are assembled or used as they are. The term assemblage is usually reserved for constructed sculpture that incorporates any of a vast array of ready-made, so-called found objects, such as old boilers, typewriters, engine components, mirrors, chairs, and table legs and other bits of old furniture. Numerous techniques are employed for joining these components, most of them derived from crafts other than traditional sculptural ones; for example, metal welding and brazing, wood joinery, bolting, screwing, rivetting, nailing, and bonding with new powerful adhesives.

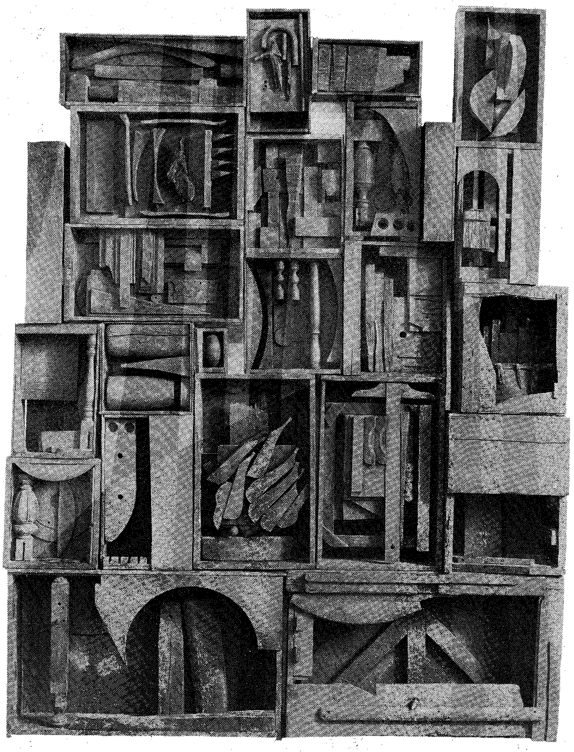
The use of constructional techniques for the production of sculpture is the major technical development of the art of sculpture in recent years. Among the reasons for its popularity are that it lends itself readily to an emphasis on the spatial aspects of sculpture that have preoccupied so many 20th-century artists; it is quicker than carving and modelling; it is considered by many sculptors and critics to be especially appropriate to a technological civilization; it is opening up new fields of imagery and new types of symbolism and form.

For constructed "gallery" sculpture, almost any materials and techniques are likely to be used, and the products are often extremely ephemeral. But architectural sculpture, outdoor sculpture, and indeed any sculpture that is actually used must be constructed in a safe and at least reasonably permanent manner. The materials and techniques employed are therefore somewhat restricted. Metal sculpture constructed by rivetting, bolting, and, above all, welding and brazing is best for outdoor use.

Direct metal sculpture. The introduction of the oxyacetylene welding torch as a sculptor's tool has revolutionized metal sculpture in recent years. A combination of welding and forging techniques was pioneered by the Spanish sculptor Julio González around 1930; and during the 1940s and 1950s it became a major sculptural technique, particularly in Britain and in the United States, where its greatest exponent was David Smith. In the 1960s and early 1970s, more sophisticated electric welding processes were replacing flame welding.

Welding
and
forging

Welding equipment can be used for joining and cutting metal. A welded joint is made by melting and fusing together the surfaces of two pieces of metal, usually with the addition of a small quantity of the same metal as a filler. The metal most widely used for welded sculpture is mild steel, but other metals can be welded. In a brazed joint, the parent metals are not actually fused together but are joined by an alloy that melts at a lower temperature than the parent metals. Brazing is particularly useful for making joints between different kinds of metal, which cannot be done by welding, and for joining nonferrous metals. Forging is the direct shaping of metal by bending, hammering, and cutting.



Assemblage.
"Black Wall," wood sculpture by Louise Nevelson, 1959. In the Tate Gallery, London. 2.84 × 2.16 × 0.65 m.
By courtesy of the trustees of the Tate Gallery, London

Direct metalworking techniques have opened up whole new ranges of form to the sculptor—open skeletal structures, linear and highly extended forms, and complex, curved sheet forms. Constructed metal sculpture may be precise and clean, as that of Minimalist sculptors Donald Judd and Phillip King, or it may exploit the textural effects of molten metal in a free, "romantic" manner.

REPRODUCTION AND SURFACE-FINISHING TECHNIQUES

Casting and molding processes are used in sculpture either for making copies of existing sculpture or as essential stages in the production of a finished work. Numerous materials are used for making molds and casts, and some of the methods are complex and highly skilled. Only a broad outline of the principal methods can be given here.

Casting and molding. These are used for producing a single cast from a soft, plastic original, usually clay. They are especially useful for producing master casts for subsequent reproduction in metal. The basic procedure is as follows. First, the mold is built up in liquid plaster over the original clay model; for casting reliefs, a one-piece mold may be sufficient, but for sculpture in the round a mold in at least two sections is required. Second, when the plaster is set, the mold is divided and removed from the clay model. Third, the mold is cleaned, reassembled, and filled with a self-setting material such as plaster, concrete, or fibreglass-reinforced resin. Fourth, the mold is carefully chipped away from the cast. This involves the destruction of the mold—hence the term "waste" mold. The order of reassembling and filling the mold may be reversed; fibreglass and resin, for example, are "laid up" in the mold pieces before they are reassembled.

Plaster piece molds are used for producing more than one cast from a soft or rigid original and are especially good for reproducing existing sculpture and for slip casting (see below). Before the invention of flexible molds (see below), piece molds were used for producing wax casts for metal casting by the lost-wax process. A piece mold is built up in sections that can be withdrawn from the original model without damaging it. The number of sections depends on the complexity of the form and on the amount of undercutting; tens, or even hundreds, of pieces may be required for really large, complex works. The mold sections are carefully keyed together and supported by a plaster case. When the mold has been filled, it can be removed section by section from the cast and used again. Piece molding is a highly skilled and laborious process.

Made of such materials as gelatin, vinyl, and rubber, flexible molds are used for producing more than one cast; they offer a much simpler alternative to piece molding when the original model is a rigid one with complex forms and undercuts. The material is melted and poured around the original positive in sections, if necessary. Being flexible, the mold easily pulls away from a rigid surface without causing damage. While it is being filled (with wax, plaster, concrete, and fibreglass-reinforced resins), the mold must be surrounded by a plaster case to prevent distortion.

The traditional method of casting metal sculpture

The lost-wax process is the traditional method of casting metal sculpture. It requires a positive, which consists of a core made of a refractory material and an outer layer of wax. The positive can be produced either by direct modelling in wax over a prepared core, in which case the process is known as direct lost-wax casting, or by casting in a piece mold or flexible mold taken from a master cast. The wax positive is invested with a mold made of refractory materials and is then heated to a temperature that will drive off all moisture and melt out all the wax, leaving a narrow cavity between the core and the investment. Molten metal is then poured into this cavity. When the metal has cooled down and solidified, the investment is broken away, and the core is removed from inside the cast. The process is, of course, much more complex than this simple outline suggests. Care has to be taken to suspend the core within the mold by means of metal pins, and a structure of channels must be made in the mold that will enable the metal to reach all parts of the cavity and permit the mold gases to escape. A considerable amount of filing and chasing of the cast is usually required after casting is completed.

While the lost-wax process is used for producing complex, refined metal castings, sand molding is more suitable for simpler types of form and for sculpture in which a certain roughness of surface does not matter. Recent improvements in the quality of sand castings and the invention of the "lost-pattern" process (see below) have resulted in a much wider use of sand casting as a means of producing sculpture. A sand mold, made of special sand held together by a binder, is built up around a rigid positive, usually in a number of sections held together in metal boxes. For a hollow casting, a core is required that will fit inside the negative mold, leaving a narrow cavity as in the lost-wax process. The molten metal is poured into this cavity.

The lost-pattern process is used for the production by sand molding of single casts in metal. After a positive made of expanded polystyrene is firmly embedded in casting sand, molten metal is poured into the mold straight onto the expanded foam original. The heat of the metal causes the foam to pass off into vapour and disappear, leaving a negative mold to be filled by the metal. Channels for the metal to run in and for the gases to escape are made in the mold, as in the lost-wax process. The method is used mainly for producing solid castings in aluminum that can be welded or rivetted together to make the finished sculpture.

Slip casting is primarily a potter's technique that can be used for repetition casting of small pottery sculptures. Liquid clay, or slip, is poured into a plaster piece mold. Some of the water in the slip is absorbed by the plaster and a layer of stiffened clay collects on the surface of the mold. When this layer is thick enough to form a cast, the excess slip is poured off and the mold is removed. The hollow clay cast is then dried and fired.

Simple casts for pottery sculpture—mainly tiles and low reliefs—can be prepared by pressing clay into a rigid mold. More complex forms can be built up from a number of separately press-cast pieces. Simple terra-cotta molds can be made by pressing clay around a rigid positive form. After firing, these press molds can be used for press casting.

Pointing. A sculpture can be reproduced by transposing measurements taken all over its surface to a copy. The process is made accurate and thorough by the use of a pointing machine, which is an arrangement of adjustable metal arms and pointers that are set to the position of any point on the surface of a three-dimensional form and then used to locate the corresponding point on the surface of a copy. If the copy is a stone one, the block is drilled to the depth measured by the pointing machine. When a number of points have been fixed by drilling, the stone is cut away to the required depth. For accurate pointing, a vast number of points have to be taken, and the final surface is approached gradually. The main use of pointing has been for the indirect method of carving.

Enlarged and reduced copies

Enlarged and reduced copies of sculpture can also be produced with the aid of mechanical devices. A sophisticated reducing machine that works on the principle of the pantograph (an instrument for copying on any predetermined scale, consisting of four light, rigid bars jointed in parallelogram form) is used in minting for scaling down the sculptor's original model to coin size.

Surface finishing. Surface finishes for sculpture can be either natural—bringing the material of the sculpture itself to a finish—or applied. Almost all applied surface finishes preserve as well as decorate.

Smoothing and polishing. Many sculptural materials have a natural beauty of colour and texture that can be brought out by smoothing and polishing. Stone carvings are smoothed by rubbing down with a graded series of coarse and fine abrasives, such as carborundum, sandstone, emery, pumice, and whiting, all used while the stone is wet. Some stones, such as marble and granite, will take a high gloss; others are too coarse-grained to be polished and can only be smoothed to a granular finish. Wax is sometimes used to give stone a final polish.

The natural beauty of wood is brought out by sandpapering or scraping and then waxing or oiling. Beeswax and linseed oil are the traditional materials, but a wide range of waxes and oils is currently available.

Ivory is polished with gentle abrasives such as pumice and whiting, applied with a damp cloth.

Concrete can be rubbed down, like stone, with water and abrasives, which both smooth the surface and expose the aggregate. Some concretes can be polished.

Metals are rubbed down manually with steel wool and emery paper and polished with various metal polishes. A high-gloss polish can be given to metals by means of power-driven buffing wheels used in conjunction with abrasives and polishes. Clear lacquers are applied to preserve the polish.

Polychrome sculpture

Painting. Stone, wood, terra-cotta, metal, fibreglass, and plaster can all be painted in a reasonably durable manner provided that the surfaces are properly prepared and suitable primings and paints are used. In the past, stone and wood carvings were often finished with a coating of gesso (plaster of Paris or gypsum prepared with glue) that served both as a final modelling material for delicate surface detail and as a priming for painting. Historically, the painting and gilding of sculpture were usually left to specialists. In Greek relief sculpture, actual details of the composition were often omitted at the carving stage and left for the painter to insert. In the 15th century, the great Flemish painter Rogier van der Weyden undertook the painting of sculpture as part of his work.

Modern paint technology has made an enormous range of materials available. Constructed sculptures are often finished with mechanical grinders and sanders and then sprayed with high-quality cellulose paints.

Gilding. The surfaces of wood, stone, and plaster sculpture can be decorated with gold, silver, and other metals that are applied in leaf or powder form over a suitable priming. Metals, especially bronze, were often fire-gilded; that is, treated with an amalgam of gold and mercury that was heated to drive off the mercury. The panels of the "Gates of Paradise" in Florence, by the 15th-century sculptor Lorenzo Ghiberti, are a well-known example of gilded bronze.

Patinas achieved artificially

Patination. Patinas on metals are caused by the corrosive action of chemicals. Sculpture that is exposed to different kinds of atmosphere or buried in soil or immersed in seawater for some time acquires a patina that can be extremely attractive. Similar effects can be achieved artificially by applying various chemicals to the metal surface.

This is a particularly effective treatment for bronze, which can be given a wide variety of attractive green, brown, blue, and black patinas. Iron is sometimes allowed to rust until it acquires a satisfactory colour, and then the process is arrested by lacquering.

Electroplating. The surfaces of metal sculpture or of specially prepared nonmetal sculpture can be coated with such metals as chrome, silver, gold, copper, and nickel by the familiar industrial process of electroplating. The related technique of anodizing can be used to prevent the corrosion of aluminum sculpture and to dye its surface.

Other finishes. The surfaces of metal sculpture can be decorated by means of numerous metalsmithing techniques—etching, engraving, metal inlaying, enamelling, and so on. Pottery sculpture can be decorated with coloured slips, oxides, and enamels; glazed with a variety of shiny or mat glazes; and brought to a dull polish by burnishing.

Other materials have often been added to the surface of sculpture. The eyes of ancient figure sculpture, for example, were sometimes inlaid with stones. Occasionally—as in Mexican mosaic work—the whole surface of a sculpture is inlaid with mother-of-pearl, turquoise, coral, and many other substances.

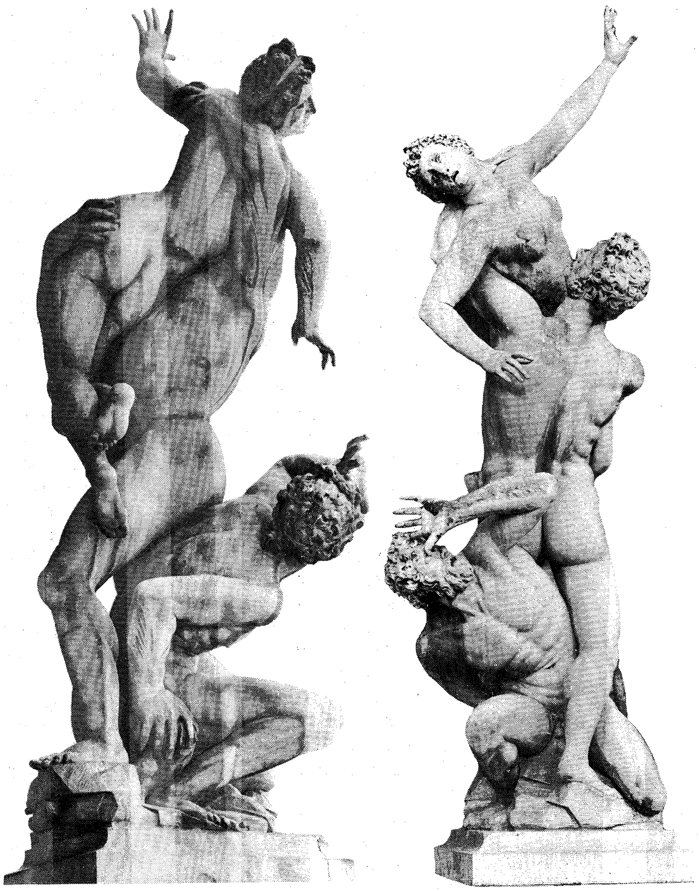
Forms, subject matter, imagery, and symbolism of sculpture

A great deal of sculpture is designed to be placed in public squares, gardens, parks, and similar open places or in interior positions where it is isolated in space and can be viewed from all directions. Other sculpture is carved in relief and is viewed only from the front and sides.

SCULPTURE IN THE ROUND

The opportunities for free spatial design that such freestanding sculpture presents are not always fully exploited. The work may be designed, like many Archaic sculptures, to be viewed from only one or two fixed positions, or it may in effect be little more than a four-sided relief that hardly changes the three-dimensional form of the block at all. Sixteenth-century Mannerist sculptors, on the other hand, made a special point of exploiting the all-around visibility of freestanding sculpture. Giambologna's "Rape of the Sabines," for example, compels the viewer to walk all around it in order to grasp its spatial design. It has no principal views; its forms move around the central axis of the composition, and their serpentine movement unfolds itself gradually as the spectator moves around to follow them. Much of the sculpture of Henry Moore and other 20th-century sculptors is not concerned with movement of this kind, nor is it designed to be viewed from any fixed positions. Rather, it is a freely designed structure of multidirectional forms that is opened up, pierced, and extended in space in such a way that the viewer is made aware of its all-around design largely by seeing through the sculpture. The majority of constructed sculptures are disposed in space with complete freedom and invite viewing from all directions. In many instances the spectator can actually walk under and through them.

The way in which a freestanding sculpture makes contact with the ground or with its base is a matter of considerable importance. A reclining figure, for example, may in effect be a horizontal relief. It may blend with the ground plane and appear to be rooted in the ground like an outcrop of rock. Other sculptures, including some reclining figures, may be designed in such a way that they seem to rest on the ground and to be independent of their base. Others are supported in space above the ground. The most completely freestanding sculptures are those that have no base and may be picked up, turned in the hands, and literally viewed all around like a netsuke (a small toggle of wood, ivory, or metal used to fasten a small pouch or purse to a kimono sash). Of course, a large sculpture cannot actually be picked up in this way, but it can be designed so as to invite the viewer to think of it as a detached, independent object that has no fixed base and is designed all around.



Sculpture in the round.
The changing appearance of freestanding sculpture observed when walking around a statue is suggested by these two views of "Rape of the Sabines," marble sculpture by Giambologna, 1583. In the Loggia dei Lanzi, Florence. Height 411 cm.
Alinari—Art Resource

Sculpture designed to stand against a wall or similar background or in a niche may be in the round and freestanding in the sense that it is not attached to its background like a relief; but it does not have the spatial independence of completely freestanding sculpture, and it is not designed to be viewed all around. It must be designed so that its formal structure and the nature and meaning of its subject matter can be clearly apprehended from a limited range of frontal views. The forms of the sculpture, therefore, are usually spread out mainly in a lateral direction rather than in depth. Greek pedimental sculpture illustrates this approach superbly: the composition is spread out in a plane perpendicular to the viewer's line of sight and is made completely intelligible from the front. Seventeenth-century Baroque sculptors, especially Bernini, adopted a rather different approach. There may be considerable recession and foreshortening in their compositions, but the forms are carefully arranged so that they present a coherent and intelligible whole from one special frontal viewpoint.

Difference between wall and niche and freestanding sculpture

The frontal composition of wall and niche sculpture does not necessarily imply any lack of three-dimensionality in the forms themselves; it is only the arrangement of the forms that is limited. Classical pedimental sculpture, Indian temple sculpture such as that at Khajurāho, Gothic niche sculpture, and Michelangelo's Medici tomb figures are all designed to be placed against a background, but their forms are conceived with a complete fullness of volume.

RELIEF SCULPTURE

Relief sculpture is a complex art form that combines many features of the two-dimensional pictorial arts and the three-dimensional sculptural arts. On the one hand, a relief, like a picture, is dependent on a supporting surface, and its composition must be extended in a plane in order to be visible. On the other hand, its three-dimensional properties are not merely represented pictorially but are in some degree actual, like those of fully developed sculpture.

Among the various types of relief are some that approach very closely the condition of the pictorial arts. The reliefs of Donatello, Ghiberti, and other early Renaissance artists make full use of perspective, which is a pictorial method of representing three-dimensional spatial relationships realistically on a two-dimensional surface. Egyptian and most pre-Columbian American low reliefs are also extremely pictorial but in a different way. Using a system of graphic conventions, they translate the three-dimensional world into a two-dimensional one. The relief image is essentially one of plane surfaces and could not possibly exist in three dimensions. Its only sculptural aspects are its slight degree of actual projection from a surface and its frequently subtle surface modelling.

Other types of relief—for example, Classical Greek and most Indian—are conceived primarily in sculptural terms. The figures inhabit a space that is defined by the solid forms of the figures themselves and is limited by the background plane. This back plane is treated as a finite, impenetrable barrier in front of which the figures exist. It is not conceived as a receding perspective space or environment within which the figures are placed nor as a flat surface upon which they are placed. The reliefs, so to speak, are more like contracted sculpture than expanded pictures.

The central problem of relief sculpture is to contract or condense three-dimensional solid form and spatial relations into a limited depth space. The extent to which the forms actually project varies considerably, and reliefs are classified on this basis as low reliefs (bas-reliefs) or high reliefs. There are types of reliefs that form a continuous series from the almost completely pictorial to the almost fully in the round.

Representation of forms in depth

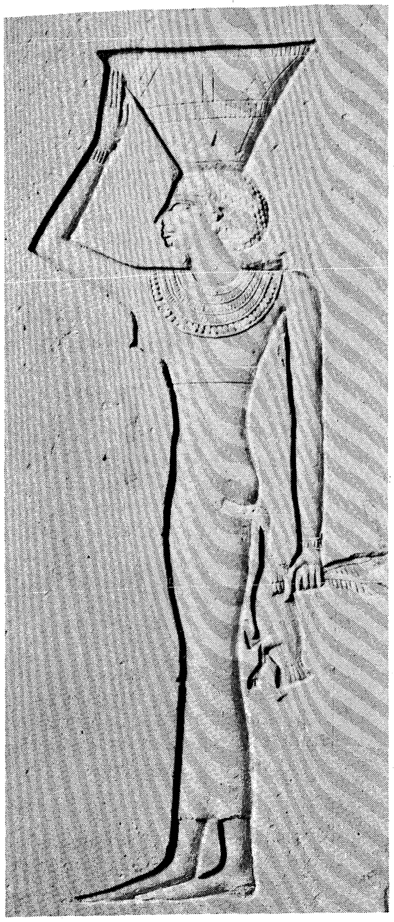
One of the relief sculptor's most difficult tasks is to represent the relations between forms in depth within the limited space available to him. He does this mainly by giving careful attention to the planes of the relief. In a carved relief the highest, or front, plane is defined by the surface of the slab of wood or stone in which the relief is carved; and the back plane is the surface from which the forms project. The space between these two planes can be thought of as divided into a series of planes, one behind the other. The relations of forms in depth can then be thought of as relations between forms lying in different planes.

Sunken relief is also known as incised, coelanaglyphic, and intaglio relief. It is almost exclusively an ancient Egyptian art form, but some beautiful small-scale Indian examples in ivory have been found at Bagrām in Afghanistan. In a sunken relief, the outline of the design is first incised all around. The relief is then carved inside the incised outline, leaving the surrounding surface untouched. Thus, the finished relief is sunk below the level of the surrounding surface and is contained within a sharp, vertical-walled contour line. This approach to relief preserves the continuity of the original surface and creates no projection from it. The outline shows up as a powerful line of light and shade around the whole design.

Figurative low relief is generally regarded by sculptors as an extremely difficult art form. To give a convincing impression of three-dimensional structure and surface modelling with only a minimal degree of projection demands a fusion of draftsmanship and carving or modelling skill of a high order. The sculptor has to proceed empirically, constantly changing the direction of his light and testing the optical effect of his work. He cannot follow any fixed rules or represent things in depth by simply scaling down measurements mathematically, so that, say, one inch of relief space represents one foot.

The forms of low relief usually make contact with the background all around their contours. If there is a slight amount of undercutting, its purpose is to give emphasis, by means of cast shadow, to a contour rather than to give any impression that the forms are independent of their background. Low relief includes figures that project up to about half their natural circumference.

Technically, the simplest kind of low relief is the two-plane relief. For this, the sculptor draws an outline on a surface and then cuts away the surrounding surface, leaving the figure raised as a flat silhouette above the background plane. This procedure is often used for the first stages of a full relief carving, in which case the sculptor will proceed to carve into the raised silhouette, rounding the forms and giving an impression of three-dimensional structure. In a two-plane relief, however, the silhouette is left flat and substantially unaltered except for the addition of surface detail. Pre-Columbian sculptors used this method of relief carving to create bold figurative and abstract reliefs.



Sunken relief.
Egyptian woman carrying food for the dead, detail of an Old Kingdom limestone relief originally from the tomb of Thetha at Thebes, *c.* 2300 BC. In the British Museum. Figure height 43 cm.
By courtesy of the trustees of the British Museum

Stiacciato relief

Stiacciato relief is an extremely subtle type of flat, low relief carving that is especially associated with the 15th-century sculptors Donatello and Desiderio da Settignano. The design is partly drawn with finely engraved chisel lines and partly carved in relief. The stiacciato technique depends largely for its effect on the way in which pale materials, such as white marble, respond to light and show up the most delicate lines and subtle changes of texture or relief.

The forms of high relief project far enough to be in some degree independent of their background. As they approach the fullness of sculpture in the round, they become of necessity considerably undercut. In many high reliefs, where parts of the composition are completely detached from their background and fully in the round, it is often impossible to tell from the front whether or not a figure is actually attached to its background.

Many different degrees of projection are often combined in one relief composition. Figures in the foreground may be completely detached and fully in the round, while those in the middle distance are in about half relief and those in the background in low relief. Such effects are common in late Gothic, Renaissance, and Baroque sculpture.

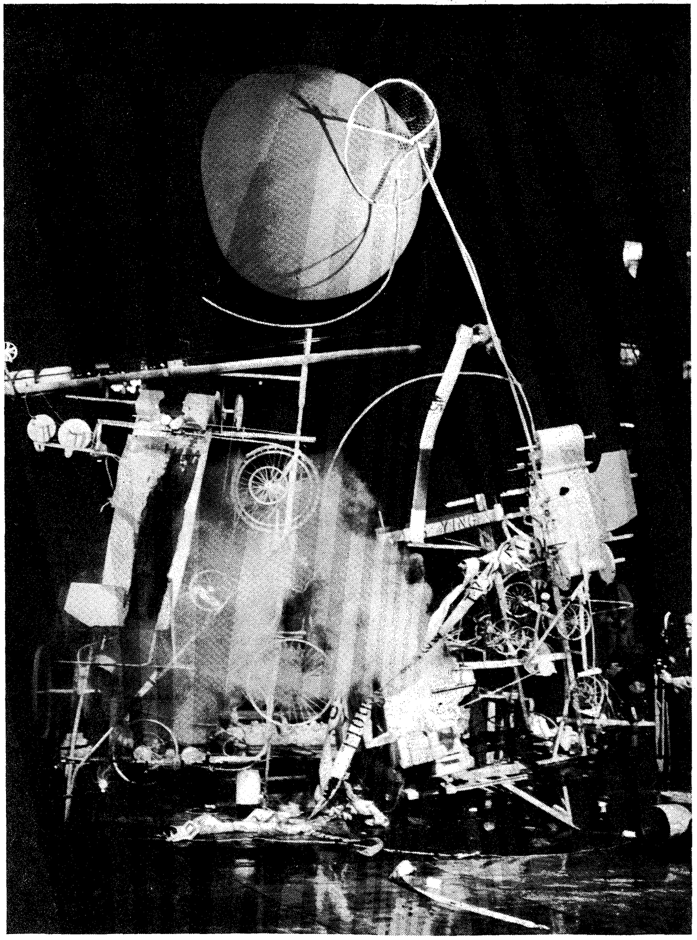
MODERN FORMS OF SCULPTURE

Since the 1950s, many new combined forms of art have been developed that do not fit readily into any of the traditional categories. Two of the most important of these, environments and kinetics, are closely enough connected with sculpture to be regarded by many artists and critics as branches or offshoots of sculpture. It is likely, however, that the persistence of the terms environmental sculpture and kinetic sculpture is a result of the failure of language to keep pace with events; for the practice is already growing of referring simply to environments and kinetics, as one might refer to painting, sculpture, and engraving, as art forms in their own right.

Traditional sculptures in relief and in the round are static, fixed objects or images. Their immobility and immutability are part of the permanence traditionally associated with the art of sculpture, especially monumental sculpture. What one refers to as movement in, say, a Baroque or Greek sculpture is not actual physical motion but a movement that is either directly represented in the subject matter (galloping horses) or expressed through the dynamic character of its form (spirals, undulating curves). In recent years, however, the use of actual movement, kineticism, has become an important aspect of sculpture. Naum Gabo, Marcel Duchamp, László Moholy-Nagy, and Alexander Calder were pioneers of kinetic sculpture in modern times, but many kinetic artists see a connection between their work and such forms as the moving toys, dolls, and clocks of previous ages.

There are now types of sculpture in which the components are moved by air currents, as in the well-known mobiles of Calder; by water; by magnetism, the speciality of Nicholus Takis; by a variety of electromechanical devices; or by the participation of the spectator himself. The neo-Dada satire quality of the kinetic sculpture created during the 1960s is exemplified by the works of Jean Tinguely. His self-destructing "Homage to New York" perfected the concept of a sculpture being both an object and an event, or "happening."

By courtesy of the Museum of Modern Art, New York; photograph, David Gahr



Kinetic sculpture.
"Homage to New York," a self-constructing and self-destroying work of art by Jean Tinguely, 1960. Small pieces now in the Museum of Modern Art, New York.

The aim of most kinetic sculptors is to make movement itself an integral part of the design of the sculpture and not merely to impart movement to an already complete static object. Calder's mobiles, for example, depend for their aesthetic effect on constantly changing patterns of relationship. When liquids and gases are used as components, the shapes and dimensions of the sculpture may undergo continual transformations. The movement of smoke; the diffusion and flow of coloured water, mercury, oil, and so on; pneumatic inflation and deflation; and the movement of masses of bubbles have all served as media for kinetic sculpture. In the complex, electronically controlled "spatio-dynamic" and "lumino-dynamic" constructions of Nicolas Schöffer, the projection of changing patterns of light into space is a major feature.

The environmental sculptor creates new spatial contexts that differ from anything developed by traditional sculpture. The work no longer confronts the spectator as an object but surrounds him so that he moves within it as he might within a stage set, a garden, or an interior. The most common type of environment is the "room," which may have specially shaped and surfaced walls, special lighting effects, and many different kinds of contents. Kurt Schwitters' *Merzbau* (destroyed in 1943) was the first of these rooms, which now include the nightmare fantasy of Edward Kienholz' tableaux, such as "Roxy's" (1961) or "The Illegal Operation" (1962); George Segal's compositions, in which casts of clothed human figures in frozen, casual attitudes are placed in interiors; and rooms built of mirrors, such as Yayoi Kusama's "Endless Love Room" and Lucas Samaras' "Mirrored Room," in both of which the spectator himself, endlessly reflected, becomes part of the total effect.

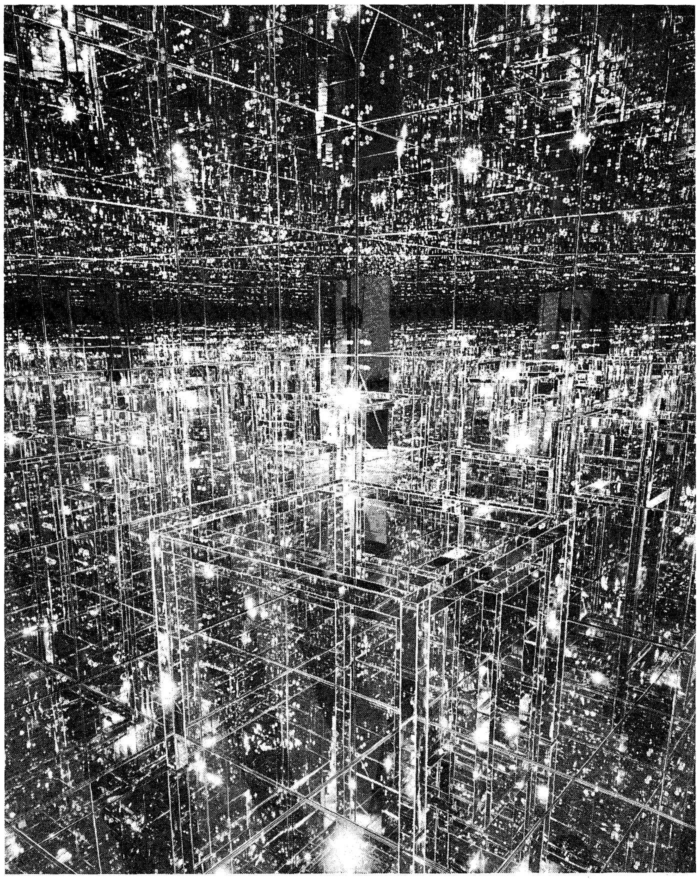
Use of real objects in environmental art

Environmental art, in common with collage and assemblage, has tended toward greater concreteness not by making a more realistic representation, as naturalistic art does, but by including more of reality itself in the work; for example, by using casts taken from the actual human body, real clothes, actual objects and casts of objects, actual lighting effects, and real items of furniture. Plastic elements may be combined with music and sound effects, dance, theatrical spectacles, and film to create so-called happenings, in which real figures are constituents of the "artwork" and operations are performed not on "artistic" materials but are performed on real objects and on the actual environment. Ideas such as these go far beyond anything that has ever before been associated with the term sculpture.

REPRESENTATIONAL SCULPTURE

Sculpture in the round is much more restricted than relief in the range of its subject matter. The representation of, say, a battle scene or a cavalcade in the round would require a space that corresponded in scale in every direction with that occupied by an actual battle or cavalcade. No such problems arise in relief because the treatment of scale and relations in depth is to some extent notional, or theoretical, like that of pictures. Then again, because a relief is attached to a background, problems of weight and physical balance and support do not arise. Figures can be represented as floating in space and can be arranged vertically as well as horizontally. Thus, in general, sculpture in the round is concerned with single figures and limited groups, while reliefs deal with more complex "pictorial" subjects involving crowds, landscape, architectural backgrounds, and so on.

The human figure. The principal subject of sculpture has always been the human figure. Next in importance are animals and fantastic creatures based on human and animal forms. Other subjects—for example, landscape, plants, still life, and architecture—have served primarily as accessories to figure sculpture, not as subjects in their own right. The overwhelming predominance of the human figure is due: first, to its immense emotional importance



Environmental sculpture.
"Mirrored Room," mirror on wood by Lucas Samaras, 1966. In the Albright-Knox Art Gallery, Buffalo. 305 × 244 cm.
Albright-Knox Art Gallery, Buffalo, gift of Seymour H. Knox; photograph, by courtesy of the Pace Gallery, New York

as an object of desire, love, fear, respect, and, in the case of anthropomorphic gods, worship; and, second, to its inexhaustible subtlety and variety of form and expression. The nude or almost nude figure played a prominent role in Egyptian, Indian, Greek, and African sculpture, while in medieval European and ancient Chinese sculpture the figure is almost invariably clothed. The interplay of the linear and modelled forms of free draperies with the solid volumes of the human body was of great interest to Classical sculptors and later became one of the principal themes of Renaissance and post-Renaissance sculpture. The human figure continues to be of central importance in modern sculpture in spite of the growth of nonfigurative art; but the optimistic, idealized, or naturalistic images of man prevalent in previous ages have been largely replaced by images of despair, horror, deformation, and satire.

Devotional images and narrative sculpture. The production of devotional images has been one of the sculptor's main tasks, and many of the world's greatest sculptures are of this kind. They include images of Buddha and the Hindu gods; of Christ, the Virgin, and the Christian saints; of Athena, Aphrodite, Zeus, and other Greek gods; and of all the various gods, spirits, and mythical beings of Rome, the ancient Near East, pre-Columbian America, Black Africa, and the Pacific Islands.

Closely connected with devotional images are all of the narrative sculptures in which legends, heroic deeds, and religious stories are depicted for the delight and instruction of peoples who lived when books and literacy were rare. The Buddhist, Hindu, and Christian traditions are especially rich in narrative sculpture. Stories of the incarnations of Buddha—*Jātaka*—and of the Hindu gods abounded in the temple sculpture of India and Southeast Asia; for example, at Sānchi, Amarāvatī, Borobuḍur, and Angkor. Sculpture illustrating the stories of the Bible is so abundant in medieval churches that the churches have been called "Bibles in stone." Sculpture recounting the heroic deeds of kings and generals are common, especially in Assyria and Rome. The Romans made use of a form known as continuous narrative, the best known example of which is the spiral, or helical, band of relief sculpture that surrounds Trajan's Column (*c.* AD 106–113) and tells the story of the Emperor's Dacian Wars. The episodes in the narrative are not separated into a series of framed compositions but are linked to form a continuous band of unbroken relief.

Portraiture. Portraiture was practiced by the Egyptians but was comparatively rare in the ancient world until the Romans made portrait sculpture one of their major artistic achievements. The features of many famous people are known to modern man only through the work of Roman sculptors on coins and medals, portrait busts, and full-length portraits. Portraiture has been an important aspect of Western sculpture from the Renaissance to the present day. Some of the best known modern portrait sculptors are Rodin, Charles Despiau, Marino Marini, and Jacob Epstein.

Modern portrait sculptors

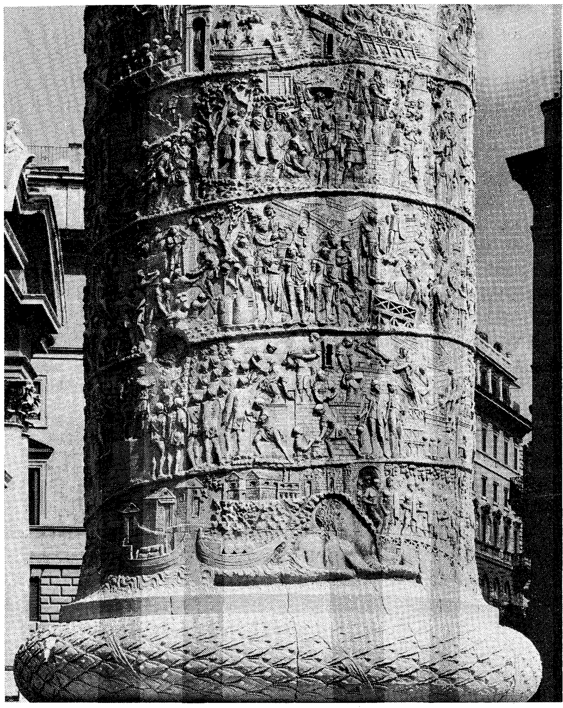
Scenes of everyday life. Scenes of everyday life have been represented in sculpture mainly on a small scale in minor works. The sculptures that are closest in spirit to the quiet dignity of the great 17th- and 18th-century genre paintings of Jan Vermeer and Jean-Baptiste-Siméon Chardin are perhaps certain Greek tombstones, such as that of the Stele of Hegeso, which represents a quiet, absorbed moment when a seated young woman and her maidservant are looking at a necklace they have just removed from a casket. Intimate scenes of the people and their activities in everyday rural life are often portrayed in medieval and Egyptian reliefs as part of larger compositions.

Animals. Animals have always been important subjects for sculpture. Paleolithic man produced some extraordinarily sensitive animal sculptures both in relief and in the round. Representations of horses and lions are among the finest works of Assyrian sculpture. Egyptian sculptors produced sensitive naturalistic representations of cattle, donkeys, hippopotamuses, apes, and a wide variety of birds and fish. Ancient Chinese sculptors made superb small-scale animal sculptures in bronze and pottery. Animals were the main subject matter for the sculpture of the nomadic tribes of Eurasia and northern Europe, for whom they became the basis for elaborate zoomorphic fantasies. This animal art contributed to the rich tradition of animal sculpture in medieval art. Animals also served as a basis for semi-abstract fantasy in Mexican, Mayan, North American Indian, and Oceanic sculpture. The horse has always occupied an important place in Western sculpture, but other animals have also figured in the work of such sculptors as Giambologna, in the 16th century, and Antoine-Louis Barye, in the 19th, as well as numerous sculptors of garden and fountain pieces. Among modern sculptors who have made extensive use of animals or animal-like forms are Brancusi, Picasso, Gerhard Marcks, Germaine Richier, François Pompon, Pino Pascali, and François-Xavier-Maxime Lalanne.

Fantasy. In their attempts to imagine gods and mythical beings, sculptors have invented fantastic images based on the combination and metamorphosis of animal and human forms. A centaur, the Minotaur, and animal-headed gods of the ancient world are straightforward combinations. More imaginative fantasies were produced by Mexican and Mayan sculptors and by tribal sculptors in many parts of the world. Fantastic creatures abound in the sculpture produced in northern Europe during the early Middle Ages and the Romanesque period. Fantasy of a playful kind is often found in garden sculpture and fountains.

In the period following World War I, fantasy was a dominant element in representational sculpture. Among its many forms are images derived from dreams, the technological fantasy of science fiction, erotic fantasies, and a whole host of monsters and automata. The Surrealists have made a major contribution to this aspect of modern sculpture.

Other subjects. Architectural backgrounds in sculpture range from the simplified baldacchinos (ornamental struc-



Narrative sculpture.
Lower portion of Trajan's Column, marble, Rome, *c.* AD 106–113. Height of one relief band about 127 cm.
GEKS

tures resembling canopies used especially over altars) of early medieval reliefs to the 17th- and 18th-century virtuoso perspective townscapes of Grinling Gibbons. Architectural accessories such as plinths, entablatures, pilasters, columns, and moldings have played a prominent role both in Greek and Roman sarcophagi, in medieval altarpieces and screens, and in Renaissance wall tombs.

Outside the field of ornament, botanical forms have played only a minor role in sculpture. Trees and stylized lotuses are especially common in Indian sculpture because of their great symbolic significance. Trees are also present in many Renaissance reliefs and in some medieval reliefs.

Landscape, which was an important background feature in many Renaissance reliefs (notably those of Ghiberti) and, as sculptured rocks, appeared in a number of Baroque fountains, entered into sculpture in a new way when Henry Moore combined the forms of caves, rocks, hills, and cliffs with the human form in a series of large reclining figures.

There is nothing in sculpture comparable with the tradition of still-life painting. When objects are represented, it is almost always as part of a figure composition. A few modern sculptors, however, notably Giacomo Manzù and Oldenburg, have used still-life subjects.

NONREPRESENTATIONAL SCULPTURE

Kinds of nonrepresentational sculpture

There are two main kinds of nonrepresentational sculpture. One kind uses nature not as subject matter to be represented but as a source of formal ideas. For sculptors who work in this way, the forms that are observed in nature serve as a starting point for a kind of creative play, the end products of which may bear little or no resemblance either to their original source or to any other natural object. Many works by Brancusi, Raymond Duchamp-Villon, Jacques Lipchitz, Henri Laurens, Umberto Boccioni, and other pioneer modern sculptors have this character. The transformation of natural forms to a point where they are no longer recognizable is also common in many styles of primitive and ornamental art.

The other main kind of nonrepresentational sculpture, often known as nonobjective sculpture, is a more completely nonrepresentational form that does not even have a starting point in nature. It arises from a constructive manipulation of the sculptor's generalized, abstract ideas of spatial relations, volume, line, colour, texture, and so on. The approach of the nonobjective sculptor has been likened to that of the composer of music, who manipulates the elements of his art in a similar manner. The inclusion of purely invented, three-dimensional artifacts under the heading of sculpture is a 20th-century innovation.

Some nonobjective sculptors prefer forms that have the complex curvilinearity of surface typical of living organisms; others prefer more regular, simple geometric forms. The whole realm of three-dimensional form is open to nonobjective sculptors, but these sculptors often restrict themselves to a narrow range of preferred types of form. A kind of nonobjective sculpture prominent in the 1950s and '60s, for example, consisted of extremely stark, so-called primary forms. These were highly finished, usually coloured constructions that were often large in scale and made up entirely of plane or single-curved surfaces. Prominent among the first generation of nonobjective sculptors were Jean Arp, Antoine Pevsner, Naum Gabo, Barbara Hepworth, Max Bill, and David Smith. Subsequent artists who worked in this manner include Robert Morris, Donald Judd, and Phillip King.

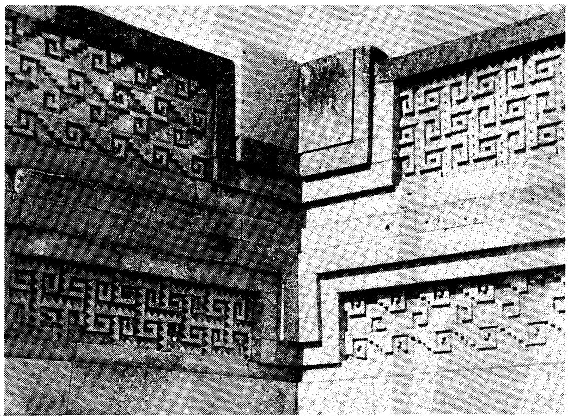
DECORATIVE SCULPTURE

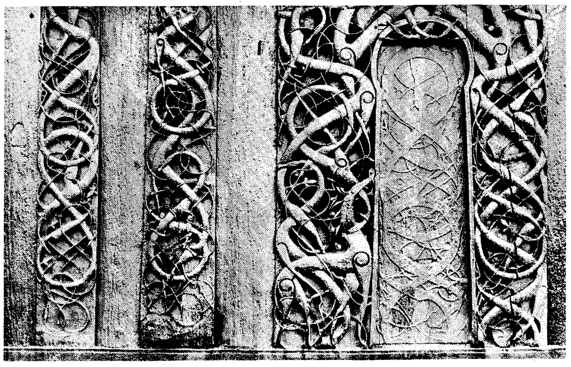
The devices and motifs of ornamental sculpture fall into three main categories: abstract, zoomorphic, and botanical. Abstract shapes, which can easily be made to fit into any framework, are a widespread form of decoration. Outstanding examples of abstract relief ornament are found on Islāmic, Mexican, and Mayan buildings and on small Celtic metal artifacts. The character of the work varies from the large-scale rectilinear two-plane reliefs of the buildings of Mitla in Mexico, to the small-scale curvilinear plastic decoration of a Celtic shield or body ornament.

Zoomorphic relief decoration, derived from a vast range of animal forms, is common on primitive artifacts and on Romanesque churches, especially the wooden stave churches of Scandinavia.

Botanical forms lend themselves readily to decorative purposes because their growth patterns are variable and

(Top) Henri Lehmann





Decorative sculpture.
(Top) Detail of the patio of the Palace of the Columns, Mitla, Oaxaca state, Mexico, Mixtec culture, 9th–16th century. (Bottom) Norse wood carving, late 11th century, detail of doorway of Urnes stave church, Norway. Doorway 3.90 × 2.80 m.

their components—leaf, tendril, bud, flower, and fruit—are infinitely repeatable. The acanthus and anthemion motifs of Classical relief and the lotuses of Indian relief are splendid examples of stylized plant ornament. The naturalistic leaf ornament of Southwell Minster, Reims Cathedral, and other Gothic churches transcends the merely decorative and becomes superbly plastic sculpture in its own right.

SYMBOLISM

Sculptural images may be symbolic on a number of levels. Apart from conventional symbols, such as those of heraldry and other insignia, the simplest and most straightforward kind of sculptural symbol is that in which an abstract idea is represented by means of allegory and personification. A few common examples are figures that personify the cardinal virtues (prudence, justice, temperance, fortitude), the theological virtues (faith, hope, and charity), the arts, the church, victory, the seasons of the year, industry, and agriculture. These figures are often provided with symbolic objects that serve to identify them; for example, the hammer of industry, the sickle of agriculture, the hourglass of time, the scales of justice. Such personifications abound in medieval and Renaissance sculpture and were until recently the stock in trade of public sculpture the world over. Animals are also frequently used in the same way; for example, the owl (as the emblem of Athens and the symbol of wisdom), the British lion, and the American eagle.

Allegory and personification

Beyond this straightforward level of symbolism, the images of sculpture may serve as broader, more abstruse religious, mythical, and civic symbols expressing some of mankind's deepest spiritual insights, beliefs, and feelings. The great tympanums (the space above the lintel of a door that is enclosed by the doorway arch) of Autun, Moissac, and other medieval churches symbolize some of the most profound Christian doctrines concerning the ends of human life and man's relations with the divine. The Hindu image of the dance of Śiva is symbolic in every detail, and the whole image expresses in one concentrated symbol some of the complex cosmological ideas of the Hindu religion. The Buddhist temple of Borobuḍur, in Java, is one of the most complex and integrated of all religious symbols. It is designed as a holy mountain whose structure symbolizes the structure of the spiritual universe. Each of the nine levels of the temple has a different kind of sculptural symbolism, progressing from symbols of hell and the world of desire at the lowest level to austere symbols of the higher spiritual mysteries at its uppermost levels.

In more individualistic societies, works of sculpture may be symbolic on a personal, private level. Michelangelo's "Slaves" have been interpreted as allegories of the human soul struggling to free itself from the bondage of the body, its "earthly prison," or, more directly, as symbols of the struggle of intelligible form against mere matter. But there is no doubt that, in ways difficult to formulate precisely, they are also disturbing symbols of Michelangelo's personal attitudes, emotions, and psychological conflicts. If it is an expression of his unconscious mind, the sculptor himself may be unaware of this aspect of the design of his work.

Many modern sculptors disclaim any attempt at symbolism in their work. When symbolic images do play a part in modern sculpture, they are either derived from obsolete classical, medieval, and other historical sources or they are private. Because there has been little socially recognized symbolism for the modern sculptor to use in his work, symbols consciously invented by individual artists or deriving from the image-producing function of the individual unconscious mind have been paramount. Many of these are entirely personal symbols expressing the artist's private attitudes, beliefs, obsessions, and emotions. They are often more symptomatic than symbolic. Henry Moore is outstanding among modern sculptors for having created a world of personal symbols that also have a universal quality; and Naum Gabo has sought images that would symbolize in a general way modern man's attitudes to the world picture provided by science and technology.

Symbolism of sculptural sites

Examples of sculpture of which the positioning, or siting, as well as the imagery is symbolic are the carved boundary stones of the ancient world; memorials sited on battlegrounds or at places where religious and political martyrs have been killed; the Statue of Liberty and similar civic symbols situated at harbours, town gates, bridges, and so on; and the scenes of the Last Judgment placed over the entrances to cathedrals, where they could serve as an admonition to the congregation.

The choice of symbolism suitable to the function of a sculpture is an important aspect of design. Fonts, pulpits, lecterns, triumphal arches, war memorials, tombstones, and the like all require a symbolism appropriate to their function. In a somewhat different way, the tomb sculptures of Egypt, intended to serve a magical function in the afterlife of the tomb's inhabitants, had to be images suitable for their purpose. These, however, are more in the nature of magical substitutes than symbols.

Uses of sculpture

The vast majority of sculptures are not entirely autonomous but are integrated or linked in some way with other works of art in other mediums. Relief, in particular, has served as a form of decoration for an immense range of domestic, personal, civic, and sacred artifacts, from the spear-throwers of Paleolithic man and the cosmetic palettes of earliest Egyptian civilization to the latest mass-produced plastic reproduction of a Jacobean linenfold panel (a carved or molded panel representing a fold, or scroll, of linen).

The main use of large-scale sculpture has been in conjunction with architecture. It has either formed part of the interior or exterior fabric of the building itself or has been placed against or near the building as an adjunct to it. The role of sculpture in relation to buildings as part of a townscape is also of considerable importance. Traditionally, it has been used to provide a focal point at the meeting of streets, and in marketplaces, town squares, and other open places—a tradition that many town planners today are continuing.

Sculpture has been widely used as part of the total decorative scheme for a garden or park. Garden sculpture is usually intended primarily for enjoyment, helping to create the right kind of environment for meditation, relaxation, and delight. Because the aim is to create a lighthearted arcadian or ideal paradisal atmosphere, disturbing or serious subjects are usually avoided. The sculpture may be set among trees and foliage where it can surprise and delight the viewer or sited in the open to provide a focal point for a vista.

Fountains, too, are intended primarily to give enjoyment to the senses. There is nothing to compare with the interplay of light, movement, sound, and sculptural imagery in great fountains, which combine the movement and sound of sheets, jets, and cataracts of water with richly imaginative sculpture, water plants and foliage, darting fish, reflections, and changing lights. They are the prototypes of all 20th-century "mixed-media" kinetic sculptures.

The durability of sculpture makes it an ideal medium for commemorative purposes, and much of the world's greatest sculpture has been created to perpetuate the memory of persons and events. Commemorative sculpture includes tombs, tombstones, statues, plaques, sarcophagi, memorial columns, and triumphal arches. Portraiture, too, often serves a memorial function.

Sculptural coins and medals

One of the most familiar and widespread uses of sculpture is for coins. Produced for more than 2,500 years, these miniature works of art contain a historically invaluable and often artistically excellent range of portrait heads and symbolic devices. Medals, too, in spite of their small scale, may be vehicles for plastic art of the highest quality. The 15th-century medals of the Italian artist Antonio Pisanello and the coins of ancient Greece are generally considered the supreme achievements in these miniature fields of sculpture.

Also on a small scale are the sculptural products of the glyptic arts—that is, the arts of carving gems and hard stones. Superb and varied work, often done in conjunction with precious metalwork, has been produced in many countries.

Finally, sculpture has been widely used for ceremonial and ritualistic objects such as bishop's croziers, censers, reliquaries, chalices, tabernacles, sacred book covers, ancient Chinese bronzes, burial accessories, the paraphernalia of tribal rituals, the special equipment worn by participants in the sacred ball game of ancient Mexico, processional images, masks and headdresses, and modern trophies and awards.

BIBLIOGRAPHY. EDWARD LANTERI, *Modelling and Sculpture: A Guide for Artists and Students,* 3 vol. (1965; previously pub. under the title *Modelling,* 3 vol., 1902–11), still an outstanding work on traditional methods; JACK C. RICH, *The Materials and Methods of Sculpture* (1947), comprehensive coverage of all except the most recent methods and materials; WILBERT VERHELST, *Sculpture: Tools, Materials, and Techniques* (1973), a wide-ranging survey with good coverage of modern materials; JOHN W. MILLS, *The Technique of Casting for Sculpture* (1967), and *Sculpture in Concrete* (1968), two useful technical handbooks; TREVOR FAULKNER, *The Thames and Hudson Manual of Direct Metal Sculpture* (1978), an informative work on a variety of historical and modern methods; UDO KULTERMANN, *The New Sculpture: Environments and Assemblages* (1968; originally published in German, 1967), a comprehensive account of these two recently developed forms of sculpture; RUDOLF WITTKOWER, *Sculpture: Processes and Principles* (1977), an authoritative account of the interaction of techniques and aesthetics in the history of sculpture; L.R. ROGERS, *Sculpture* (1969), and *Relief Sculpture* (1974), two books dealing with the principles and techniques of sculpture and their bearing on its appreciation as an art form.

(L.R.R.)

The History of Western Sculpture

Sculpture may be broadly defined as the art of representing observed or imagined objects in solid materials and in three dimensions. Like Western painting, Western sculpture has tended to be humanistic and naturalistic, concentrating upon the human figure and human action studied from nature. Early in the history of the art there developed two general types: statuary, in which figures are shown in the round, and relief, in which figures project from a ground.

Western sculpture in the ancient world of Greece and Rome and from the late Middle Ages to the end of the 19th century twice underwent a progressive development, from archaic stylization to realism; the term progressive here means that the stylistic sequence was determined by what was previously known about the representation of the human figure, each step depending upon a prior one, and not that there was an aesthetic progression or improvement. Modern criticism has sometimes claimed that much was lost in the change. In any event, the sculptors of the West closely observed the human body in action, at first attempting to find its ideal aspect and proportions and later aiming for dramatic effects, the heroic and the tragic; still later they favoured less significant sentiments, or at least more familiar and mundane subjects.

The pre-Hellenic, early Christian, Byzantine, and early medieval periods contradicted the humanist-naturalist bias of Greece and Rome and the Renaissance; and in the 20th century that contradiction has been even more emphatic. The 20th century has seen the move away from humanistic naturalism to experimentation with new materials and techniques and new and complex imagery. With the advent of abstract art, the concept of the figure has been broadened to encompass a wide range of nonliteral representation; the notion of statuary has been superseded by the more inclusive category of freestanding sculpture; and, further, two new types have risen to prominence: kinetic sculpture, in which actual movement of parts or of the whole sculpture is considered an element of design; and environmental sculpture, in which the artist either alters a given environment as if it were a kind of medium or provides in the sculpture itself an environment for the viewer to enter.

This article is divided into the following sections:

European Metal Age cultures

AEGEAN AND EASTERN MEDITERRANEAN

Aegean civilization is a general term for the prehistoric Bronze Age cultures of the area around the Aegean Sea covering the period from *c.* 3000 BC to *c.* 1100 BC, when iron began to come into general use throughout the area. From the earliest times these cultures fall into three main groups: (1) the Minoan culture (after the legendary king Minos) of Crete, (2) the Cycladic culture of the Cyclades islands, and (3) the Helladic culture of mainland Greece (Hellas). For convenience, the three cultures are each divided into three phases, Early, Middle, and Late, in accordance with the phases of the Bronze Age. The culture of Cyprus in the eastern Mediterranean, although it commenced somewhat later than those of the Aegean, came to parallel them by the Middle Bronze Age. The Late Bronze Age phase of the mainland is usually called Mycenaean after Mycenae, the chief Late Bronze Age site in mainland Greece.

Cultures of the Aegean

The first centre of high civilization in the Aegean area, with great cities and palaces, a highly developed art, extended trade, writing, and use of seal stones, was Crete. Here from the end of the 3rd millennium BC onward a very distinctive civilization, owing much to the older civilizations of Egypt and the Middle East but original in its character, came into being.

The Cretan (Minoan) civilization had begun to spread by the end of the Early Bronze Age across the Aegean to the islands and to the mainland of Greece. During the Late Bronze Age, from the middle of the 16th century onward, a civilization more or less uniform superficially but showing local divergences is found throughout the Aegean area. Eventually people bearing this civilization spread colonies eastward to Cyprus and elsewhere on the southern and western coasts of Asia Minor as far as Syria; also westward to Tarentum in southern Italy and even perhaps to Sicily. In the latter part of this period, after about 1400 BC, the centre of political and economic power, if not of artistic achievement, appears to have shifted from Knossos in Crete to Mycenae on the Greek mainland. (Ed.)

The Early Bronze Age (3000–2000 BC). *Early Minoan.* The early Minoan period saw a thousand years of peaceful development, which eventually gave place to the full flowering of the Minoan spirit, the Middle Minoan period. Pottery was preeminent among the Early Minoan arts.

Early Cycladic. The Early Cycladic culture developed on parallel lines to the Early Minoan. Thanks to obsidian from Melos, marble from many islands, and local sources of gold, silver, and copper, the Cycladic islanders rapidly became prosperous. As in Crete, the Early Bronze Age merged without incident into the Middle Bronze Age.

The Early Cycladic period is celebrated principally for its statuettes and vases carved from the brilliant coarse-crystalled marble of these islands. The statuettes, mostly of goddesses, are among the finest products of the Greek Bronze Age. They owe their charm to the extreme simplification of bodily forms. The typical "Cycladic idol" is a naked female, lying with her head back, her arms crossed over her breasts. These figures vary in size from a few inches to more than six feet in length (Figure 1).

Early Helladic and Early Cypriot. Mainland Greece probably received its Bronze Age settlers from the Cyclades, but the two cultures soon diverged. A prosperous era arose about 2500 BC and lasted until about 2200. Sculpture was overshadowed by pottery, metalwork, and architecture among the Early Helladic arts. In the Early Cypriot, the only surviving sculptures are a series of steatite cruciform figures of a mother goddess (3000–2500 BC) stylized in much the same way as contemporary Cycladic idols, from which they may have been derived.

The Middle Bronze Age (2000–1600 BC). *Middle Minoan.* The Middle Minoan period differs principally from the Early Minoan in the creation of palaces and a palatial life and art. Large-scale sculpture seems not to have found much favour in Crete, although fragments of life-size figures from this period were discovered in the Cyclades in the late 20th century. Miniature sculpture of the highest quality, some of it of fired sand and clay, was produced from at least as early as 1700 BC. Good examples are two female figures (called "Snake Goddesses") from Knossos, dated about 1700 BC (Archaeological Museum, Iráklion, Crete). These women stand with their arms in front of them, holding sacred snakes; they wear a flounced skirt and tight belt, and their breasts are bare.

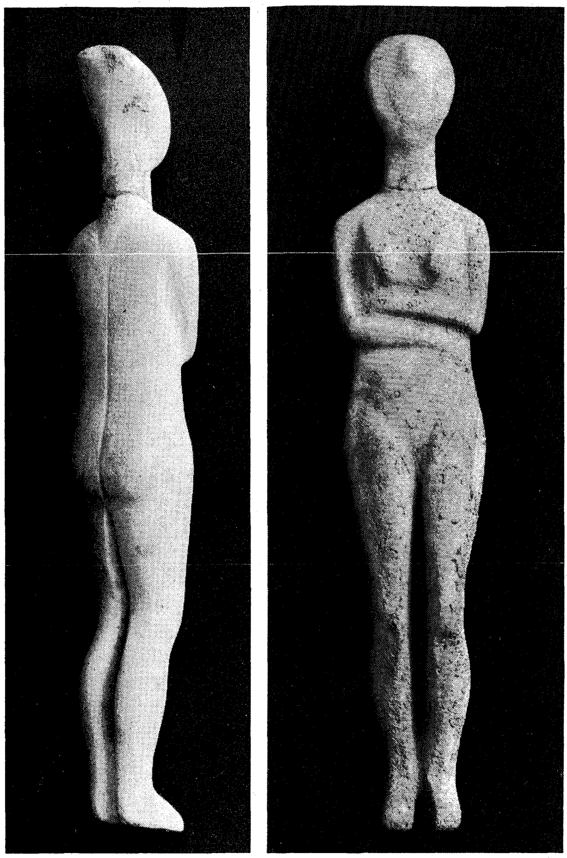


Figure 1: Marble Cycladic idol from Amorgos, Greece, *c.* 2500 BC. In the National Archaeological Museum, Athens.
Emile Serafis

Minoan "Snake Goddess" figures

Middle Cycladic, Middle Helladic, and Middle Cypriot. During the Middle Cycladic period, the Cyclades suffered a diminution in prosperity and seem to have become politically subordinate to Crete. Two waves of Indo-European peoples seem to have descended on the Greek mainland, one about 2200 BC and the other about 2000 BC. They destroyed much and for long contributed little to Greece's artistic heritage. The pottery of this period, however, is of high quality. The Middle Cypriot period was a development of the Early Cypriot. As on the mainland, no important art apart from pottery has survived.

The Late Bronze Age (1600–1100 BC). *Late Minoan.* Prosperity and artistic achievement remained at a high level until about 1450 BC, when all the great centres of Cretan culture were destroyed by earthquakes (probably connected with a cataclysmic eruption of the volcanic island of Thera). After these disasters, only the palace at Knossos was restored for occupation. About 1375 BC, however, the palace at Knossos was destroyed by fire. Thereafter Crete was a second-class power and became somewhat of a cultural backwater. Miniature sculpture was still popular. No longer in faience, figures were increasingly made of bronze, ivory, and terra-cotta. Some of the bronzes, cast solid by the "lost wax" process (using a wax model), are very fine, the earliest being the best. The subjects include male worshippers wearing boots, tight belt, and kilt; women (perhaps goddesses) dressed like the faience snake goddesses of the Middle Minoan period; and animals, especially bulls.

Carved-stone vases

Carved-stone vases were made between 1600 and 1450 BC. Elegant vessels were carved from such diverse materials as marble, obsidian, and steatite (Figure 2). Others, of soft stone, were made in the shape of bulls' heads, aston-

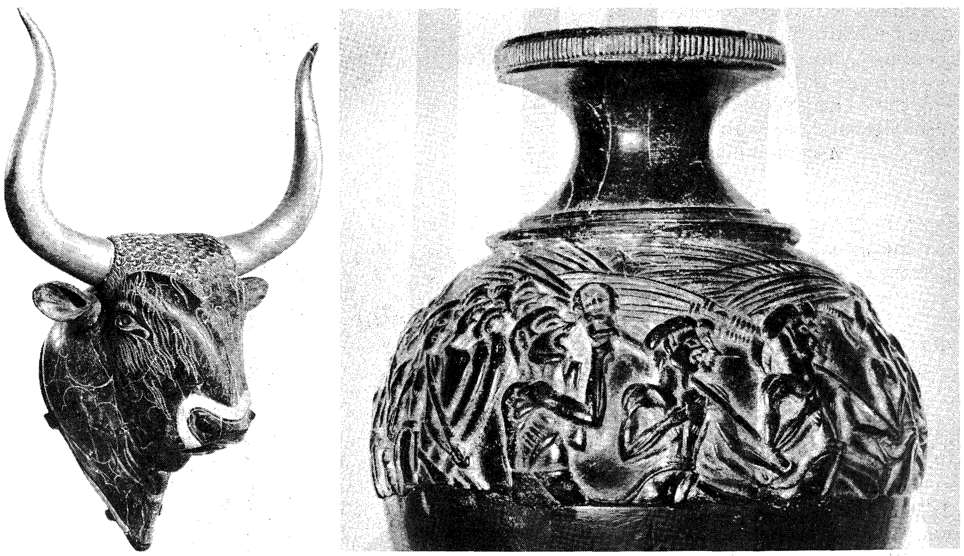


Figure 2: *Carved stone Minoan vessels of the Late Bronze Age (1600–1100 BC).* (Left) Serpentine rhyton (drinking vessel) in the form of a bull's head, steatite with gold-plated horns (now restored), from the Little Palace at Knossos, Crete, *c.* 1500 BC. In the Archaeological Museum, Iráklion, Crete. Height without horns 20.6 cm. (Right) Extant portion of the "Harvester Vase," from Ayía Triádha, steatite, *c.* 1600 BC. In the Archaeological Museum, Iráklion. Diameter 14 cm.

Alison Frantz

ishingly true to life, or were carved in relief, with religious or court ritual scenes, and covered with gold leaf.

The art of the seal engraver flourished until 1375 BC. Religious subjects, scenes of the bullring, and depictions of animals in their natural setting were popular. Even the exaggerations of the style reflect careful observation of the movements of the animals and their idiosyncratic anatomy, but they also relate the forms depicted to the shape of the stone—the curve of a bull's back or horns to that of the edge, for instance (Figure 3).

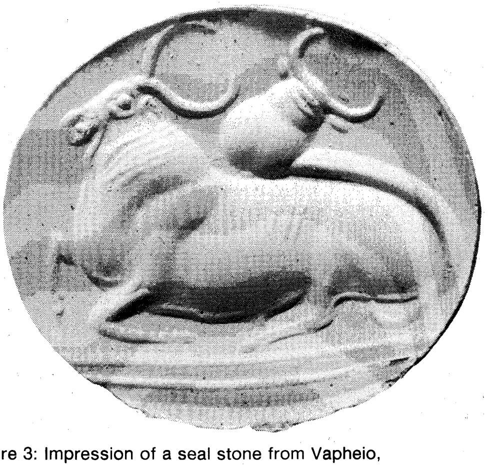


Figure 3: Impression of a seal stone from Vapheio, Greece, dating from *c.* 1500 BC.

From *Crete and Mycenae* published by Thames & Hudson, London, and Harry N. Abrams, New York; photograph, Hirmer Fotoarchiv, Munchen

Mycenaean. Mainland Greece enjoyed renewed contacts with Crete *c.* 1600 BC, and a rich culture, based on the Late Minoan, rapidly came into being. The Mycenaeans gained control of Crete *c.* 1450 BC, and between 1375 and 1200 BC they became masters of an empire that stretched from Sicily and southern Italy in the west to Asia Minor and the Levant coast in the east. About 1200 BC, however, many of the Mycenaean strongholds were destroyed by fire. There were signs of a renaissance, but the end of Mycenaean civilization came *c.* 1100 BC.

The Lion Gate at Mycenae

The Mycenaeans seem to have had more of a taste for monumental sculpture than had their Minoan mentors. Of the few surviving examples, the best known is a relief over the Lion Gate at Mycenae (*c.* 1250 BC), in which two lions confront each other across an architectural column (Figure 4). Probably heraldic in concept, this design is comparable with those on tiny seals and ivories of Cretan inspiration. Sculpture on a small scale, in ivory, bronze, and terra-cotta, generally Minoan in character, remained popular.

Late Cypriot. Cyprus reached its highest degree of prosperity in the Late Cypriot period, due to increased exploitation of its copper mines. There were close commercial relations not only with the Levant coast, as before, but also with Egypt, Crete, and Mycenaean Greece (the latter being close from 1400 BC). About 1200 BC Mycenaean Greeks, refugees from their homeland, settled in Cyprus. They introduced their skills and produced many luxury articles in a mixed Mycenaean-Cypriot style. Cyprus escaped the invasions that finally destroyed Mycenaean and Minoan culture, but its own culture did not last much longer. By 1050 BC, for reasons that are not clear, it, too, had ceased to exist.

From *Crete and Mycenae* published by Thames & Hudson, London, and Harry N. Abrams, New York; photograph, Hirmer Fotoarchiv, Munchen

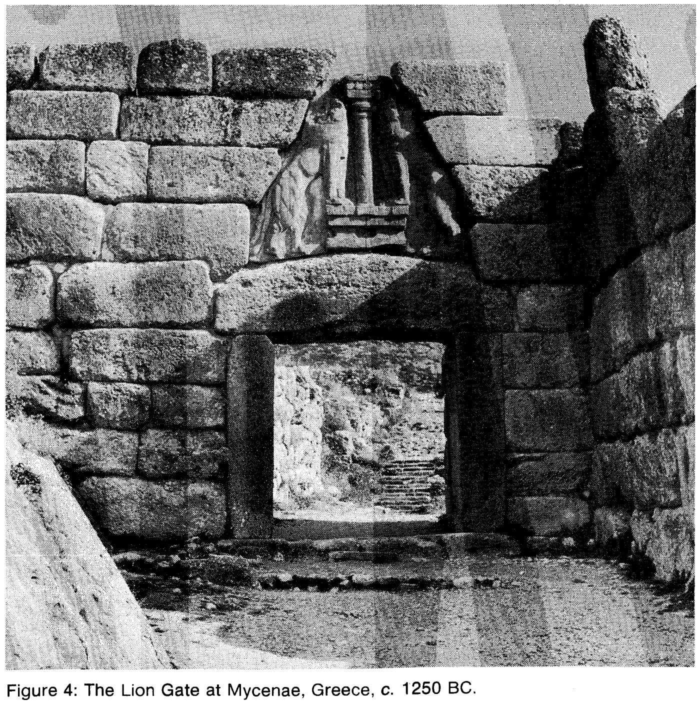


Figure 4: The Lion Gate at Mycenae, Greece, *c.* 1250 BC.

As in Crete, large-scale sculpture was rejected in favour of small-scale work. A bronze figure of a horned god (shortly after 1200 BC) from Enkomi (Cyprus Museum, Nicosia) shows a successful blend of Mycenaean and Cypriot elements. A good example of these characteristics is a carved ivory gaming box (British Museum), also from Enkomi, whose style shows a blend of Mycenaean and Middle Eastern motifs. (R.A.Hi./Ed.)

WESTERN MEDITERRANEAN

Like central and northern Europe, although to a lesser degree, the western Mediterranean was considerably behind the eastern Mediterranean, where civilization, the arts, and writing were born much earlier. The development of the metallurgical industry did not occur simultaneously in the various regions of the western Mediterranean, but it did bring important innovations in the mode of living and, of course, in the arts.

The Chalcolithic (Copper-Stone) era began in Spain at the end of the 3rd millennium BC at Los Millares, near Almería, and in Italy at the beginning of the 2nd millennium with the Remedello civilization. Bronze appeared not long afterward, around 1800 BC, in Italy and Sardinia. The Bronze Age in Italy gave way to the Iron Age at the beginning of the 1st millennium BC, but elsewhere, as in Sardinia or Spain, it lasted longer. The Iron Age flourished on the Illyrian coasts and in Italy from 900 to 800 BC; it also lasted varying lengths of time according to locale. After this, one may speak of the civilizations of Magna Graecia, of Rome, or of Etruria.

During the metal ages, popular migrations, commerce, and wars increased, which resulted in the rise of cities and of fortified works for their protection and defense, such as the talayots (round or quadrangular towers) of the Balearic Isles and the nuraghi (round towers) of Sardinia. With respect to the plastic arts, one particularly remarkable phenomenon was the birth and multiplication of megalithic human representations, which gained in number and importance from the 3rd to the 1st millennium BC. The Neolithic monuments, menhirs (single, vertical megaliths) and dolmens (structures of two vertical stones capped by a horizontal one), which had arisen in the megalithic era, continued to appear in the Copper and Bronze Ages, but then—here and there in Spain, Sardinia, Corsica, Liguria, and in the south of France—stelae-menhirs (carved or inscribed stone slabs used for commemorative purposes), like the stammerings of Western figure sculpture, imitated the human form. They maintained certain stylistic relations with rock engravings of mountainous regions, such as the Val Camonica.

Human representations

Bronze Age cultures. *Sardinia and Corsica.* The nuraghic civilization had an original sculpture expressed in a large production of bronze statuettes, about 500 of which have been found in nuraghi, temples, houses, and tombs (Figure 5). These figurines represent all classes of the proto-Sardinian populations—military chiefs, soldiers, priests, and women, as well as heroes and gods—in what seems to the modern viewer to be an engagingly direct but also sophisticated geometric style. The greatest number of these bronzes are today in the Museo Archeologico Nazionale in Cagliari, Sardinia. Some have been discovered in Etruscan tombs of Vetulonia and Vulci and have been dated to the period extending from the 9th to the 6th century BC.

Corsican menhir, or stela, statuary

Corsican menhir, or stela, statuary constitutes a group of special interest. The stone is imbued with life by a sculptural art that involves roughing-in of the head, animation of the upper portion of the body, and placement of a few elements of ornamentation or weaponry (sculpted in relief or, more rarely, engraved) on the schematically anthropomorphic image. These primitive statues are masculine and, no doubt, represent family or tribal heads made heroic or divine. This megalithic stela statuary art appears not only on Corsica but also in various other countries and regions of the western Mediterranean, including Spain, Sardinia, Liguria, and, in southern France, Provence, Aveyron, Hérault, and Gard, though to a lesser degree. The advance of this type of megalithic sculptural art is difficult to follow, but it is clear that these different groups are related, with close affinities existing between the stelae-menhirs of Corsica and those of the Ligurian coast. Such art is everywhere the expression of a patriarchal society seeking to impose on men's vision, massively and not without grandeur, the image of the departed ancestors.

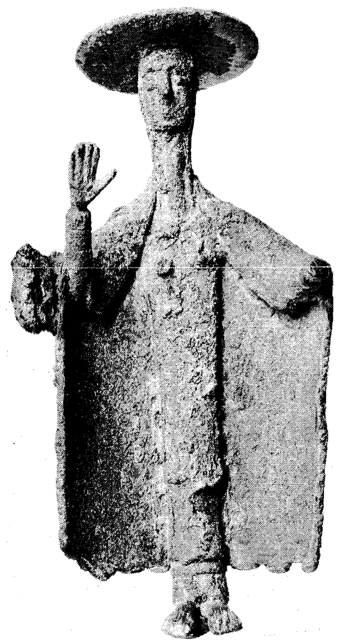


Figure 5: Bronze statuette, nuraghic civilization, from an unknown site in Sardinia, c. 7th century BC. In the Nationalmuseet, Copenhagen. Height 20.3 cm.

By courtesy of the Nationalmuseet, Copenhagen

Italy. From the Bronze Age of far northern Italy there survives an exceptional collection of rock engravings, a remarkable extension of an art that, in fact, had been represented in the prehistoric era and had not yet vanished completely. About 20,000 rock engravings have been found between altitudes of 5,000 and 5,600 feet (1,500 and 1,700 metres) in the Val Camonica, north of the town of Brescia. This art is found again further west, in the Maritime Alps of France on Monte Bego, between altitudes of 6,600 and 8,900 feet, and less remarkably elsewhere. What is exceptional about the carvings of the Val Camonica is that they represent a variety of subjects—rituals, battles, hunting, and daily labour—and that these were treated as compositions.

Rock engravings

Although engraving played a minor role in the case of the menhir statuary mentioned earlier, relations do exist between the sculpted works and the Camunian images of Monte Bego. The same representations of collar torques appear on the menhir statuary of Gard, Aveyron, and Tarn, on the one hand, and on certain monumental engravings of the Val Camonica, on the other. Some kind of relationship thus unites the arts of rock engraving and stela statuary in the Bronze Age.

Iron Age cultures. *Italy.* The Italian peninsula, which in the Bronze Age had been only one among many centres of civilization, took on a special importance in the Iron Age. Widespread and powerful cultural and artistic centres grew up there, first in the Villanovan civilization and later in the Etruscan; their influence was disseminated into the surrounding areas.

At the beginning of the 1st millennium BC there began to develop in the Po plain, in Tuscany, Latium, and some areas of Lucania, a new cremating civilization, which draws its name from that of the Villanova necropolis, discovered near Bologna. It is obviously related to the so-called Urnfield civilization that, at the end of the Bronze Age and beginning of the Iron Age, extended over central and eastern Europe and had developed a metal art with geometric and abstract ornamentation. The ashes of the dead were placed in urns thrust in level with the soil. From the Urnfield civilization arose two others: the Hallstatt civilization, which spread into the Balkans, northern and

Villanovan civilization

central Europe, and France, beyond the Pyrenees; and in Italy the Villanovan civilization and the civilizations that, to the east and west of the Po plain, were related to it, the so-called Golasecca civilization in the great lakes region and the Este civilization in the Venice area.

These Italic civilizations of the Early Iron Age, which appeared at the beginning of the 1st millennium BC and lasted for varying lengths of time, multiplied the number of dwelling sites. Originating as outposts established on naturally strong positions, they began to resemble towns as population increased.

The cinerary urn, which was made first of terra-cotta and later of bronze, assumes, by its form, a symbolic value (Figure 6). Biconical in form and covered with an overturned cup, later with a helmet, it schematically represents the appearance of the human body. Sometimes, as in examples from Latium and Tuscany, the funerary vessel is in the form of a hut or cabin—the house of the dead person whose remains it holds. The ornamentation, painted or engraved on the vases and engraved or in relief on metal objects, is in a geometric, nonfigurative style. Human or animal forms appear only rarely—in the decoration of small utilitarian objects such as vase handles and horse bits. It is a severe art, therefore, which essentially limits itself to linear exercises. Even motifs such as the disk, the solar boat, and the birds that encircle them, inherited from a more distant past and possessing primitively religious value, take on a stylized air and become abstract figures.

By courtesy of the Archeological Museum, Florence

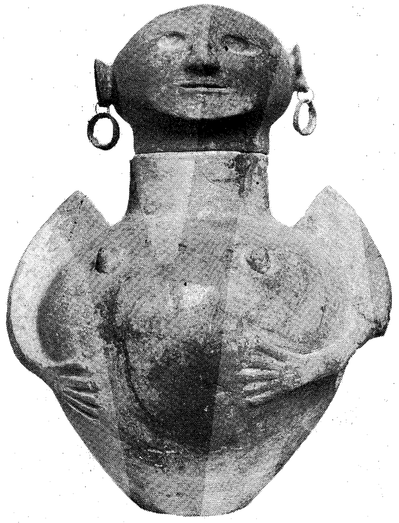


Figure 6: Villanovan canopic vase from Chiusi, Italy, terra-cotta, 6th century BC. In the Museo Archeologico, Florence. Height 44 cm.

A naturalistic note is provided by the imagery that decorates, in zones of superimposed relief, bronze vessels called situlae, a kind of pail found in Eastern countries and in the eastern Alps. These situlae were made in Venetian workshops in particular and were very popular in the neighbouring areas. They rapidly underwent an Etruscan influence, however, that tended to give prominence in the chased ornamentation to human figures at feasts, games, or funerals, as in the masterpiece known from the place of its discovery as the Certosa Situla (Museo Civico Archeologico, Bologna). (R.Bl./Ed.)

Etruria, Latium, and the Faliscan districts fall into three main areas of artistic production: northern, central, and southern, each centred upon cities with a distinctive artistic style. In the southern areas the chief centres were Caere and Veii, in which the Etruscan style most closely approached the Greek. In central Etruria, Vulci was evidently the leading art centre, although Tarquinia was unsurpassed in the beauty of its wall paintings. There were several potteries in Vulci, and the greater part of the central Etruscan bronzes, artistically the best, were produced there. The north was dominated by Clusium, although Perugia seems to have been important along with lesser centres at Volterra and Fiesole.

The very earliest examples of Etruscan statuary are flat, rectilinear figurines from Vetulonia and Capodimonte di Bolsena. These figures occur in later contexts in the Regolini-Galassi and Bernardini tombs, both of which contain pieces in a more advanced style that cannot have developed much later. These are statuettes of women with pigtails and long skirts depicted in a manner that suggests a north Syria influence, although this female type, frequently copied in ivory and amber, is certainly of local origin.

The earliest evidence of Greek influence is the presence of centaurs, perhaps transmitted on Corinthian vases. Their style in Etruria is Orientalizing, with a slim body and elongated legs, perhaps reflecting Cretan influence. These and other mythical creatures found great favour with the Vulci stonemasons. To archaic works of early Etruscan sculpture certain Greek parallels can be found in the late 7th and early 6th centuries, and in general characteristics the works still followed the Greek Archaic Daedalic tradition. The next change in style took place *c.* 550 BC, when art became distinctively Ionian. These new influences can be seen earliest in such pieces of bronze work as the Loeb Tripod from San Valentino near Perugia and the Monteleone chariot platings (in the Metropolitan Museum, New York City), but they soon become apparent also in the relief designs on *bucchero pesante* (heavily embossed black pottery) and in architectural reliefs like those from Tarquinia. By the end of the 6th century BC Veii possessed an excellent school of terra-cotta sculptures in Ionian styles. The statues of Apollo and of a votaress suckling a child are elaborately stylized in features, draperies, and muscles. Clay statuary, still retaining traces of former painting, was made in many Etruscan centres. Examples in the more mature classical style that began in the last quarter of the 5th century are the satyr-and-maiden groups from Satricum (modern Conca) in the Museo Nazionale di Villa Giulia, Rome, which contains a rich collection of architectural terra-cottas from Caere, Falerii, Veii, Satricum, and other sites.

Ionian influences

These pieces of statuary were designed to stand on temple roofs, and the socketed bases by which they were fixed have survived. Terra-cotta sculpture was also used for antefixes for these temples but above all for funerary sculpture. Sarcophagi with the sculptured figures of the husband and his wife reclining on the lids seem to have begun late in the 6th century, the date of the haunting sarcophagus from Caere (Villa Giulia, Rome). Bronze sculptures were also produced from the end of the 6th century, beginning with the famous She-Wolf, the symbol of modern Rome (Musei Capitolini, Rome), and the later Chimera from Arezzo (Museo Archeologico, Florence) or the so-called Mars of Todi (Vatican Museums) of the early 4th century BC.

In spite of great achievements in sculpture in the round, most of what has survived is in low relief, and a series of fine 6th–5th-century relief sarcophagi from Clusium, depicting dances, funeral games and banquets, or the journey of the dead to the underworld, are a major source of information on Etruscan everyday life. Superbly carved gravestones of the late mid-6th century are known from Clusium and Settimello, but the disk- and horseshoe-shaped gravestones of the Bologna, Fiesole, and Populonia graves have crude reliefs. (W.Cu./Ed.)

Sculpture developed but did not seek, as in Greece, to represent the idealized body of athletes and gods, attempting instead to represent the figure and features of the deceased (Figure 7). There was a continuing taste for real or fantastic animals such as lions, panthers, and sphinxes, and the Etruscan imagination seems to have been haunted by these beasts and demons, the vigilant guardians of the tombs.

Iberia. Whether in the form of great statuary or small votive images, Iberian figurative art was essentially religious and intended to represent sacred animals, deities, and their worshippers. Although much influenced by Greek and other sources, these works are vigorous and original, as may be seen from "La dama de Cerro de los Santos" in the Museo Arqueológico Nacional at Madrid and "La dama de Elche" in the Prado at Madrid (Figure 8). In the

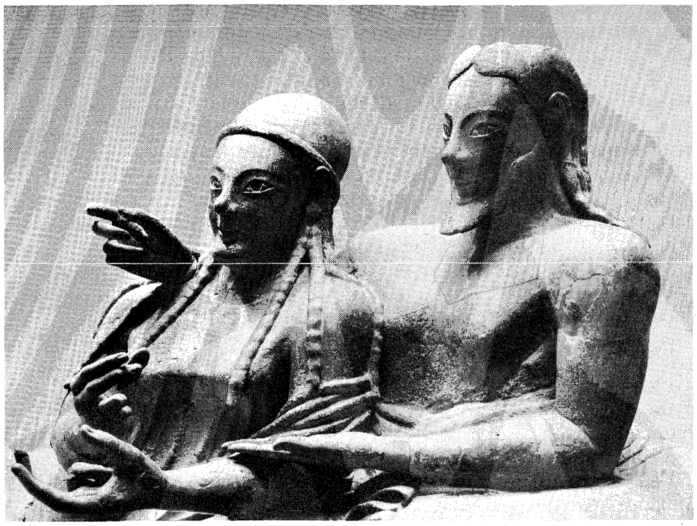


Figure 7: Detail of "Reclining Couple," clay sarcophagus from Caere, Italy, late 6th century BC. In the Museo Nazionale di Villa Giulia, Rome. Length of entire sarcophagus 2.01 m.

Alinari—Art Resource/EB Inc.

latter, a hieratic visage, with a severity not unlike some of the ideal heads of classical Greece, is adorned with a superabundance of heavy Iberian jewels. (R.Bl./Ed.)

Ancient Greek

Greek art no doubt owed much indirectly to the Minoan-Mycenaean civilization (now known in its later stages to have been Greek), which disintegrated at the end of the 2nd millennium BC, partly under the impact of a series of invasions from the Balkans. The period covered by this section, however, begins about 900 BC with the kaleidoscopic rearrangement of invaders and earlier inhabitants into a new pattern, which was followed by a steady artistic development—continuing without interruption down to the conquest of Greece by Rome in 146 BC. Even this diverted, rather than interrupted, the flow, and Greek artists continued to be predominant under the Roman Empire and beyond that into the Byzantine. But after Greece had become a Roman province, Greek art fell increasingly under the patronage of Romans and was devoted either to expressing Roman ideals or to reproducing older works of art. It is therefore reasonable to regard the later years of the 1st century BC, when the Roman Empire was forming, as the later limit of the period.

Five stages of Greek art

Within this period it is convenient to distinguish five stages of development. Their names are modern and arbitrary; the divisions between them are not equally sharp and do not apply equally to all parts of the Greek world, but they serve as a general guide to successive trends.

The first is the Geometric period (so-called from the rectilinear character of its art) from about 900 to about 800 BC, when Greece was self-contained and contact with the outside world was rare.

The second, the Orientalizing period, for about a century and a half from 800 BC, is one of contact with the East, a contact that had been broken by the upheavals at the end of the 2nd millennium.

The third period, the Archaic, from about 650 to about 480 BC, is characterized by the gradual absorption of Oriental elements and the rise and development of archaic Greek art.

The fourth period, from about 480 to about 330 BC, is known as the Classical; its beginning is marked by the rise of the sculptors Myron, Phidias, and Polyclitus and the painter Polygnotus, and its end, by the work of Scopas, Praxiteles, and Lysippus. (The word classical, which originally meant simply first-class, can also be used either in a narrower sense than this to denote only the Phidian age—*i.e.,* 50 years in the middle of the 5th century BC—or in a broader sense to cover the whole of post-Mycenaean Greek art from Geometric to late Roman.)

The fifth period is the Hellenistic, from about 330 BC, when the conquests of Alexander the Great opened new areas to the Greeks and the division of his kingdom among his Greek successors after his death in 323 diffused Greek art over the greater part of the known world, down to the late 1st century BC. Hellenistic symbolism and Hellenistic technical skill continued as living traditions under the Romans.

Statues were of limestone, marble, bronze, gold and ivory, terra-cotta, and wood. After the Archaic period the use of wood and of limestone seems to have been rare, as was the use of terra-cotta for statues of large size, although it should be noted that sculpture in the first and last of these materials tended to be ephemeral. The group of Orpheus and the two harpies that was restored at the J. Paul Getty Museum, Malibu, California, in the 1980s is astonishing not only for its quality but also for its size, and yet many other such figures may have been produced. Full-size statues of gold and ivory were rare at all times because of their cost; statues with gilded wooden bodies and marble extremities were sometimes made instead. For statuettes, ivory and amber, limestone, marble, wood, gold, silver, bronze, and terra-cotta were used; of these, terra-cotta was by far the most common, bronze and marble less so, and the rest rare. Extremely valuable because they can often be dated with accuracy are the types of sculpture used for the decoration of buildings: acroteria (*i.e.,* figures on the tops or ends of gables); figures in the low triangular field of the pediment under the gable (both of these are usually almost in the round); sculptured panels (metopes) of the Doric frieze, which are usually in high or very high relief; and the continuous Ionic frieze, which is usually in low relief.

Holle Bildarchiv, Baden-Baden, West Germany

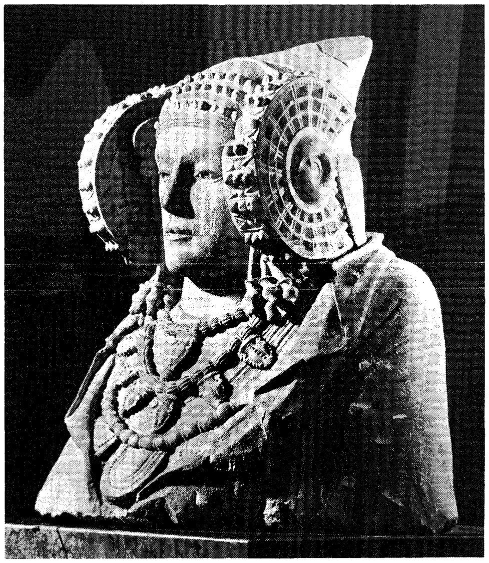


Figure 8: "La dama de Elche," painted limestone bust from Elche, Alicante, Spain, 5th century BC. In the Prado, Madrid. Height 26.0 cm.

Sources of modern knowledge of Greek sculptures

Of the many thousands of statues produced during the period in which Greek art flourished, not more than a few dozen survive, and those mostly mutilated. Knowledge of the history of Greek sculpture depends partly on these and partly on the architectural sculptures—both of high importance, since they are original. Much can also be learned about the general development of sculptural style from the small bronzes, often of very high quality, and from the terra-cottas. Of the small bronzes many, and of the terra-cottas very many, have survived, but they were made by independent artists and did not copy contemporary statues closely. The great bulk of evidence comes from copies made by Greeks, for Roman patrons, of originals now destroyed. Such evidence is invaluable but not entirely reliable. There is also literary evidence, but much of this is also second-hand or dates from long after the period in which the sculptures in question were made.

THE GEOMETRIC PERIOD

In the 9th century BC Greece was settling down again after upheavals and migrations both into and out of the mainland. It seems that invaders from the north brought with them the germs of an artistic style that developed into the Greek Geometric tradition.

In addition to the pottery, the Geometric period produced some terra-cottas and many small bronzes. The bronzes tended to be flat at first but became more solid and less angular as casting direct from wax models superseded cutting from bronze plates. Birds and other animals, especially horses, were popular and often admirably done; men, perhaps because their form commanded less imaginative interest, were not so successfully rendered; in the later stages of geometric art, groups of some complexity were attempted—a doe with her fawn, a man fighting (or greeting) a centaur, even a lion hunt complete with dogs. (B.As./Ed.)

THE ORIENTALIZING PERIOD

Sculpture of the Orientalizing period was profoundly affected by technical and stylistic influences from the East. In about 700 BC, the Greeks learned from their Eastern neighbours how to use molds to mass-produce clay relief plaques. Widely adopted, this technique helped to establish in Greece a stereotyped convention for figure representation, even in freestanding, unmolded sculptures; and a strong Eastern stylistic influence ensured that the convention was Oriental in flavour—in most cases a frontal pose with stiff patterned hair and drapery rendered in a strictly decorative manner. The adoption of this convention, which has come to be known as Daedalic style (after Daedalus, the legendary craftsman of Crete, where the style especially flourished), put an end to the development of naturalism and freedom in miniature sculptures that had shown promise in the Geometric period, and eventually became representative of even major Greek sculpture in the mid-7th century BC.

The Daedalic style

By courtesy of the Metropolitan Museum of Art, New York, Fletcher Fund, 1932

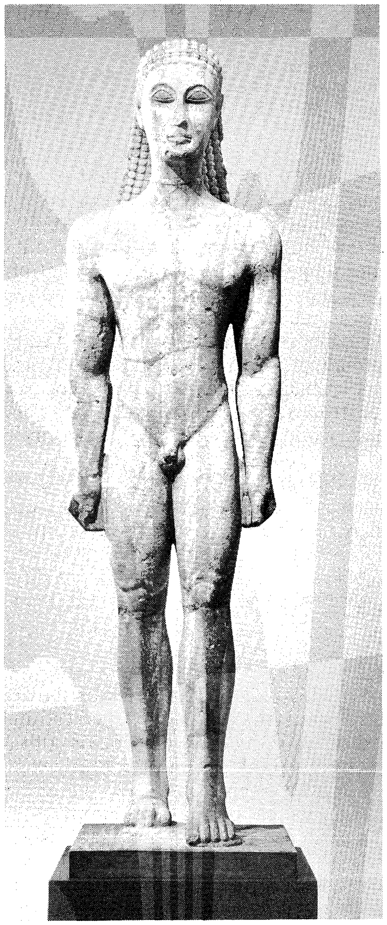


Figure 9: Marble kouros, *c.* 600 BC. In the Metropolitan Museum of Art, New York City. Height 1.86 m.

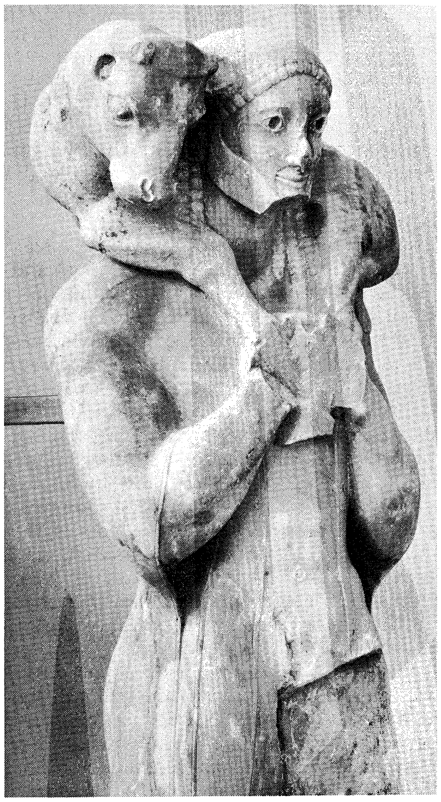


Figure 10: "The Calf Bearer," marble statue, *c.* 570 BC. In the Acropolis Museum, Athens. Height 1.65 m.

Hirmer Fotoarchiv, Munchen

In about 640 BC, however, a second Eastern influence began to be felt. As with the gigantic architecture of Egypt, the Greeks were impressed with the monumentality of Egyptian statuary, larger than life-size and executed in hard stone instead of the limestone, clay, or wood to which the Greeks had been accustomed. The Greeks learned the techniques of handling the harder stone in Egypt, and at home they turned to the fine white marble of the Cyclades islands (mainly Paros and Naxos) for their materials. It was at this time that the first truly monumental examples of Greek sculpture appeared. The idiom and proportions were at first still Daedalic. By about 630 BC, however, first in the islands and later in mainland Greece, they were carving freestanding figures of naked men that were copies of types formerly seen only in minor art and that owed something in proportion and details of pose to the common Egyptian standing figures. This new series of life-size or larger marble youths (kouroi) reveals rapid developments in technique and style, notably a transition from the Daedalic past to greater naturalism through the new monumental manner (Figure 9). The earliest of these figures were, as might be expected, dedications in sanctuaries, especially on the island of Delos, but some were grave markers, as on another island, Thera. At the same time, the older style was used for relief decoration of temples in Crete and Greece, particularly at Mycenae.

The kouros figure

THE ARCHAIC PERIOD

The kouroi, which had become standardized as freestanding statues of naked youths with hands to sides and one leg advanced, were the most representative examples of Archaic sculpture (Figure 9). At first their proportions were based on theory rather than observation; much the same was true of the anatomical details, which were treated as separate patterns applied to the figure with no proper understanding of their physiological relationships. Growing awareness of natural forms, although still without systematic study of the model, together with technical

The Archaic kouros and kore

mastery, led to a realism that is striking in comparison with the Daedalic pieces of the Orientalizing period (Figure 10). Still, the overriding considerations of proportion and pattern were never subordinated to nature. Only in the years just before the Persian invasion of 480 BC did some sculptors recognize the organic structure of the body and succeed in showing a truly relaxed pose, with the weight shifted onto one leg and the hips and torso consequently tilted to break the rigid symmetry of the characteristic kouros of the Archaic period (Figure 11).

In the female counterpart of the kouros, the kore, Archaic sculptors were again preoccupied with proportion and pattern—the pattern of drapery rather than of anatomy (Figure 12). Ionian (Chios, Samos) and island (Naxos) sculptors took the lead in developing decorative schemes for rendering the fall and splay of the folds of the loosely draped Ionic dress (chiton) and overmantle (himation). These patterns, like the anatomy of the kouroi, suggest nature rather than copy it; the strict logic of dressmaking is never observed by the sculptor, who uses the natural gesture of pulling a long skirt up and to one side first to produce a pleasing pattern of folds and only later to reveal the contours of the legs and body beneath. Most of the korai, like the kouroi, stood as dedications in sanctuaries, the richest series being from the Acropolis at Athens (these were overthrown by the Persians and then piously buried by the returning Athenians). Few of these statues were grave markers.

Archaic architectural sculpture

In the addition of sculpture to architecture, the determining factor was usually its position on the building. On a Doric temple, for instance, the metope frieze offered a series of rectangular plaques for reliefs that could accommodate two or three figures. There was a tendency in the Archaic period to let the action run on from one metope to the next, regardless of the intervening triglyph, a practice that was later abandoned. Above the frieze, the pediments formed by the gabled roof provided an awkward field—a long, low triangle. The sculptors of early temple pediments met the problem by depicting separate groups of different sizes (Figure 13), as at Corcyra (Archaeological Museum, Kérkira, Greece), or by devising monster bodies to fill the shallow corners, as in Athens (Acropolis Museum, Athens). Later, the advantages of using fighting groups with falling and fallen bodies were discovered; this type is represented at Athens and Aegina (Munich). The later Archaic pedimental figures were executed virtually in the round, standing against or just free from the background of the gable. Because these figures, unlike the kouroi and korai, were often in violent action, it may have been through meeting the problems of architectural sculpture that the artist arrived at a better understanding of the dynamics of the human body.

Hirmer Fotoarchiv, Munchen

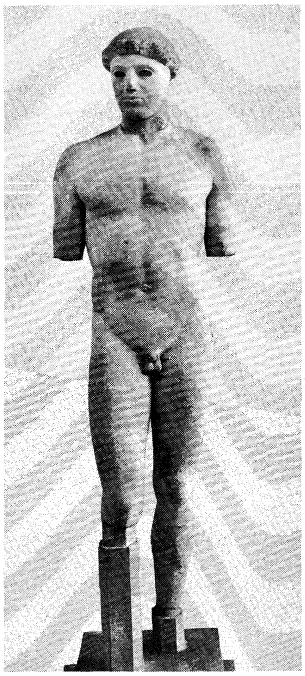


Figure 11: The "Kritios Boy," marble kouros, *c.* 490–480 BC. In the Acropolis Museum, Athens. Height 86 cm.

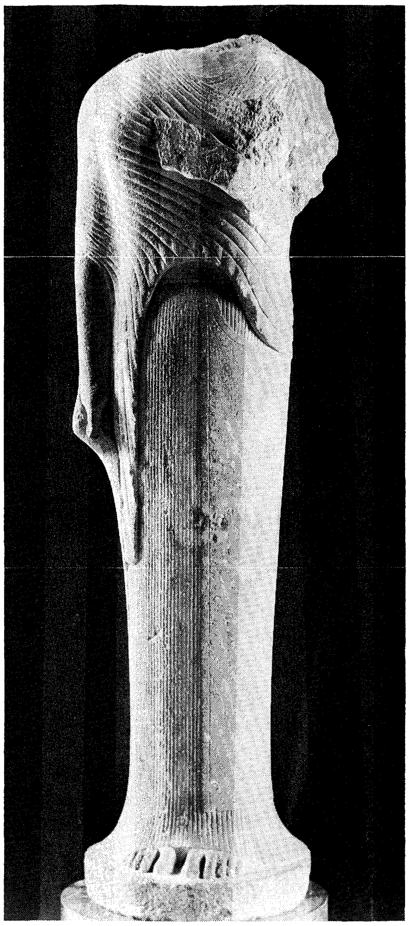


Figure 12: Marble statue of a woman dedicated by Cheramyes to Hera, found in the Heraeum on Samos, Greece, *c.* 560 BC. In the Louvre, Paris. Height 1.92 m.

Archives Photographiques

Work in relief also was used on gravestones, chiefly in Athens, for decorative bases of columns and for the frieze decoration on Ionic buildings, of which the best examples are from the Siphnian Treasury at Delphi (Archaeological Museum, Delphi), constructed shortly before 525 BC. The shallow relief on these works is little more than drawing rendered partly in the round; but the sculptor soon learned how, even in the shallowest relief, to indicate depth by overlapping figures and by bringing details up into the front plane. A dark-painted background helped the illusion; but the effect of the lavish use of colour on flesh, drapery, and backgrounds cannot now be readily appreciated since so little of it has survived in more than ghostly traces.

THE CLASSICAL PERIOD

The Severe style

Early Classical (c. 500–450 BC). This brief period is more than a mere transition from Archaic to Classical; in the figurative arts a distinctive style developed, in some respects representing as much of a contrast with what came afterward as with what went before. Its name—Severe style—is in part an indication that the "prettiness" of Archaic art, with its patterns of drapery and its decisive action, has been replaced by calm and balance. In vase painting and in sculpture, this new tone is evident in the composition of scenes and in details such as drapery, where the fussy pleats of the Archaic chiton give place to the heavy, straight fall of an outer robe called the peplos. The finest artists transformed the verve of the late Archaic

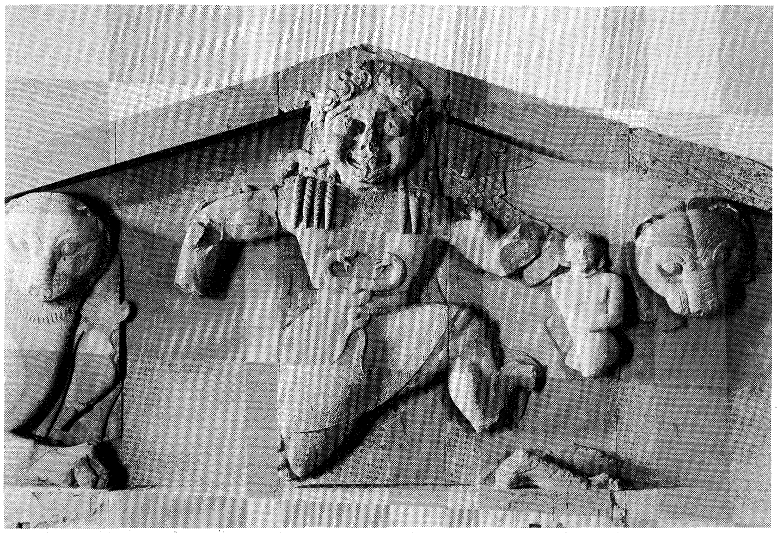


Figure 13: Gorgon from the west pediment of the Temple of Artemis, Corcyra (Corfu), Greece, limestone, *c.* 580 BC. In the Archaeological Museum, Kérkira, Greece.

By courtesy of the Deutsches Archaologisches Institut, Athens

style into more delicate expressions of emotion, and some were clearly checking their work more deliberately against the living model.

The early Classical period saw an impressive series of sculptural works that were excellent in their own right and significant in the continuing development of technical expressive skill and naturalism such as the relief carvings of the so-called Ludovisi Throne (Figure 14). Moreover, for the first time individual artists—and their contributions to technical and stylistic development—can in some cases be positively identified through Roman copies and written descriptions of their works.

Works of the Olympia Master

The finest examples of early Classical architectural sculpture are the works of the Olympia Master, an unidentified artist who decorated the pediments and frieze (Archaeological Museum, Olympia) of the Temple of Zeus at Olympia. In the east pediment, which shows men and women preparing for a chariot race, his figures display the sobriety and calm characteristic of the early Classical period. The men stand in the new, relaxed pose (the weight of the body being carried mainly by one leg) that was to be used by most sculptors throughout the period; and the women wear the peplos, its broad, heavy folds lending severity to the static composition. The west pediment, with a scene of struggling men and centaurs, has something of the rigid formality of the Archaic spirit, but here—and in the metopes that show the labours of Heracles—the artist has acutely observed differences of age in the human bodies and differences of expression—pain, fear, despair, disgust—in the faces (Figure 15). This was something new in Greek sculpture, and, in fact, cannot be readily matched in other works of this period.

The works of Myron

In freestanding sculpture—at this time, more commonly bronze than marble—the works of Myron (of Eleutherae, in Attica), identified through copies, were among the most celebrated of the period. Myron's most famous work is the "Discobolos" (discus thrower), of which a Roman copy (Museo Nazionale Romano) survives. Another of Myron's works surviving in copy is a sculpture of Athena with the satyr Marsyas (Athena in Städtische Galerie, Frankfurt am Main; Marsyas in Lateran Museums, Rome). The interplay of mood and action between the figures in this freestanding group is new, foreshadowed only by the now lost group of the Tyrantslayers erected in Athens at the end of the 6th century. Because bronze was often looted and corrodes easily, the majority of freestanding sculptures from this period have been lost. Some, however, have been rediscovered in the 20th century, the "Poseidon" (National Archaeological Museum, Athens) and the "Charioteer" from Delphi (Archaeological Museum, Delphi), for instance, although they have been eclipsed in fame by the still more remarkable pair of warriors dredged from the

Alinari—Art Resource/EB Inc.

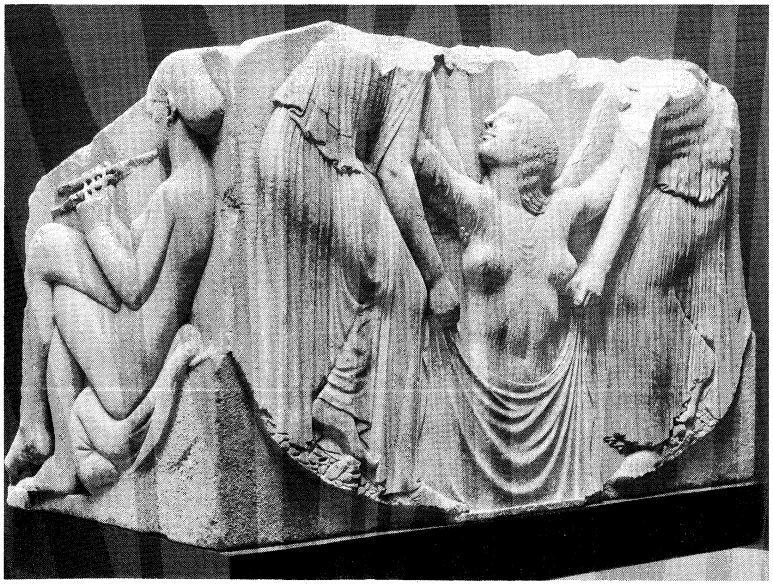


Figure 14: "Ludovisi Throne," *c.* 460 BC. In the Museo Nazionale Romano.

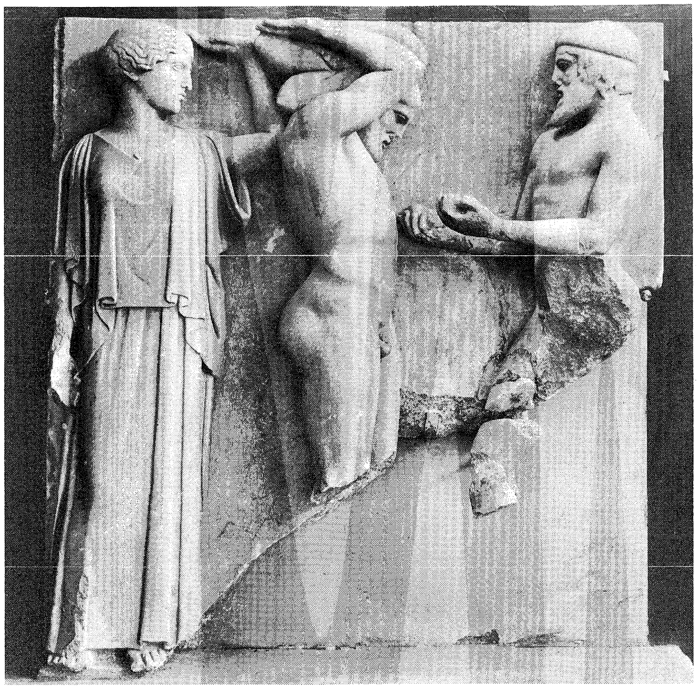


Figure 15: "Atlas Brings Heracles the Apples of the Hesperides in the Presence of Athena," marble metope from the east end of the Temple of Zeus at Olympia, *c.* 460 BC. In the Archaeological Museum, Olympia, Greece. 1.60 m × 1.42 m.

By courtesy of the Deutsches Archaologisches Institut—Athens

sea in 1972 and displayed in the Museo Nazionale, Reggio di Calabria. The finer of these latter bronzes (Figure 16), although it probably represents a mortal, has a supernatural glamour and a ferocity quite unlike the calm solemnity conventionally admired in Classical works. This derives partly from the glowing surface of the swelling musculature and the use of inlay for eyes, teeth, and lips.

The idealization of man

High Classical period (c. 450–400 BC). Since Roman times, Greek art of the second half of the 5th century BC has been generally regarded as the high point in the development of the Classical tradition. It was the most refined expression of the Greek view of their gods as men and of their men as partaking of the divine. The aesthetic result of this concept was that the bestial or supernatural was abjured in representations of the divine; thus, even a Greek monster, such as the centaur, seems plausible as an image combining humanity and divinity. To some degree, the idealization of human figures was facilitated by the Greeks' traditional concern with proportion and pattern. As a result of the value placed on the ideal image, the representation of extremes (of age or youth, for example, or of deep emotion) and of individuality was ignored or little practiced. Even figures engaged in violent or painful action have a calm, detached expression that modern observers may find chilly and unfeeling. Another reflection of the value placed on the ideal image is an increasing preoccupation with the "heroic nude." From an early phase of Greek art, the artist had shown his interest in man as man rather than as individual. In the Archaic period, the artist studied the visual pattern of the naked male body. When anatomical competence was complete, it was still the abstraction, the pattern, that dictated that his subjects be nude; for it is certain that the average Greek dressed for everyday life and for battle and that only in the exercise ground or the racetrack was the naked body freely revealed.

During the high Classical period, Athens resumed a position of importance as an artistic centre of the Greek world after years of inactivity. Once most of the Greek homelands were secure from the Persian threat, the funds that had been provided to Athens by the Greek states to lead their defense were turned by the statesman Pericles to the embellishment of Athens itself, and a program of rebuilding temples in the city and countryside was begun. This task attracted sculptors, masons, and other artists to Athens from all over the Greek world. It is largely the work of these artists, under the guidance of Athenian masters, that determined what is now recognized as the high Classical style.

The statues of Phidias and Polyclitus

Of the several types of sculpture that flourished during the high Classical period, major statuary is least represented in surviving examples. Phidias, the most influential sculptor of the period, made two huge cult images plated with gold and ivory, the statue of Athena for the Parthenon and a seated statue of Zeus for the temple at Olympia that was one of the seven wonders of the ancient world. These works amazed and overawed viewers through all antiquity, but no adequate copies survive. Another important sculptor of the period, whose work can be seen through copies, was Polyclitus, from Argos. Polyclitus embodied his views on proportion in his "Doryphoros" ("Spear Bearer"), called "The Canon" because of its "correct" proportions of one ideal male form. Unlike freestanding statues, architectural sculpture from the high Classical period has survived in abundance. The Parthenon sculptures must have been executed by many different hands, but, because the overall design was by Phidias, the composition and details undoubtedly reflect his style and instructions. The pedimental figures and frieze, especially, display the Classical qualities of idealization (Figure 17). These allow an approximate assessment of Phidias' style and the importance of his contribution to the establishment of the Classical idiom. About the time that full employment for sculptors in Athens on the Parthenon came to an end, there began a distinguished series of carved relief gravestones for Athenian cemeteries. The general type had been familiar in Archaic Athens, and the practice continued in other parts of Greece through the early Classical period, mainly in the islands and in Boeotia. The new Attic series, with calm and dignified groups of figures in generalized settings of domesticity or

SCALA/Art Resource, NY

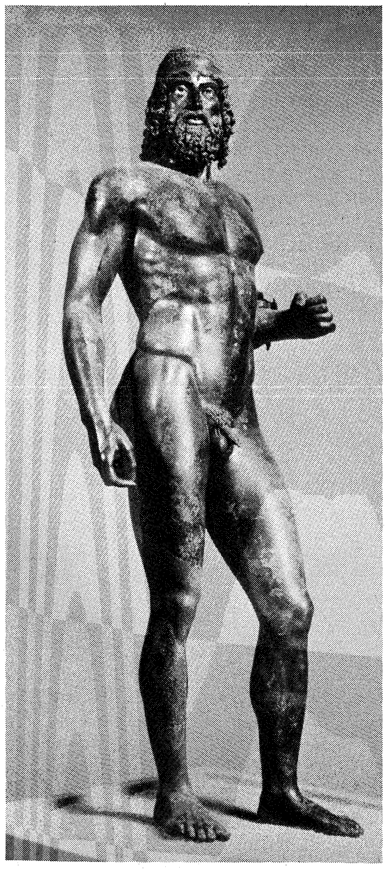


Figure 16: Greek warrior, 5th century BC, one of a pair of bronze statues found in the Mediterranean off Riace, Italy. In the Museo Nazionale, Reggio di Calabria, Italy. Height 2.00 m.

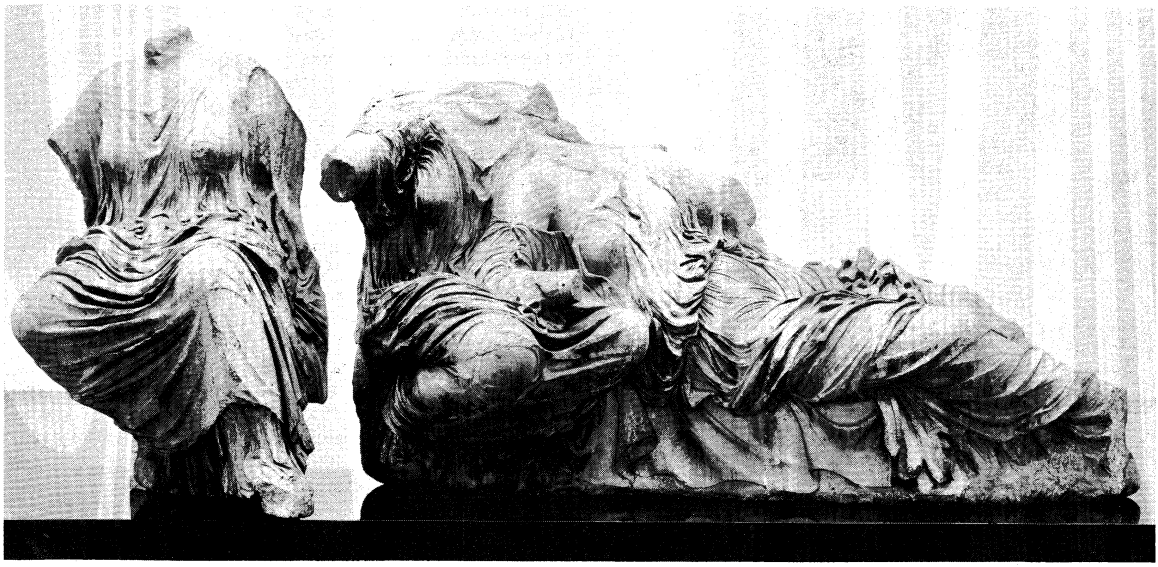


Figure 17: Probably "Leto, Dione, and Aphrodite," marble figures from the east pediment of the Parthenon on the Athenian Acropolis, *c.* 432 BC. In the British Museum, London. Over life-size.

By courtesy of the trustees of the British Museum

leave-taking, exploited effectively the rather impersonal calm in figure and features of the Classical conventions.

The other important class of sculpture, much of which has survived in the original, is the dedicatory—votive reliefs or major works like the "Nike" ("Victory") found at Olympia, made by Paeonius. This work, and others that belong to the last years of the century, such as the frieze from the balustrade of the temple of Athena Nike on the Acropolis at Athens, give a clear indication of progress and change in sculptural style. In the representation of the female body, never before a subject of particular interest to the sculptor (with the distinguished exception of the Olympia Master), true femininity was at last achieved through observation; in these works the figures are no longer like male bodies with the more obvious female characteristics added, which had generally been true of earlier works. Drapery, which had for its patterns been an important element of female figures in the Archaic period, has a heaviness, almost a life of its own in the Parthenon sculptures. By the end of the century, in the Nike balustrade, it is shown pressed tight against the body revealing the forms of the limbs and torso clearly beneath, with brittle, dramatic folds standing clear of the surface. This last style, together with the new approach to the rendering of women's bodies, led quickly to a deliberately sensual effect in statuary and hastened the decline of the unemotional Classical conventions.

Developments in rendering the female body

Late Classical period (c. 400–323 BC). The 4th century saw a dramatic increase of wealth in Greece but less in the hands of the warring states of the 5th century and more concentrated on the periphery of the Greek world—with the western colonies, the eastern Greeks, who continued in close touch with the friendlier Persian provinces, and the increasingly powerful Macedonian kingdom in the north. Macedonian power, culminating in Alexander the Great's annexation of the whole Persian Empire in the third quarter of the 4th century, was to transform Greek art as effectively as it did Greek life and politics. Even before Alexander's accession, however, the seeds of change were sown. The many new centres and patrons for artists may have made it easier for them to break with Classical conventions established in 5th-century Athens or by dominant 5th-century artists like Polyclitus. The trend was toward greater individuality of expression, of emotion, and of identity, leading eventually to true portraiture. The last was encouraged by the ambitions and pride of rulers such as the Macedonian kings or by the royal houses of Hellenized provinces in the western Persian Empire. To the same sources can be traced the new interest in monumental tomb construction. Men were aspiring more openly to divinity, and Greek art was no barrier to its explicit expression. It is clear, however, that artists were conscious of the values that were set in the 5th century, and by no means did they act as revolutionaries in styles or techniques. The development of Greek art was swift but smooth, and personalities lent impetus to the development rather than changing its flow dramatically.

The sculpture of Praxiteles

Three names dominate 4th-century sculpture, Praxiteles, Scopas, and Lysippus. Each can be appreciated only through ancient descriptions and copies, but each clearly contributed to the rapid transition in sculpture from Classical idealism to Hellenistic realism. Praxiteles, an Athenian, demonstrated a total command of technique and anatomy in a series of sinuously relaxed figures that, for the first time in Greek sculpture, fully exploited the sensual possibilities of carved marble. His Aphrodite (several copies are known), made for the east Greek town of Cnidus, was totally naked, a novelty in Greek art, and its erotic appeal was famous in antiquity. The "Hermes Carrying the Infant Dionysus" (Archaeological Museum, Olympia) at Olympia, which may be an original from his hand, gives an idea of how effectively a master could make flesh of marble (Figure 18). The reputation of Scopas, from the island of Paros, came from the intensity of expression with which he imbued his figures. Fragments of his work at Tegea (National Archaeological Museum, Athens) show his technique in the deep-sunk eye sockets that characterize his faces and that transform the hitherto passionless features of Classical sculpture into studies of intense emotion. Praxiteles and Scopas seem to typify the new spirit that can readily be discerned in surviving original sculptures. The "Demeter of Cnidus" (British Museum, London; perhaps by the Athenian sculptor Leochares) is Classical in mood, but the features are Praxitelean; and in the reliefs on the Mausoleum (British Museum, London) at Halicarnassus (on which both Scopas and Leochares are said to have worked), the vigour of the battle scenes is heightened by both the intensity of the features and a new, rather flamboyant use of drapery. On Athenian grave reliefs the Classical calm gave place to expressions of controlled but deep emotion. These are styles that can be recognized in places far from Greek soil, as in the relief sarcophagi fashioned by the Greeks for the kings of Sidon in Phoenicia.

The sculpture of Lysippus

Lysippus, from Sicyon in the northern Peloponnese, was Alexander's favourite sculptor. He was true to the Classical tradition in demonstrating his views on proportion by sculpturing athlete figures in different poses, although his types have heavier bodies and smaller heads than those of the Classical standard set down by Polyclitus. But he adds something to these single figure studies; for the first time they are composed in such a way that the viewer is

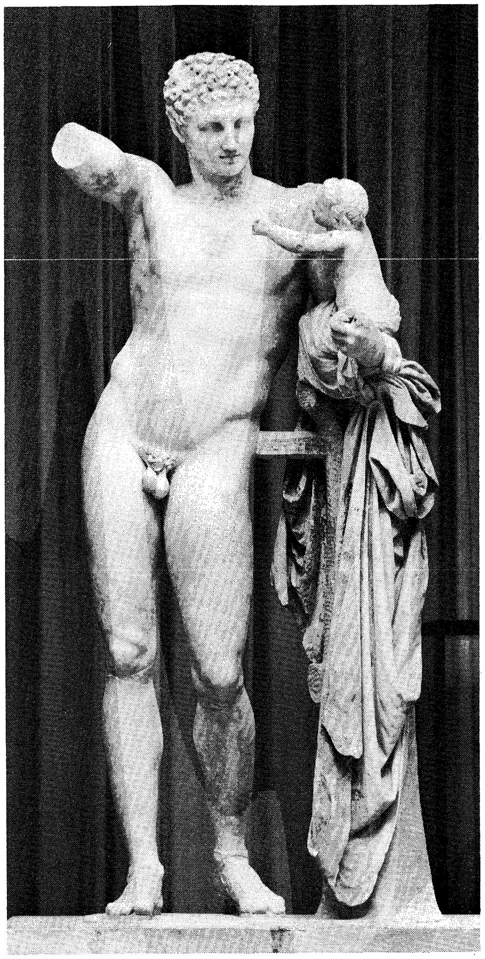


Figure 18: "Hermes Carrying the Infant Dionysus," marble statue by Praxiteles, *c.* 350–330 BC (or perhaps a fine Hellenistic copy of his original). In the Archaeological Museum, Olympia, Greece. Height 2.15 m.

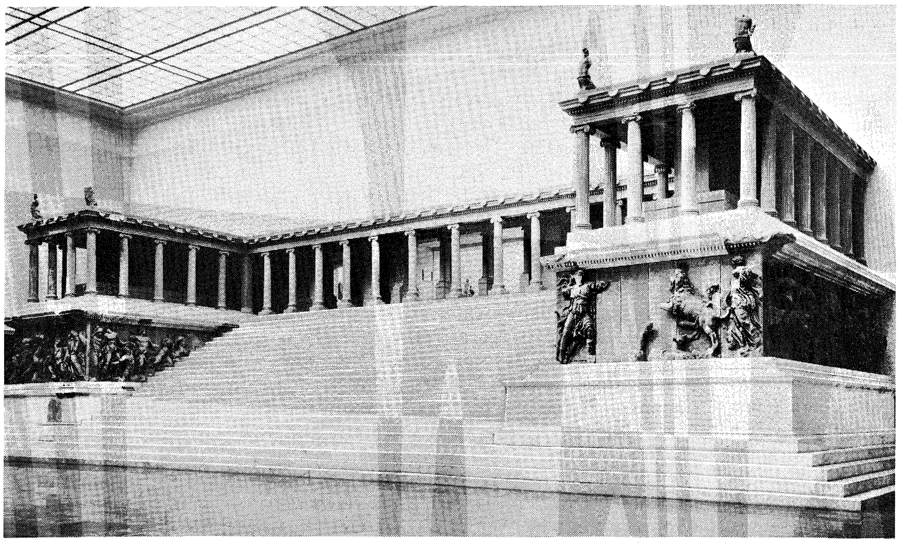
Hirmer Fotoarchiv, Munchen

invited to move around them, and they are not tied to a single optimum viewpoint, as even Praxiteles' figures had been. This was an important innovation in the history of sculpture.

Another innovation, in the development of which Lysippus must also have played a vital part, is portraiture; he carved likenesses of Alexander. Nevertheless, portraits of contemporaries were still exceptional, and many early portraits are semi-idealized studies of the great philosophers, statesmen, or poets of the Classical period. And yet, it is clear that by now the use of live models was commonplace, as can be judged from the works or copies that survive and from stories of Praxiteles' use of his mistress Phryne as a model or of Lysippus' brother taking casts from life. By the time of Alexander most of the important problems in the realistic or dramatic treatment of features, pose, and drapery had been solved, leaving to later generations an opportunity only to exaggerate anatomy or expression or to devise sculptural groups of yet greater complexity. Fourth-century sculptors, led by Praxiteles, Scopas, and Lysippus, gathered and expressed the best of what had been learned before of anatomy, pattern, and composition; by adding emotional appeal they can be said to have achieved the logical culmination of the Classical tradition, in which Phidian sculpture in the 5th century was but one brilliant and influential episode.

Hellenistic period. Styles of Hellenistic sculpture were determined by places and schools rather than by great names. Pergamene sculpture is exemplified by the great reliefs from the altar of Zeus (Figure 19), now in East Berlin, and copies of dedicatory statues showing defeated Gauls (Figure 20, bottom). These, like the well-known "Nike of Samothrace" (Figure 21), are masterful displays of vigorous action and emotion—triumph, fury, despair—and the effect is achieved by exaggeration of anatomical detail and features and by a shrewd use of the rendering of hair and drapery to heighten the mood. The "Laocoon" group (Vatican Museums), a famous sculpture of the Trojan priest and his two sons struggling with a huge serpent, probably made by Rhodian artists in the 1st century AD but derived from examples of suffering figures carved in the 1st century BC, is a good example of this applied to a freestanding group (Figure 20, left); and the "Belvedere Torso" (Vatican Museums), much admired in Renaissance Italy, of the effective emphasis of anatomy (Figure 20, right). In vivid contrast, a fully sensual treatment of the female nude was achieved by careful surface working of the marble, and the accentuation of femininity by the incorporation of sloping shoulders, tiny breasts, and high full hips. It is the Hellenistic Aphrodite, such as the "Venus de Milo" (Figure 22), who proliferates in Roman copies. The sculptural groups such as Laocoon were novel, demanding a palatial or sanctuary setting and far removed from earlier two-figure groups or the more nearly comparable but one-view pedimental compositions. The new realism extended to the portrayal of old age, decrepitude,

(Right) By courtesy of the Staatliche Museen zu Berlin; photograph, (left) EDI Studio, Barcelona



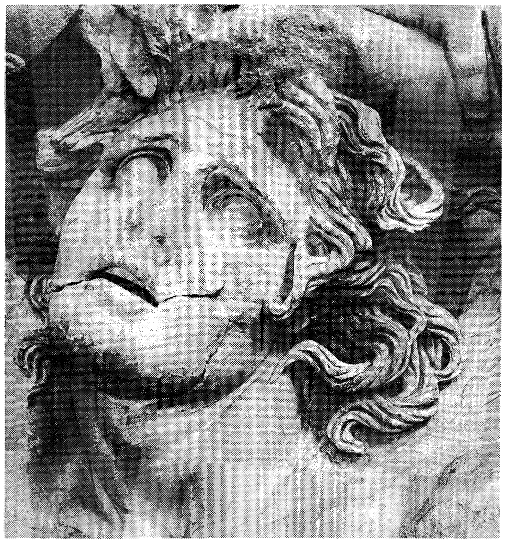


Figure 19: *Hellenistic relief sculpture.*
Great altar of Zeus at Pergamum. In the Staatliche Museen zu Berlin. (Left) Reconstruction of the west front. (Right) Frieze detail of Alcyoneus seized by the hair, from the marble relief on the east side of the great altar of Zeus, *c.* 180 BC.

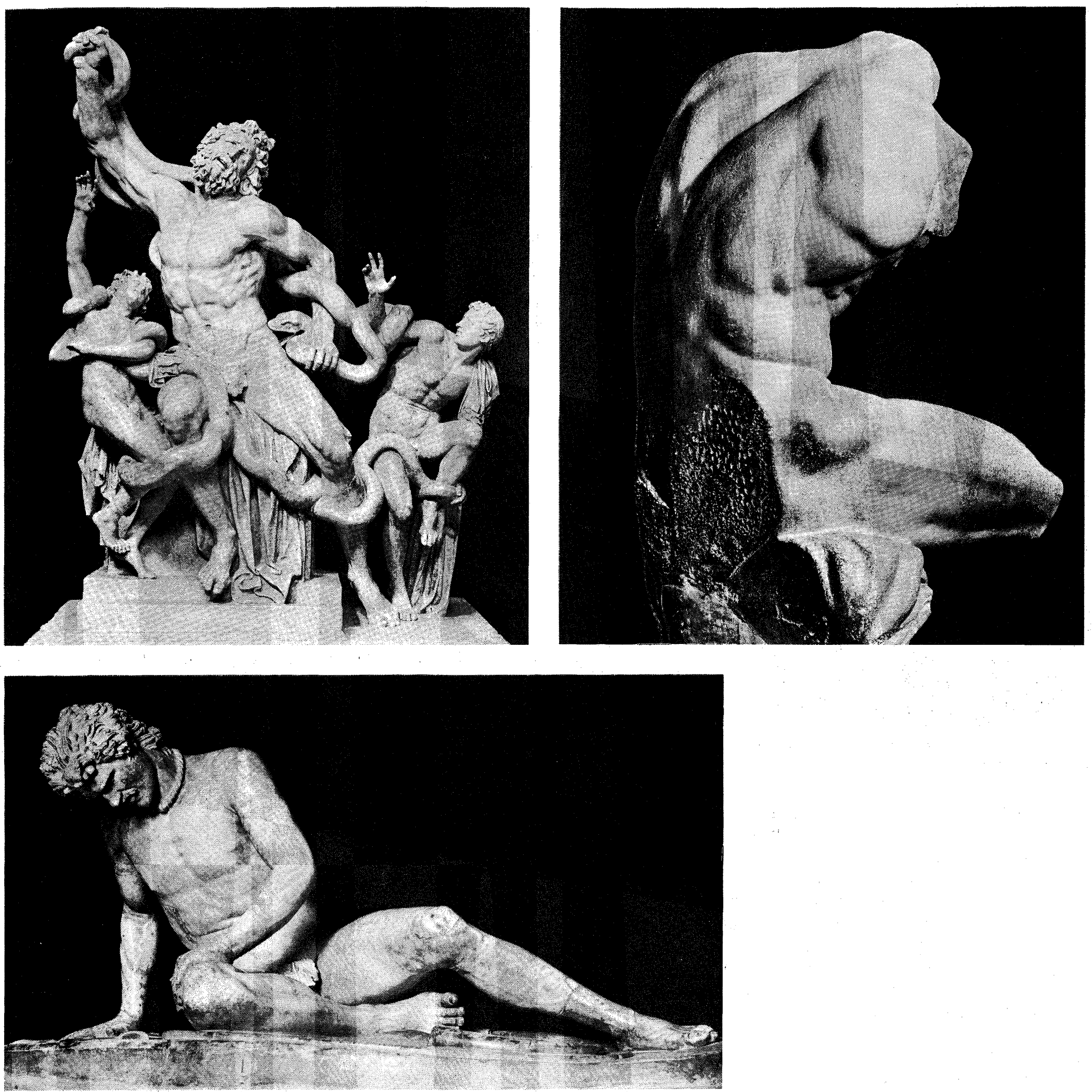


Figure 20: *Vigorous action and dramatic emotion in Hellenistic sculpture.*
(Top left) "Laocoon," marble sculpture by Agesander, Athenodorus, and Polydorus of Rhodes, 1st century AD. In the Vatican Museums. Height 2.41 m. (Top right) "Belvedere Torso," marble by Apollonius, 1st century BC. In the Vatican Museums. (Bottom) "Dying Gaul," marble Roman copy after a bronze original, from Pergamum, *c.* 230–220 BC. In the Museo Nazionale Romano.

(Top) Alinari—Art Resource/EB Inc., (bottom) Anderson—Alinari from Art Resource/EB Inc.

New realism in sculpture

disease, low life, and even the grotesque. Alexandria, in its major and minor (clay) works of sculpture, seems to have been one of the important schools in this genre. For the first time in Greek art, babies were rendered as other than reduced adults. In portraiture, the idealizing tendencies of the 4th century were still strong, and portraits of kings or poets were overlaid by conceptions of kingship or artistry. It was to take Roman patronage to enforce a more brutal realism in portraiture of contemporaries.

Copies and stylistic revivals

Two of the most significant developments in Hellenistic sculpture, however, had nothing to do with the evolution of new styles or types of compositions. The first was the production of accurate copies of earlier works, which began by about 100 BC, in part occasioned by the demand from the Roman West. This production stimulated interest in the styles of the great Classical sculptors and helped to determine the decidedly Classical atmosphere of early imperial art. The second, related development is the creation of original works deliberately in the style of the late Archaic, early Classical, or full Classical periods. This archaizing can be seen as both a reaction against the more exuberant Hellenistic sculptural styles and a response to the new interest in the Classical past.

Greece and Rome. It was Hellenistic art that the great Roman Republic and its early empire came to know and to covet. It was already to some degree familiar to them from the work of the western Greeks in Italy and Sicily, and the Romans formed a closer acquaintance with it in the court of Alexandria and from the profits of their diplomacy and warfare. The flow of works of art and artists to the west began, and the classical styles of early imperial Rome are exactly those of the late Hellenistic Greek world, in many instances executed by the same artists. Thus, in the early empire the majority of known artists' signatures are those of Greeks. The adoption of Greek art by the Roman Empire ensured its continuity in the Western tradition and its eventual transmission, through the Renaissance revival, to the modern world.

(Jo.Bo./Ed.)

Roman and Early Christian

There are many ways in which the term ancient Roman art can be defined, but here, as commonly elsewhere, it is used generally to describe what was produced throughout the part of the world ruled or dominated by Rome until

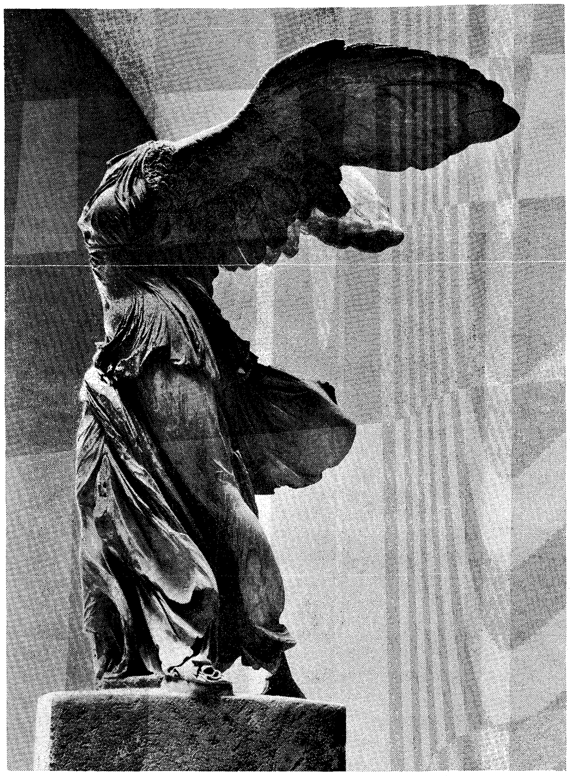


Figure 21: "Nike of Samothrace," marble statue, *c.* 200 BC. In the Louvre, Paris. Height 2.44 m.

J.E. Bulloz

around AD 500, including Jewish and Christian work that is similar in style to the pagan work of the same period.

Greek influences

The Romans were always conscious of the superiority of the artistic traditions of their neighbours. Such works of art as were made in or imported into Rome during the periods of the monarchy and the early republic were produced almost certainly by Greek and Hellenized Etruscan artists or by their imitators from the cities of central Latium; and throughout the later republican and the imperial epochs many of the leading artists, architects, and craftsmen had Greek names and were Greek, or at any rate Greek-speaking. References in ancient literature and signatures of artists preserved in inscriptions leave no doubt on this point. According to tradition, the earliest image of a god made in Rome dated from the 6th century BC period of Etruscan domination and was the work of Vulca of Veii. A magnificent terra-cotta statue of Apollo found at Veii (Figure 23) may give some notion of its character. In the 5th, 4th, and 3rd centuries BC, when Etruscan influence on Rome was declining and Rome's dominion was spreading through the Italian peninsula, contacts with Greek art were no longer chiefly mediated via Etruria but, instead, were made directly through Campania and Magna Graecia; paintings and "idealizing" statues of gods and worthies mentioned in literature as executed in the capital during this period were clearly the works of visiting or immigrant Greek artists. The plundering of Syracuse and Tarentum at the end of the 3rd century BC marked the beginning of a flow of Greek art treasures into Rome that continued for several centuries and played a leading role in the aesthetic education of the citizens.

Literature shows that by the middle of the 2nd century BC the Roman forum was thronged with honorific statues of Roman magistrates, which, although none of them has survived, may be assumed to have been carved or cast by Greeks because no native Roman school of sculptors of that time is known. And it is significant that the earliest account of Roman realistic portraits of private individuals is contained in the Greek historian Polybius' description of ancestral *imagines* ("masks") displayed and worn at patrician funerals—a description written about the middle of the 2nd century BC, when the tide of Greek artistic influence was sweeping into Rome and Italy from countries east of the Adriatic, where a highly realistic late-Hellenistic portrait art, which sometimes depicted Roman or Italian subjects, had already blossomed.

Three art forms that expressed the Roman spirit

The first appearance of three art forms that expressed the Roman spirit most eloquently in sculpture can be traced to the Hellenistic Age. These forms are realistic portraiture showing a preference for the ordinary over the heroic or legendary, in which every line, crease, wrinkle, and even blemish was ruthlessly recorded; a continuous style in narrative art of all types; and a three-dimensional rendering of atmosphere, depth, and perspective in relief work and painting. Of these three art forms there is no evidence in the early art of pre-Hellenistic central Italy; and it would be safe to guess that, if Rome had not met them in the homelands of Greek art, it would never have evolved them in its great art of imperial times. But Rome's own contributions to art, if of a different order, were vitally important. Its historical aims and achievements furnished late Hellenistic artists with a new setting and centre, new subjects, new stimuli, a new purpose, and a new dignity. Rome provided the external circumstances that enabled architects, sculptors, painters, and other craftsmen to exploit on a much more extensive scale than before artistic movements initiated in the Hellenistic world, and Rome became a great new patron of art and a great new wellspring of inspiration and ideas.

THE LAST CENTURY OF THE REPUBLIC

Funerary masks

Ancestral *imagines,* or funerary masks, made of wax or terra-cotta, had become extremely individualized and realistic by the middle of the 2nd century BC. The source of this realism is in the impact on Rome of late-Hellenistic iconography; although this use of masks was rooted

J.E. Bulloz

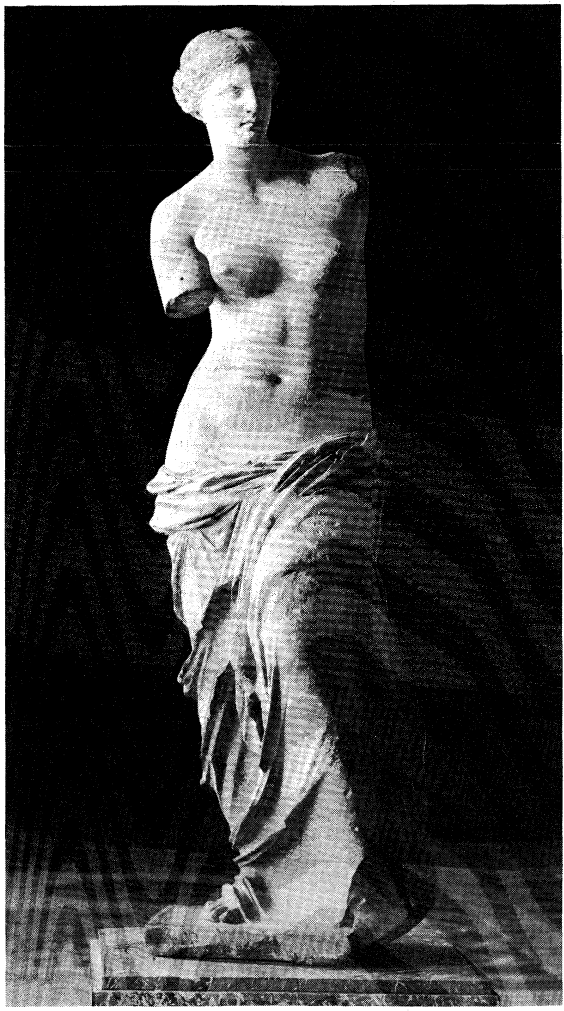


Figure 22: "Venus de Milo," marble statue of Aphrodite from Melos, *c.* 150 BC. In the Louvre, Paris.

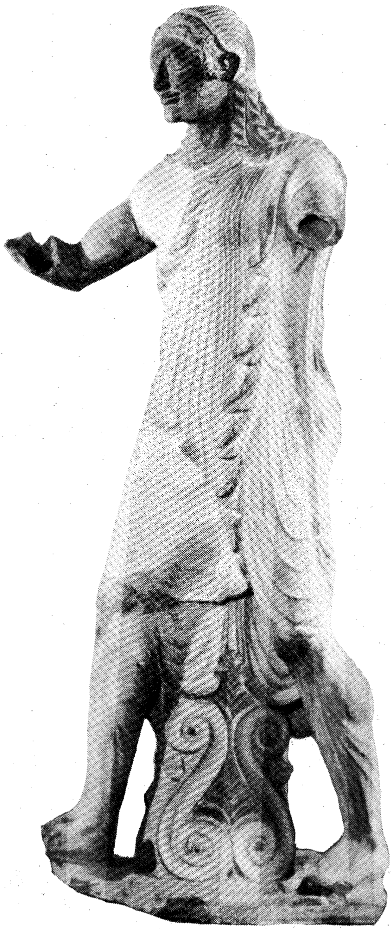


Figure 23: "Veii Apollo," clay statue, *c.* 500 BC. In the Museo Nazionale di Villa Giulia, Rome. Height 1.75 m.
Pellegrini—Crimoldi

in ancient Roman social and religious practice, there is no basis for a belief that the Romans and Etruscans had, from early times, been in the habit of producing death masks proper, cast directly from the features of the dead. It was undoubtedly their funerary customs that predisposed the Romans to a taste for portraits; but it was not until around 100 BC that realistic portraiture, as an art in its own right, appeared in Rome as a sudden flowering, and to that time belong the beginnings of the highly realistic heads, busts, and statues of contemporary Romans—in marble, stone, or bronze—that have actually survived. Coin portraits of public personages, whose names and dates are recorded, greatly assist in determining a chronological sequence of the large-scale likenesses, the earliest of which can be attributed to the period of Sulla (82–79 BC). The style reached its climax in a stark, dry, linear iconographic manner that prevailed around 75–65 BC and that expressed to perfection current notions of traditional Roman virtues; of this manner, a marble head of an elderly veiled man in the Vatican is an outstanding illustration (Figure 24, top left). Shortly thereafter, an admiration for earlier phases of Greek art came into fashion in the West, and verism was toned down at the higher social levels by a revival of mid-Hellenistic pathos and even by a classicizing trend that was to stamp itself upon Augustan portraits. Meantime, in sepulchral custom, the ancestral bust had become an alternative to the ancestral mask, a development exemplified in a marble statue of a man wearing a toga and carrying two such busts in the Capitoline Museums at Rome (Figure 24, bottom); and portrait busts and figures carved on numerous stone and marble grave stelae (slabs or pillars used for commemorative purposes), characteristic of the late republican epoch, suggest the persistence of a preference for severe pose in middle-class and humbler circles. Furthermore, there are some 1st-century-BC portraits that suggest that the making of death masks proper (arguably a sophisticated idea) was occasionally practiced at this time (Figure 24, top right). None of the vivid Etruscan portraits, such as a bronze orator popularly called the "Arringatore" (Museo Archeologico) at Florence (Figure 25) and a terra-cotta married pair on the lid of a cinerary chest (for ashes of the dead) in the Museo Etrusco Guarnacci, at Volterra, is earlier than *c.* 100 BC; works of that type may be reckoned as provincial imitations of the new metropolitan, 1st-century-BC portrait style.

Narrative reliefs

There are no narrative reliefs from Rome that can confidently be assigned to a date before 100 BC. The only definitely dated 2nd-century-BC relief depicting an episode from contemporary Roman history, a frieze with the Battle of Pydna on Lucius Aemilius Paulus' victory monument at Delphi, was worked in 168 BC in Greece. The most familiar republican example of this form of art as practiced in the West is frieze decoration (partly in the Louvre, and

By courtesy of (top right) the Deutsches Archaologisches Institut Rom; photographs, (top left) Anderson—Alinari from Art Resource/EB Inc., (bottom) Alinari—Art Resource/EB Inc.

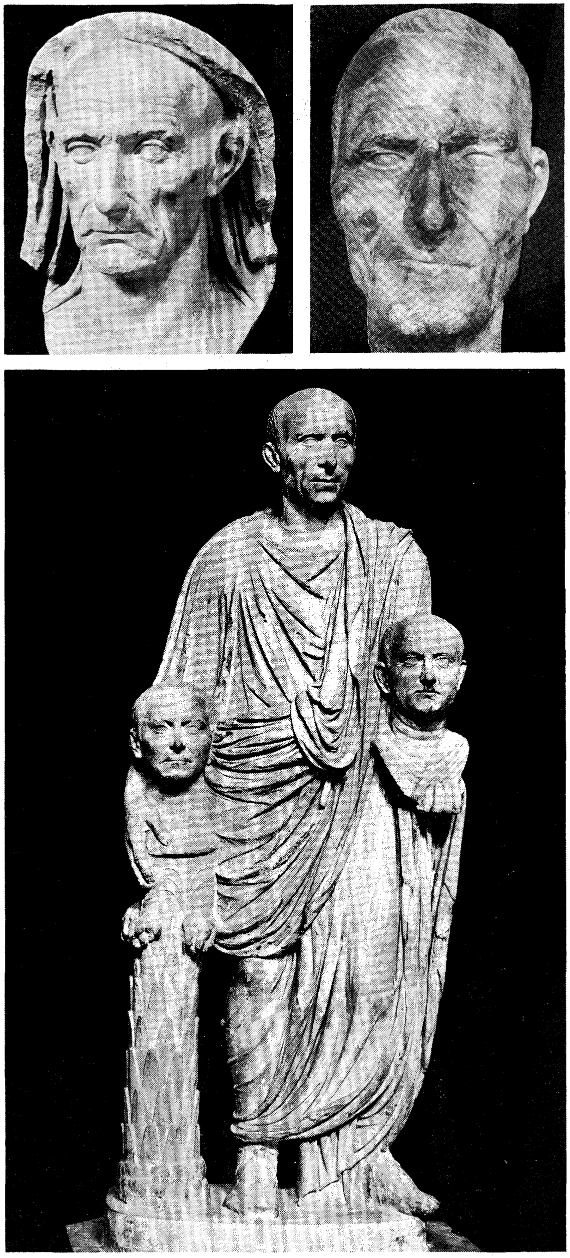


Figure 24: *Roman marble portraits of the Republic.*
(Top left) Head of an elderly veiled man, *c.* 75-65 BC. In the Vatican Museums, Rome. Life-size. (Top right) Portrait in death-mask style, 1st century BC. In the Museo di Antichità, Turin, Italy. Life-size. (Bottom) A Roman patrician with the busts of his ancestors, 1st century AD. In the Capitoline Museums, Rome. Life-size.

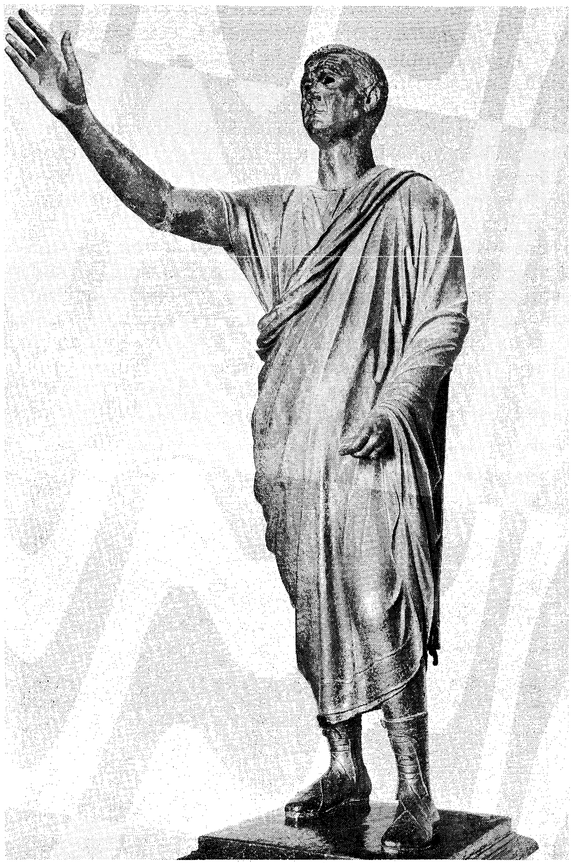


Figure 25: Bronze statue of an orator ("Arringatore"), *c.* 150 BC. In the Museo Archeologico, Florence. Height 1.80 m.

Anderson—Alinari from Art Resource/EB Inc.

partly in the Glyptothek at Munich) from the so-called Altar of Ahenobarbus, which has been shown to have no sure connection either with an altar or with any of the Ahenobarbi. In these, prosaic documentation of Roman census procedure is juxtaposed with depictions of Greek sea nymphs, a conjunction of literalism and borrowed poetry typical of subsequent Roman art.

Funerary narrative sculpture

Funerary narrative sculpture of the late republic is exemplified in a monument of the Julii, at Saint-Rémy (Glanum), France. The base of this structure carries four great reliefs with battle and hunt scenes that allude not only to the mundane prowess of the family but also to the otherworldly victory of the souls of the departed over death and evil, since figures of the deceased, accompanied by personifications of death and victory, merge into one of the battle scenes. It is possible that these highly pictorial reliefs were partly based on lost Hellenistic monumental paintings, for southern Gaul had direct connections with Greek lands east of the Adriatic.

THE EMPIRE

Augustan Age. The hallmark of portraits of Augustus is a naturalistic classicism. The rendering of his features and the forking of his hair above the brow is individual. But the Emperor is consistently idealized and never shown as elderly or aging. A marble statue from Livia's Villa at Prima Porta (in the Vatican), which presents him as addressing, as it were, the whole empire, is the work of a fine Greek artist who, while adopting the pose and proportions of a classical Hellenic statue, perfectly understood how to adopt these to the image that Augustus cultivated as emperor (Figure 26). On his ornate cuirass (armour protecting the chest and back), Augustus' aims and achievements are recorded symbolically in a series of figure groups. A marble portrait statue found on the Via Labicana (Museo Nazionale Romano) represents the Emperor as heavily draped and veiled during the act of sacrificing as *pontifex maximus* ("chief priest"); and a bronze head from Meroe in The Sudan (British Museum), the work of a Greco-Egyptian portraitist, depicts him as a Hellenistic king. Of the female portraits of the period, one of the most charming is a green basalt head (Louvre) of the Emperor's sister, Octavia, with the hair dressed in a puff above the brow and gathered into a bun behind—a popular coiffure in early Augustan times. In many respects, the noblest of all Roman public monuments that were adorned with sculpture is the Ara Pacis Augustae ("Augustus' Altar of Peace"), founded in 13 BC and dedicated four years later (Figure 27). It stood in the Campus Martius and has been restored, with different orientation, not far from its original site. On its reliefs—significantly of Luna marble, a white marble quarried in Italy and not, as had earlier been the case, imported from Greece—it set a standard of distinction surpassed by no later work, with the harmonious blending of contemporary history, legend, and personification, of figure scenes and decorative floral motifs. The altar proper was contained within a walled enclosure, measuring about 38 by 34 feet ($11\frac{1}{2}$ by $10\frac{1}{2}$ metres), with entrances on east and west. On the upper part of the external faces of the south and north precinct walls ran a frieze representing the actual procession (of Augustus, members of his family, officers, priests, magistrates, and the Roman people) to the altar's chosen site on its foundation day (July 4, 13 BC), when sacrifice was offered in thanksgiving for the Emperor's recent return to Rome from the provinces. On either side of the western entrance was a depiction of Augustus' prototype Aeneas sacrificing on his homecoming to the promised land of Italy, and, since Augustus was also hailed as Rome's second founder, a depiction of the suckling of the twins, Romulus and Remus, by the she-wolf. The eastern entrance was flanked by personifications of Roma and of Mother Earth with children on her knees flanked by figures symbolizing air and water (Figure 27, bottom). On the exterior of the

The Ara Pacis Augustae

Alinari—Art Resource/EB Inc.

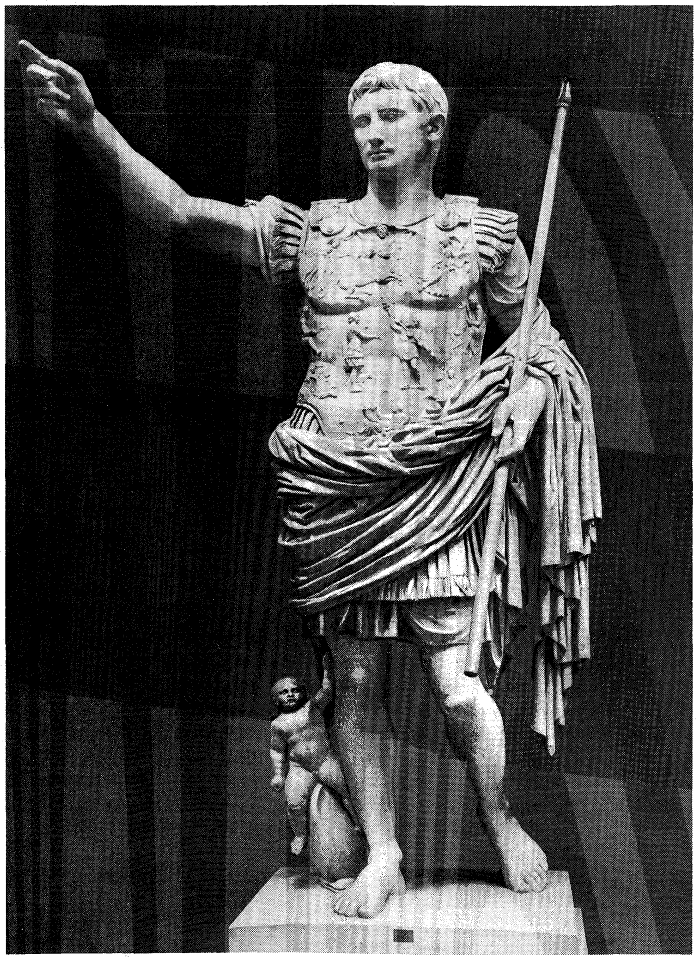


Figure 26: Augustus of Prima Porta, marble statue, *c.* 20 BC. In the Vatican Museums, Rome. Height 2.03 m.

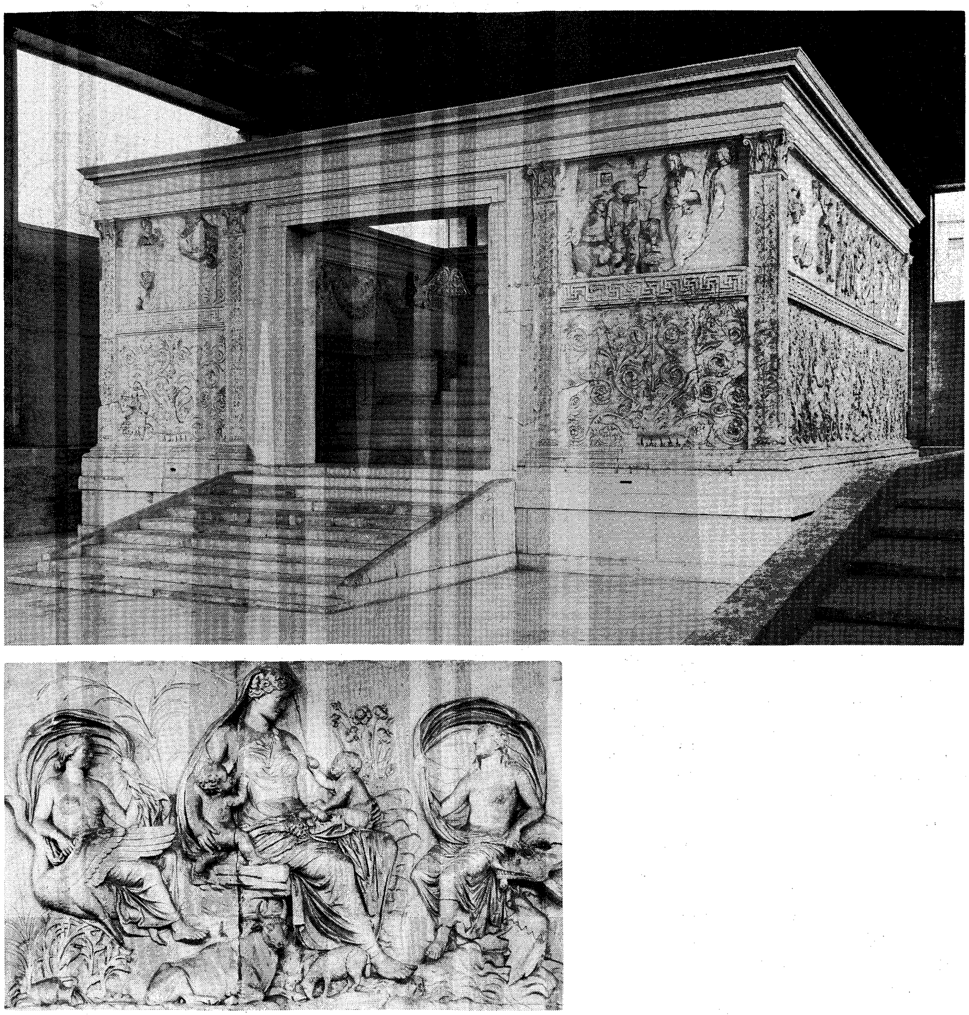


Figure 27: Ara Pacis, Rome, 13 BC. (Top) View of whole altar. (Bottom) "Mother Earth with Air and Water," marble relief on the east exterior wall of the Ara Pacis. Height 1.57 m.

(Top) Alinari—Art Resource/EB Inc., (bottom) Fototeca Unione, Rome

walls, beneath all these figure scenes, was a magnificent dado filled with a naturalistic pattern of acanthus, vine, and ivy, perhaps a translation into marble of a gorgeous carpet or tapestry used in the ceremony. Swags of fruit and flowers that decked the interior faces of the precinct walls may represent real swags that were hung on the temporary wooden altar erected for the foundation sacrifice. The procession was continued in a much smaller frieze on the inner altar, from which figures of Vestal Virgins and of sacrificial victims and their attendants have been preserved. Delightful studies of imperial and other children and such homely incidents as conversations between persons taking part in the procession introduce an element of intimacy, informality, and even humour into this solemn act of public worship. The Ara Pacis, in fact, sums up all that was best in the new Augustan order—peace, serenity, dignity without pompousness, moderation and absence of ostentation, love of children, and delight in nature. The style of the altar's floral decoration strongly suggests that the sculptors who carved it were Greeks from Pergamum.

Julio-Claudian period. The imperial portraiture of Tiberius and Caligula was generally precise but academic work, but some of the female court portraits reflect not only the fashions for elegant simplicity and extreme elaboration in female coiffure but also a subtle poetry. Two possible extremes of tone are clearly marked by the contrasting busts of Claudius and Nero, the former uncomfortably uncompromising, the latter flatteringly Hellenic. In the relatively few public monuments dating from this period to include sculpture, none reveals any novel development.

Flavian period. In the emperor Vespasian's portraits, something of the old, dry style returned. This can be observed in his striking likeness on one of two historical reliefs (Vatican Museums) that were unearthed in Rome near the Palazzo della Cancelleria. A similarly sketchy and impressionistic handling of the hair is found on the emperor Titus' portraits, whereas the third Flavian emperor, Domitian, affected a more pictorial hairdo in imitation of the coiffure introduced by Nero. Still more picturesque are the female hair styles of the time, which display piles of corkscrew ringlets or tight, round curls (Figure 28). The Cancelleria reliefs date from the close of Domitian's reign and depict, respectively, Vespasian's triumphal entry and reception in Rome in AD 70 and Domitian's *profectio* ("setting out"), under the aegis of Mars, Minerva, and Virtus, for one of his northern wars. They are worked in a two-dimensional, academic, classicizing style that is in marked contrast with the vivid, three-dimensional rendering of space and depth, with brilliant interplay of light and shade, on the panels of the Arch of Titus in the Roman Forum. The latter reliefs, which present two excerpts from Titus' triumph in Palestine, were carved in the early 80s. The late Domitianic classicizing manner appears again in the frieze of the Forum Transitorium, which the emperor Nerva completed. This conflict of relief styles within the Flavian period is but one illustration of the ceaseless, unpredictable ebb and flow of different aesthetic principles throughout the history of imperial art.

Age of Trajan. In portraits of Trajan, a deepening of the bust, which was already seen in the later Flavian period, was carried a stage further; there is a new fluidity in the molding of the face; in the hair, which is plastered down across the brow, there is a partial revival of the late republican linear style. Aesthetically, one of the finest known likenesses of the Emperor is a marble head from

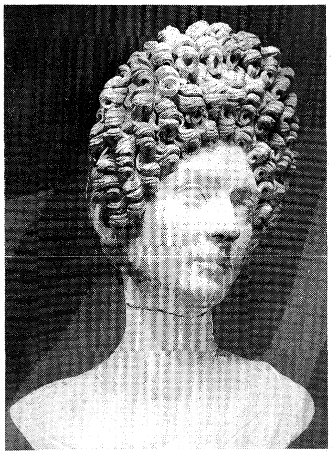


Figure 28: Portrait of a woman of the Flavian period, marble, *c.* AD 90. In the Capitoline Museums, Rome. Life-size.

Cesare Faraglia

Ostia (Ostia Museum). On his monumental column there is a series of less idealized and probably more faithful renderings of his features. The coiffures of Trajanic ladies are, if anything, even more elaborate and extravagant than those of their Flavian predecessors.

Reliefs of Trajan's Column

The reliefs of Trajan's Column, illustrating the two Dacian campaigns of 101–102 and 105–106 and winding up the shaft in a spiral band of Parian marble three feet (one metre) wide, are generally recognized to be the classic example of the continuous method of narration in Roman art. The episodes merge into one another without any punctuation, apart from an occasional tree; Trajan appears again and again in different situations, activities, and costumes. A statuesque figure of Victory separates the histories of the two wars. There are 23 spirals and about 2,500 figures. A high level of technical accomplishment is maintained throughout, and the interest and excitement of the theme never flag. Since the figures of men and animals had to be distinguished from a distance, they are inevitably overlarge in proportion to their landscape and architectural settings; and in order to avoid awkward empty spaces along the upper edges of the band and to preserve an allover, even, tapestry-like effect, background figures in the scenes are reared in bird's-eye-view perspective above the heads of those in the foreground. These carvings must be visualized as once brightly painted, with weapons and horse trappings added in metal. The sources of the scenes were possibly wartime sketches made by army draftsmen at the front, but the fusing together of those isolated pictures into a single scroll was the work of a single master artist, perhaps Apollodorus of Damascus, who designed the whole complex of Trajan's forum, basilica, and column.

The column (the interior of which contains a spiral staircase; Figure 29) had first been intended primarily as a lookout post for viewing Trajan's architectural achievements—his forum and its adjacent markets, to accommodate which he sliced away the slope of the Quirinal Hill. By the time of its dedication in 113, when the relief bands had been added and an eagle planned for the top of the capital, it had become a war memorial. Finally, it became Trajan's future tomb, crowned by his statue (which was later replaced by that of St. Peter) and containing a funerary chamber for the urns holding his and his consort's ashes.

Arch of Constantine

To the last years of Trajan's reign or to the early years of that of his successor should be attributed four horizontal panels that adorn the main passageway and the attic ends of the Arch of Constantine in Rome. If fitted together they would form a continuous frieze of three main scenes, which are, from left to right, an imperial triumphal entry, a battle, and the presentation to the Emperor of prisoners and the severed heads of captives by Roman soldiers. It seems clear that these sculptures were made between around 115 and 120, perhaps for the Temple of Divus Trajanus and Diva Plotina that was erected by Hadrian just to the north of the column. The presence on this frieze of chain-mail corselets, rarely seen on Trajan's Column, seems to indicate that that type of armour, so common under the Antonines (see below *Antonine and Severan periods*), first came into general use in late Trajanic or early Hadrianic times. These reliefs do not depict realistic fighting, as do those of the column, but a kind of ideal or dramatized warfare, with the Emperor himself participating in the melee and the soldiers wearing plumed and richly embossed parade helmets; the scenes melt into one another with total disregard of spatial and temporal logic.

Trajan's arch at Beneventum

A third example of Trajanic monumental sculpture is the relief decoration of the Arch of Trajan at Beneventum (Benevento), which is covered with pictorial slabs, the subjects of which are arranged to carry out a carefully balanced and nicely calculated order of ideas. Those on the side facing the city and on one wall of the passageway present themes from Trajan's policy and work for Rome and Italy; those on the side toward the country and on the other wall of the passageway allude to his achievements abroad. With two exceptions, where a pair of scenes forms a single picture, each panel is a self-contained unit. The reliefs already show something of the classicizing, two-dimensional character of Hadrianic work. Indeed, it seems likely that, although the arch itself was either decreed or dedicated in 114 or 115, some of the panels in which Hadrian is given a peculiar prominence were not carved until the early years of the latter's principate.

The frieze of a great, circular Tropaeum Trajani, set up in the Dobruja (Romania) to commemorate victories over the Dacians, contains a series of metopes (a decoration in a Doric frieze) carved with figure scenes in a naïve, flat, linear style that suggests the hands of army artists of provincial origin.

Age of Hadrian. In the iconography of the age of Hadrian, certain Hellenizing features—the wearing of a

Anderson—Alinari from Art Resource/EB Inc.

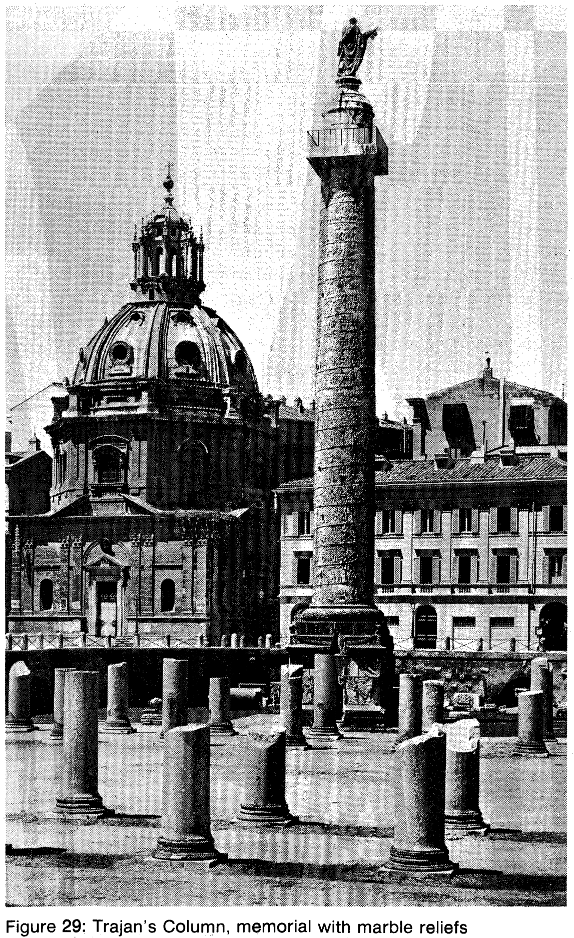


Figure 29: Trajan's Column, memorial with marble reliefs illustrating the two Dacian wars of 101–102 and 105–106; AD 106–113. In Trajan's Forum, Rome.

short Greek beard by the males and the adoption by the females of a simple, classicizing coiffure—are harmonized with new experiments. The depth of the bust increases, there is greater plasticity in the modelling of the face, the men's curly hair and beards are pictorially treated, and the irises and pupils of the eyes are marked in. Many marble portraits of the Emperor survive from all over the empire, but of his likenesses in bronze only one is extant—a colossal head recovered from the Thames River in London (British Museum), torn from a statue erected in the Roman city and probably the work of a good Gaulish sculptor. Portrait statues of Hadrian's Bithynian favourite, Antinoüs, reveal a conscious return in the pose and proportions of the body to Classical Greek standards, combined with a new emotionalism and sensuousness in the rendering of the head.

Monumental reliefs of the age of Hadrian

The monumental reliefs of Hadrian's day cannot vie with those of his predecessors. The most interesting and perhaps the earliest of them are two horizontal slabs once exposed in the Roman Forum but later transported to the shelter of the Curia. Both carry on one side similar figures of victims for the Suovetaurilia sacrifice and on the other side different historical scenes: in the one case, Hadrian doling out the *alimenta* ("poor relief") to Roman citizens, in the presence of a statuary group of Trajan and Italia with children; in the other case, the burning of debt registers. At one end of each of these scenes is carved a figure, on a base, of the legendary Greek musician Marsyas, whose statue in the Forum may once have been in part enclosed by the panels. In the background of both historical pictures are carved in low relief various buildings in the Roman Forum that can be identified. The two scenes display the characteristically Hadrianic two-dimensional style, as do three large panels (Palazzo dei Conservatori, Rome), with the Emperor's head restored and depicting an imperial triumphal entry, an *adlocutio,* and an apotheosis, respectively—somewhat rigid, academic works. Eight medallions gracing the Arch of Constantine give pleasantly composed and lively, if Hellenizing, pictures of sacrifice and hunting (Figure 30). Some of them depict Antinoüs accompanying the Emperor, whose portraits have been recut as likenesses of Constantine the Great and of his colleague Licinius. Finally, historical reliefs found at Ephesus (now in the Neue Hofburg, Vienna)—one of the very few examples of provincial state reliefs that have survived—may be claimed as late Hadrianic (not as of the period of Marcus Aurelius, to which many critics have assigned them).

In Rome and Italy during the second quarter of the 2nd century, interment began to supersede cremation as a method of disposing of the dead, and Hadrian's reign saw the beginnings of a long line of carved sarcophagi that constituted the most significant class of minor sculptures down to the close of the ancient Greco-Roman world.

Antonine and Severan periods. Portraits of Antonine imperial persons, of which a bronze equestrian figure of Marcus Aurelius on the Capitol (Figure 31) and a great marble bust of Commodus as Hercules in the Palazzo dei Conservatori are perhaps the most arresting examples, display a treatment of hair and beard, deeply undercut and drilled, that grew ever more pictorial and baroque as the 2nd century advanced. This produced an impression of nervous restlessness that contrasts with the still, satin smoothness of the facial surfaces, particularly in the iconography of Commodus. To all this picturesqueness, Septimius Severus added yet another ornamental touch—the dangling, corkscrew forelocks of his patron deity, Sarapis. The female hairstyles of the time are characterized first by a coronal of plaits on top (Faustina the Elder), next by rippling side waves and a small, neat bun at the nape of the neck (Faustina the Younger, Lucilla), and then by stiff, artificial, "permanent" waving at the sides and a flat, spreading "pad" of hair behind (Crispina, Julia Domna).

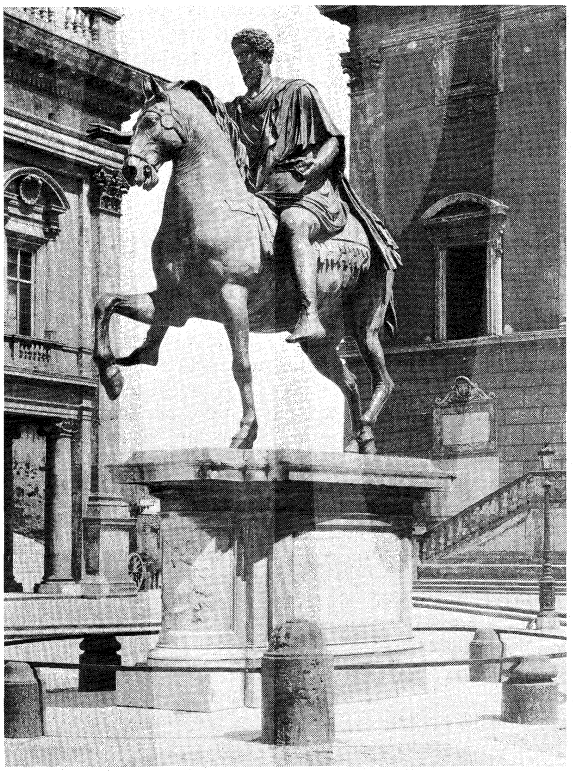


Figure 31: Bronze equestrian statue of Marcus Aurelius, in the Piazza del Campidoglio, Rome, *c.* AD 173. Height 5.03 m.

Alinari—Art Resource/EB Inc.

State reliefs

Of the state reliefs of this epoch, the earliest are on the base (in the Vatican) of a lost column set up in honour of Antoninus Pius and Faustina the Elder. The front bears a dignified, classicizing scene of apotheosis: a powerfully built winged figure lifts the Emperor and Empress aloft, while two personifications, Roma and Campus Martius, witness their departure. On each side is a *decursio,* or military parade, in which the riders farthest from the spectator appear not behind the foot soldiers but high above their heads—a remarkable instance of the bird's-eye-view per-

Alinari—Art Resource/EB Inc.

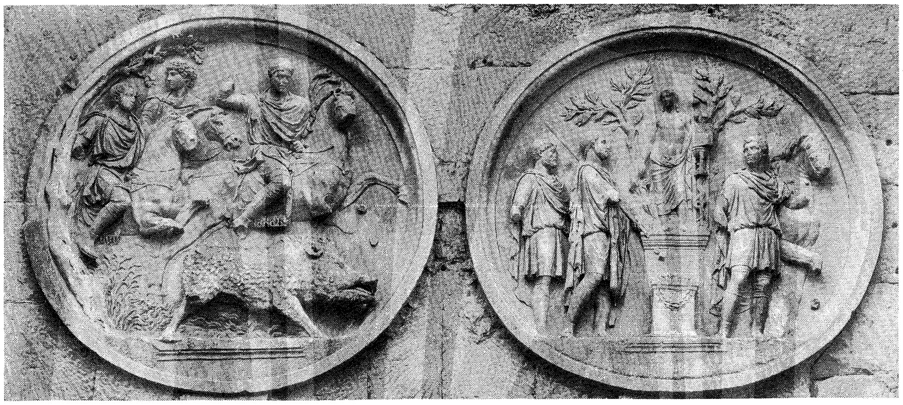


Figure 30: Medallions from the Arch of Constantine, Rome; medallions date from AD 117–138.

spective carried to its logical conclusions. All the figures in these side scenes are disposed on projecting ledges, a device employed again about 20 years later on Marcus Aurelius' Column. Eleven rectangular sculptured panels—similar to those on the Arch of Trajan at Beneventum but displaying greater crowding of figures, livelier movement, and a pronounced effect of atmosphere and depth—depict official occasions and ceremonies in the career of Marcus. Three of these are in the Palazzo dei Conservatori, Rome (Figure 32); the other eight are on the attics (low stories or walls above the cornice of the facade) of the Arch of Constantine. These two sets of panels represent two separate series and may have been carved for two (now lost) distinct triumphal arches. The contrast in style between the spiral reliefs of Marcus Aurelius' Column, put up under Commodus and depicting Marcus Aurelius' northern campaigns, with those of its Trajanic predecessor, is a measure of the change of mood that the Roman world experienced during the course of the 2nd century. The diminished proportions of the squat, doll-like figures, their herding together in closely packed, undifferentiated masses, their angular, agitated gestures, and the stress laid throughout on the horror and tragedy of war suggest that the empire is facing an unknown future with diminished security and that man is at the mercy of some unaccountable power, the supreme embodiment of which is an awe-inspiring winged, dripping figure, personifying the rainstorm that saved the Roman army from perishing from thirst. Again, in the imperial *adlocutiones* that punctuate this frieze, where the Emperor stands in a strictly frontal pose high above the heads of his audiences, can be seen a remarkable return (but probably not a conscious return) to the conventions employed in primitive art for expressing the concept of the ruler as transcendental being.

The spirit of the times is reflected no less vividly in carved sarcophagi. Their themes—familiar myths, battles, hunts, marriages, and so on—allude allegorically to death and the destiny of the soul thereafter. The classicizing, statuesque tradition is also maintained in late 2nd- and early 3rd-century columned sarcophagi, originating in the workshops of Asia Minor but freely imported into, and sometimes imitated in, Rome and Italy. On such pieces single figures or small groups of figures occupy niches between colonnettes. Among the most impressive examples is a great sarcophagus at Melfi, in Puglia, Italy, with a couch-shaped lid, on which the figure of a girl lies prostrate in the sleep of death.

Mansell—Alinari from Art Resource/EB Inc.

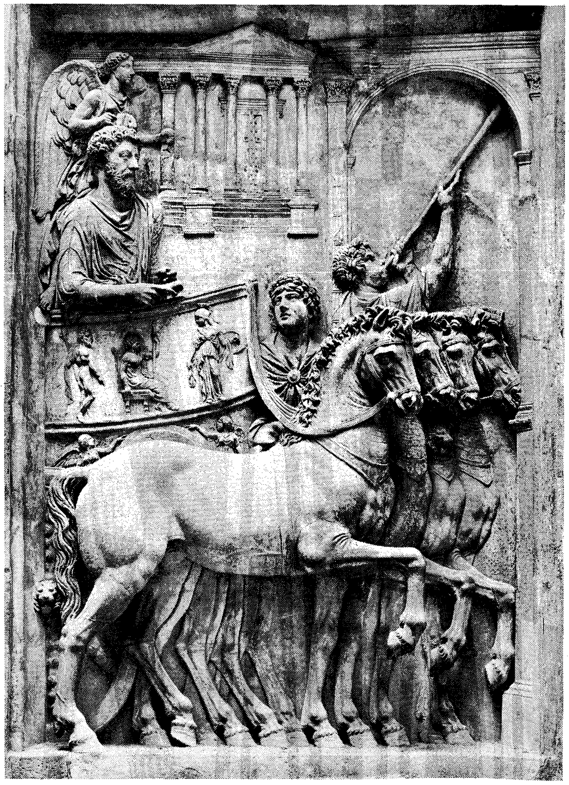


Figure 32: Marcus Aurelius in a quadriga (four-horsed chariot) entering Rome in triumph, from a marble relief in the Palazzo dei Conservatori, Rome, c. AD 176. Height 3.23 m.

The novel features that have been noted in the reliefs of Marcus Aurelius' Column were worked out more completely in those of the official monuments set up to honour Septimius Severus, both in Rome and abroad. In the arch erected in 203 at the northern end of the Roman Forum are found crowded masses of small figures in broad bands of relief, perhaps reflecting a style of documentary painting; in the smaller Porta Argentariorum in Rome, erected by bankers and cattle dealers in honour of the Emperor in the following year, there are stiff, hieratic, funeral poses; and above all in the still more remarkable four-way arch set up at Leptis (Lepcis) Magna in Tripolitania to commemorate a visit of about 203 is a pier decorated with a stylized bird's-eye view of an Oriental city under siege and (also on the piers) weirdly elongated representations of captives. The deeply undercut and drilled vine-scroll ornament here and in the Severan basilica nearby is similar to that found in Asia Minor, whence sculptors had doubtless been imported.

The tension between naturalism and schematization

3rd and 4th centuries. A new tension between naturalism and schematization marks the history of late-antique portraiture. In likenesses of Alexander Severus, the facial planes are simplified, and the tumbling curls of the 2nd-century baroque have been banished in favour of a skullcap treatment of the hair and sheathlike rendering of the beard. Toward the middle of the 3rd century, under Philip the Arabian and Decius, this clipped technique in hair and beard was combined with a return to something of the old, ruthless realism in the depiction of facial furrows, creases, and wrinkles. For a time, Gallienus reinstated the baroque curls and emotional expression, but in the later decades of the century the schematic handling of hair, beards, and features reappeared. Finally, in the clean-shaven faces of Constantine the Great and his successors of the 4th and early 5th centuries, the conception of a portrait as an architectonic structure came to stay; and the naturalistic, representational art of the Greco-Roman world was exchanged for a hieratic, transcendental style that was the hallmark of Byzantine and medieval iconography (Figure 33). The hair is combed forward on the brow in rigid, striated locks, and the eyes are unnaturally enlarged and isolated from the other features. The face is so formalized that the identification of any given portrait becomes a problem. A colossal bronze emperor (near the church of S. Sepolcro, Barletta), for example, has been given the names of several different rulers of the late 4th and early 5th centuries. Throughout these centuries the favourite female coiffure shows a plait or twisted coil of hair carried across the back and top of the head from neck to crown, while under Constantine there was a brief revival of the two Faustinas' styles.

Throughout the 3rd and 4th centuries, carved sarcophagi carry on the story of relief work. Aesthetically, the most notable 3rd-century example is an allover tapestry-like battle piece (Ludovisi Collection, Museo Nazionale Romano, Rome), which possibly was made for Decius' son Hostilian.

Of 3rd-century state reliefs in Rome, virtually nothing has survived. Narrow historical friezes carved for the Arch of Constantine, completed for the celebrations of his *decennalia* (10th anniversary of his reign) in 315, show dwarfish, dumpy, niggling figures. Both these reliefs and those of the slightly earlier Arch of Galerius at Thessalonica look as though they had been worked by artists whose experience had been confined to the production of small-scale sculptures. The last examples of Roman carving are reliefs on the base of an obelisk of Theodosius in the Hippodrome at Constantinople, where the Emperor and members of his court, ranged in rigid, hieratic poses, watch the shows. Few original portions are extant of the spiral relief bands that entwined columns of Theodosius and Arcadius in Constantinople.

Minor forms of sculpture. Of the minor forms of sculpture, none is more attractive than the art of modelling—

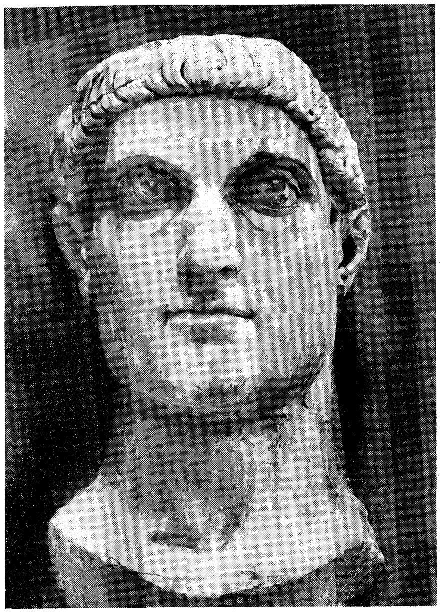


Figure 33: Marble colossal head of Constantine the Great, part of the remains of a giant statue from the Basilica of Constantine (formerly the Basilica of Maxentius) in the Roman Forum, Rome, *c.* AD 313. In the Capitoline Museums, Rome. Height of the head 2.41 m.

Hirmer Fotoarchiv, Munchen

Stucco work

in relief or in the round—in fine, white stucco. Decorative stucco work was cheaper and easier to produce than carving in stone or marble. Soft and delicate in texture, it was equally elegant whether left white or gaily painted. In domestic architecture it was a useful alternative or accessory to painting; notable are such examples as a pure white, exquisite vault decoration showing ritual scenes with small-scale figures, from a late republican or early imperial house near the Villa Farnesina in Trastevere (Museo Nazionale Romano); handsome pairs of large white griffins, framed in acanthus scrolls against a vivid red ground, in the late republican House of the Griffins on the Palatine; and a frieze depicting the story of the *Iliad,* in white figures on a bright blue background, in the House of the Cryptoporticus, or Homeric house, at Pompeii. For the use of this technique in palaces, the figure work in Domitian's villa at Castel Gandolfo in the Alban hills can be cited; it can be found in such public buildings as the Stabian and Forum baths and the Temple of Isis at Pompeii. The loveliest and most extensive stucco relief work in a semiprivate shrine is that in the underground basilica near the Porta Maggiore, Rome, where the scenes all allude to the world beyond the grave, to the soul's journey to it, and to the soul's preparation for it in this life (Figure 34). Some of the best surviving stuccos are in tombs: the tomb of the Innocentii and the tomb of the Axe under the church of S. Sebastiano on the Via Appia; the tombs of the Valerii and the Pancratii on the Via Latina (in the latter, stucco work is attractively combined with painting in the flat); and the tomb of the Valerii under St. Peter's, Rome, where the interior walls of both the main and subsidiary chambers are almost completely covered with recesses, niches, and lunettes (semicircular or crescent-shaped spaces) containing stucco figures. The Vatican tomb of the Valerii must be reckoned as a classic place for the study of this delightful and all too scantily represented branch of Roman art.

Ivory and wood carving

Ivory was another popular material for minor sculpture. It was worked in the round, in relief, and in such forms as small portraits, figurines, caskets, and furniture ornaments, of which the carved plaques composing the throne, or "Cathedra of Maximianus," at Ravenna (probably 6th century) provide a notable instance. The consular and other diptychs comprise one of the most distinctive types of ivory relief work in the 4th and 5th centuries. Among them are masterpieces that kept alive the traditions of Hellenistic carving, such as a diptych of the Symmachi and Nicomachi families (one leaf of which is in the Victoria and Albert Museum, London, and the other in the Musée de Cluny, Paris), and some outstandingly fine examples of late antique portraiture, such as the Probus diptych at Aosta (Cathedral treasury) with a double portrait of Honorius, the Felix diptych in Paris (dated 428), and one of Boethius, consul in 487, at Brescia (Civico Museo dell'Età Cristiana). Fine examples of wood carving are panels with biblical scenes on the 5th-century door of the church of Sta. Sabina on the Aventine.

Gem carving

Many types of carving in precious stones were practiced by Roman-age craftsmen, and it is to them that the credit goes for the majority of intaglios that have survived from ancient times. (Intaglios are engraved or incised figures depressed below the surface of the stone so that an impression from the design yields an image in relief.) The widespread taste for them is reflected in the many existing glass-paste imitations reproducing their subjects, which include portraits of both imperial and private persons, and a large variety of divine and mythological groups and figures, personifications, animals, etc. Many bear the signatures of Greek artists.

The most impressive series of Roman gems consists of cameos representing imperial persons. These are miniature reliefs cut in precious stones with different coloured strata (so that the relief is of a different colour from the ground), whereas intaglios, like the ancient seals mentioned earlier, were reliefs, as it were, in reverse, cut into the surface so that a true relief only emerges from an impression. Among the earliest surviving examples of the great imperial cameos are the Blacas onyx (British Museum, London), portraying Augustus in the guise of Jupiter; the Gemma Augustea, a sardonyx (an onyx with parallel layers of sard) in the Kunsthistorisches Museum, Vienna, and the Grand Camée de France, a sardonyx in the Bibliothèque Nationale, Paris, which were probably carved under Caligula and present, respectively, the apotheosis of Augustus and of Tiberius, the latter with Divus Augustus, also; and a sardonyx cameo of Claudius with Jupiter's aegis (Royal Art Collection at Windsor Castle). Late antique examples of the craft are a rectangular sardonyx (city library at Trier), portraying Constantine the Great and members of his house and an onyx with busts of Honorius and Maria (Rothschild Collection, Paris).

Other varieties of carving in precious stones are represented by a miniature head of a girl (British Museum) wearing the hair style of Messalina and Agrippina the Younger, which is cut in *plasma;* an onyx vase in the Braunschweigisches Landesmuseum, Braunschweig, possi-

Alinari—Art Resource/EB Inc.

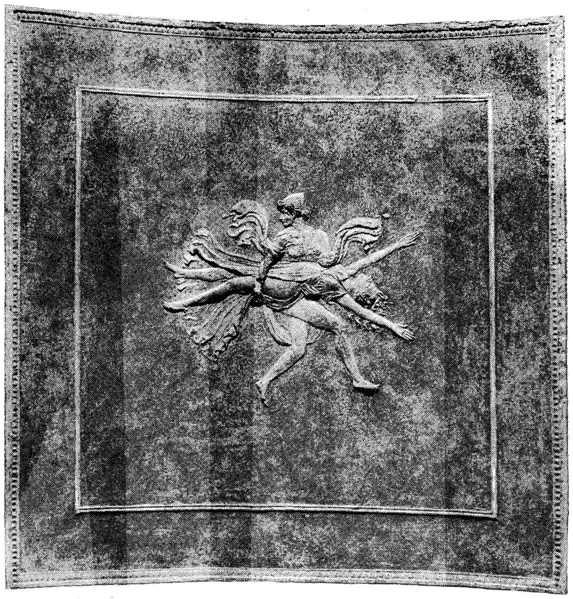


Figure 34: Detail of stucco decoration in a vaulted chamber of a subterranean basilica near Porta Maggiore, Rome, mid-1st century AD.

bly of the 1st century, depicting an emperor and empress as Triptolemus and Demeter; and a late-antique vase, carved in honey-coloured agate and decorated front and back with a naturalistic vine and with the head of Pan, cupped in acanthus, on either shoulder (Walters Art Gallery, Baltimore).

"Cameo glass"

Closely akin to cameos and vessels cut in precious stones are their substitutes in opaque "cameo glass," worked in two layers, with the designs standing out in white against a dark-blue or bright-blue background. To this class belong a blue vase from Pompeii (Museo Archeologico Nazionale), with Cupids gathering grapes; the Auldjo Vase (British Museum, London), with an exquisitely naturalistic vine; and the celebrated Portland Vase, also in the British Museum, the scenes on which have always been the subject of scholarly controversy but are generally supposed to depict myths relating to the afterlife. Similar imitations of carving in precious stones are late antique *diatreta* ("cage cups"), the decoration of which is cut back from the outer surface of the mold-cast blank. This openwork ornamentation sometimes represents the crisscross meshes of a net, while on other vessels it consists of an elaborate figure scene, the design in either case being very deeply undercut and, for the most part, only connected with the background by short shanks of glass. Of the figured examples, the most spectacular surviving specimens are a dark-blue *situla* ("bucket") with a hunting scene (treasury of St. Mark's, Venice) and a dull-green cup presenting the story of Lycurgus ("Rothschild Vase," British Museum). Of the other types of glass with figured decoration, molded cups with gladiatorial and circus scenes are characteristic of the early-imperial period; and the 4th-century glassworker's craft is represented by vessels with cut or incised designs. Among the most important centres of glass production under the empire were Syria, Alexandria, and the Cologne region.

Figured pottery

Figured terra-cotta tablewares (terra sigillata—a term often incorrectly stretched to cover plain wares) were cheaper versions of costly decorated silverwares. During the last century of the republic and in the early decades of the 1st century of the empire, Arretium (Arezzo) was the most flourishing centre of the manufacture of a fine type of red-gloss pottery (Figure 35). As signatures on the pots reveal, Italian firms often employed Greek and Oriental craftsmen, and the mythological and floral themes of the vessels' molded ornamentation owe much to the inspiration of Hellenistic art.

By courtesy of the trustees of the British Museum

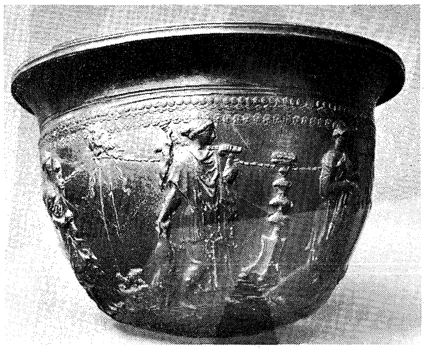


Figure 35: Arretine ware bowl with a design of the seasons in relief. Made in the factory of Gnaeus Ateius, *c.* 10 BC. In the British Museum. Height 17.8 cm.

Gaulish pottery

From shortly before the mid-1st century AD onward, the markets enjoyed by Italian fabrics were captured by the products of potteries now established in southern, central, and eastern Gaul. These manufactured cheaper, more mass-produced, and aesthetically inferior red-gloss and black-gloss wares, popularly known as "Samian," some varieties of which continued into the 4th century. The decoration of Gaulish pots was, for the most part, molded; but some vessels carry applied motifs made in individual molds, and others show designs incised to counterfeit cut glass. Yet another type of ornament was carried out in the barbotine technique, by which relief work was produced by trailing liquid clay across the surface of the pot. As regards the content of the decoration, themes from daily life were added to traditional subjects based on Greco-Roman mythology and on natural history. The *E* barbotine hunt cups (produced mainly at Castor, Northamptonshire) are the highlight of the native Romano-British potter's craft.

A late-antique class of red-gloss pottery, known as late A ware, with scenes in relief from Greek mythology and from Roman spectacles, was manufactured in a southern Mediterranean area, probably Egypt. (J.M.C.T./Ed.)

EARLY CHRISTIAN

Early in the 20th century it was thought that Christian art began after the death of Christ or, at least, in the second half of the 1st century AD. But later discoveries and studies showed that a truly Christian art—that is, with a style quite distinctive from Pagan Roman art—did not exist before the end of the 2nd or beginning of the 3rd century. When it ended, or rather developed into something else, is harder to say. Early Christian art penetrated all the provinces of the Roman Empire, adapting itself to existing pagan art. It subsequently created its own forms, which varied according to local stylistic evolution. The new capital at Constantinople (ancient Byzantium), founded by the emperor Constantine the Great (306–337), was to be an important centre of art. The art produced there, now known as Byzantine art, extended throughout the entire Christian East. It is customary to distinguish early Christian art of the West or Latin part from the Christian arts of regions dominated by the Greek language and to consider the latter as proto-Byzantine, while acknowledging, however, a certain latitude in the initial date of this separation: 330, the foundation of Constantinople; 395, the separation of the Greek part of the empire from its Latin sector; or, finally, the reign of Justinian (527–565). The transition from the earlier to the later art discussed in the next sections took place at different times in different locations; therefore, there can be no precise chronological boundary. Only after Justinian's reign did many Eastern regions submit to the ascendancy of the art of Constantinople, following until the 6th and even the 7th century the paths traced by Christian art in its beginnings. In the West the end of Early Christian art is easier to determine. Closely tied to Roman art, it finished with the collapse of the empire at the end of the 5th century. Then, transformed into a multitude of regional art styles, it assimilated various influences from the East and from the barbaric peoples who superseded their Roman masters.

The vague boundaries of this art in time and space make a definition of its character difficult. Its style evolved from the current Greco-Roman art. The new elements lay not in form but in content: places of worship very different from pagan temples, iconography drawn from the Scriptures. As the hold of the church over public and private life grew, these new elements tended to set traditional subjects completely aside. Early Christian art, while deeply rooted in Greco-Roman art, became a new entity, as distinct from ancient art as from that of the Middle Ages. An obvious difference is the absence of monumental public sculpture. Early Christian sculpture was limited to small pieces and private memorials and only gradually became incorporated into ecclesiastical architecture.

Sarcophagi. The imagery of sarcophagi followed an evolution similar to that of the catacomb paintings. The same biblical and Gospel subjects were introduced into pagan or neutral compositions. In the second or third quarter of the 3rd century, the oldest Christian sarcophagi were hardly distinguishable from the pagan. On one at Sta. Maria Antiqua, Rome, a seated philosopher reading a scroll, a praying figure, and a "Good Shepherd" are "Christianized" by the scenes that accompany them on either side: Jonah resting and the Baptism of Christ. Thus, a sarcophagus from the Via Salaria (Rome, Vatican Museums), which represents the same subjects except for the truly Christian scenes, can be called "Christian" only with reservation.

During the 4th century this iconography was enriched and became more strictly narrative; the miracles of Christ, fully described, were included, the crossing of the Red Sea

Further development of Christian iconography on carved sarcophagi

was often depicted in a long frieze, and the episodes of the Passion of Christ—his arrest, his trial before the Jewish council, his presentation to Pilate, and the Way of the Cross—often extended along the faces of the sarcophagi. The Crucifixion itself was represented only by a bare cross, surmounted by a crown enclosing the monogram of Christ: thus, the symbolic image of the triumph over death. This hesitation to portray the dead Christ on the Cross, an ignominious mode of punishment reserved by the Romans for slaves and abject criminals, disappeared only gradually during the course of the 5th and 6th centuries.

The largest group of Early Christian sarcophagi was found in Rome and its vicinity, although others were found elsewhere in the Mediterranean region. The classicizing style of the first half of the 3rd century became vulgar and a little crude around 300, but it became progressively refined in the time of Constantine and his sons. To the years from 340 to 370 belong the best Roman works: the sarcophagi called the "Two Brothers" (Museo Cristiano), that of Junius Bassus, dated 359, another with columns (both in the grotto of St. Peter's, Rome), that of the "Three Good Shepherds" (Vatican Museums), and, finally, one in S. Sebastiano, Rome, which contains several rare scenes from the story of Lot. While bearing witness to a renaissance of Classical style, they are laden with a new spirituality. A final flourishing occurred near the end of the 4th century in Milan with the decoration of a sarcophagus (S. Ambrogio), which combined an elegant finesse in the figures (due probably to Greek influence) to the vigour of the Roman style.

The sarcophagi of the Middle East and of Ravenna belong principally to the 5th and 6th centuries and to a different artistic tradition. Those of Constantinople and of Asia Minor are fewer in number and lack stylistic homogeneity. Several examples (*e.g.,* sarcophagus of a child and another of the Apostles, end of the 4th century, both in the Arkeoloji Müzeleri, Istanbul) have a harmonious beauty inspired by Classical Greek art; others are in a totally different and more popular style. The sarcophagi of Ravenna, which first appear at the end of the 4th century, stand midway between the Greek art of the East and Latin art. That of Bishop Liberius (4th–5th centuries) of Ravenna at the church of S. Francesco is close to the classicizing Roman sarcophagi in the handling of figures, while the composition—Christ and the Apostles isolated under arcades—finds its models in Asia Minor. Successive waves of Eastern influence affected local style, producing in the 5th century an art distinct from that of the rest of Italy and the Middle East.

Ivory carving. The Christianization of the decorative arts was a slower process than that of monumental art. The presence of pagan imagery on small, movable objects, usually intended for secular use, was less shocking than on the walls and floors of religious buildings. Because many of these objects were made of precious materials, most of them have disappeared. Only ivories are preserved in considerable number. On a small coffer from Brescia (Civico Museo Romano), second half of the 4th century, Gospel scenes cover the four sides and the top, surrounded by a border of biblical subjects similar to those whose presence has been noted in the paintings of the catacombs and on the sarcophagi. The figures are characterized by a gentle beauty and are close to those of certain Roman sarcophagi of the middle and third quarter of the 4th century. Ivories such as the holy women at the tomb and the Ascension of Christ in the Museo d'Arte Antica, Castello Sforzesco in Milan and in the Bayerisches Nationalmuseum in Munich (Figure 36); six miracles of Christ, divided between the two leaves of a diptych, now in Berlin and in Paris; a coffer in London (British Museum) that bears one of the oldest, if not the oldest, representations of Christ on the cross; and a reliquary found at Pola, Istria, Yugoslavia—all belong to a group of ivories that were produced either in Rome or in northern Italy from the end of the 4th to the middle of the 5th century. In the second half of the 5th century the quality of ivory carving declined in the West; it improved, however, at Constantinople and perhaps other eastern cities, such as Antioch and Alexandria.

Ivories in classical style

(He.S.)

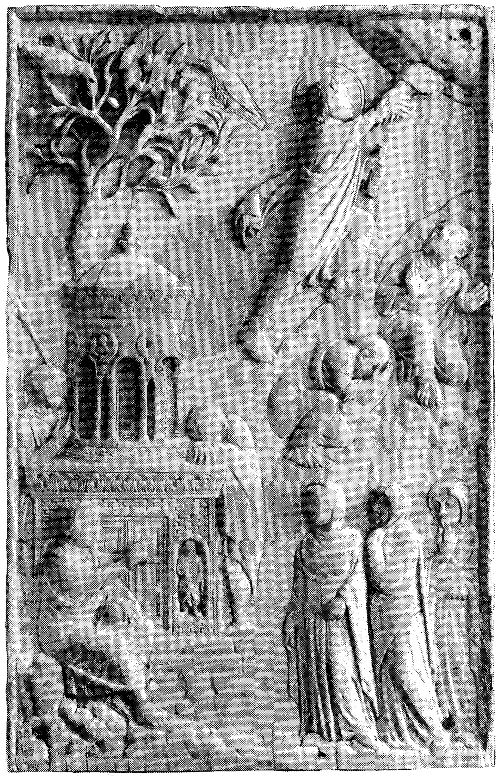


Figure 36: "The Woman at the Tomb and the Ascension," ivory, *c.* 400. In the Bayerisches Nationalmuseum, Munich. 18.7 cm × 11.7 cm.

By courtesy of the Bayerisches Nationalmuseum, Munich

The Middle Ages

EASTERN CHRISTIAN

The Byzantine era really began with the transference of the capital of the Roman Empire from Rome to the site of ancient Byzantium on the Bosporus in the year AD 330, the new capital thereafter being called Constantinople, after its founder, the emperor Constantine I. Constantine had 17 years earlier been responsible for recognizing Christianity, and from the outset he made it the official religion of the new city. The art dedicated to the service of the faith, which had already begun to develop in the days when Christians were oppressed, received official recognition in the new centre and was also subjected to a number of new influences, so that it owed a debt on the one hand to Italy and Rome and on the other to Syria and Asia Minor, where Oriental elements were prominent. It must not be forgotten that the population of Constantinople and its neighbourhood was Greek, not Latin, so that the poetic and philosophical outlook of the Greek world was itself a very considerable influence.

Constantinople and the Byzantine Empire. Sculpture underwent changes very similar to those in architecture. The decorative work in Hagia Sophia illustrates its nature. In the Classical world naturalistic representation had prevailed; at Hagia Sophia the forms are still basically representational, but they are treated in an abstract manner, more advanced in degree than at St. Polyeuktos. Capitals of the period are similarly stylized even when they use bird or animal forms, for these are usually treated as part of an overall balanced pattern. With this tendency toward stylization in architectural sculpture, it is not surprising to find that three-dimensional, representational sculpture was progressively going out of fashion. Portrait sculptures had been made of most of the early emperors, and the texts report that a mounted figure of Justinian I topped a column in front of Hagia Sophia. But that was the last of the series; figural compositions in high relief had adorned sarcophagi, and similar reliefs had found a place on the walls of churches, but virtually none of these dates from later than Justinian's reign. Instead, flat slabs with low-

Developing stylization in architectural sculpture relief ornament akin to that on the capitals and cornices of Hagia Sophia, some of it even purely geometric, came into vogue. These slabs were used for the lower sections of windows or to form a screen between the body of the church and the sanctuary; they were later to develop into the high structures called iconostases, which eventually became universal in Orthodox churches.

Ivories. The minor sculptural arts are essential to any treatment of medieval sculpture in general, partly because more is known about them and partly because some of the most able masters of the period preferred to work on small-scale objects, and patronage was ready to support them. Most important are the ivories. They comprise a wide variety of types, ranging from small pyxides—circular vessels used in the liturgy—to large-scale works made up of a number of separate panels, like the famous throne of Maximian (Figure 37), the Archbishop of Ravenna, at Ravenna (*c.* 550; Museo Arcivescovile, Ravenna). Most usual, however, were the flat plaques used as diptychs, book covers, etc. Considerable numbers of these, dating mostly from the late 5th and early 6th centuries, have been preserved. After about the middle of the 6th century, however, ivories become rarer: very few can be dated to the period between the reign of Justinian and the revival of Byzantine art in the 9th century.

Diptychs, or two-panel ivories, seem to have been very popular both for use as book covers and for ceremonial purposes. The most impressive of them were imperial. In these each leaf was made up of five panels; on the central one was a portrait of the emperor; at the sides were standing figures of the consuls; below were scenes, usually of tribute bearers; and above were angels upholding a bust of Christ. They thus illustrated the Byzantine ideas of hierarchy, Christ above and the world below, dominated by the emperor as Christ's vice-regent. The finest of them, known as the Barberini ivory (Figure 38, left), is in the Louvre and probably depicts Anastasius I (491–518); another, of his wife, the empress Ariadne, is divided between several collections.

Consular diptychs

More numerous today are the diptychs that were issued by the consuls on coming to office. Their fabrication ceased when the office of consul was abolished by Justinian in 541; though by no means are all the consuls portrayed before that time, leaves of the diptychs issued by a large number of them survive. Each leaf consisted of a single plaque. The earlier ones, like that of Probus (408), are still Roman in style; but those dating from just before and just after 500, which constitute the majority, are in a different style, either more ornate or very much simpler. The more elaborate ones are well represented by leaves of the consul Flavius Anastasius (517), in the Cabinet des Médailles, Bibliothèque Nationale, Paris (Figure 38, right); they show the consul enthroned, with lively circus scenes below. The plainer type is represented by a consular diptych of Justinian dated 521 (six years before his accession as emperor), now exhibited in the Castello

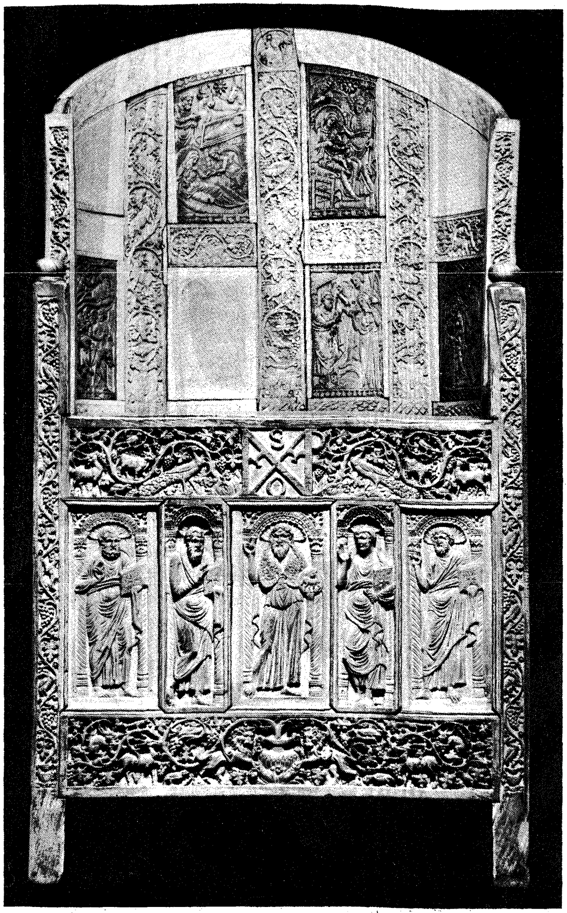


Figure 37: Ivory throne of Maximian, Archbishop of Ravenna, *c.* 550. In the Museo Arcivescovile, Ravenna. 149.9 × 60.0 cm.
Hirmer Fotoarchiv, Munchen

By courtesy of (left) the Louvre, Paris; photograph (right) Hirmer Fotoarchiv, Munchen

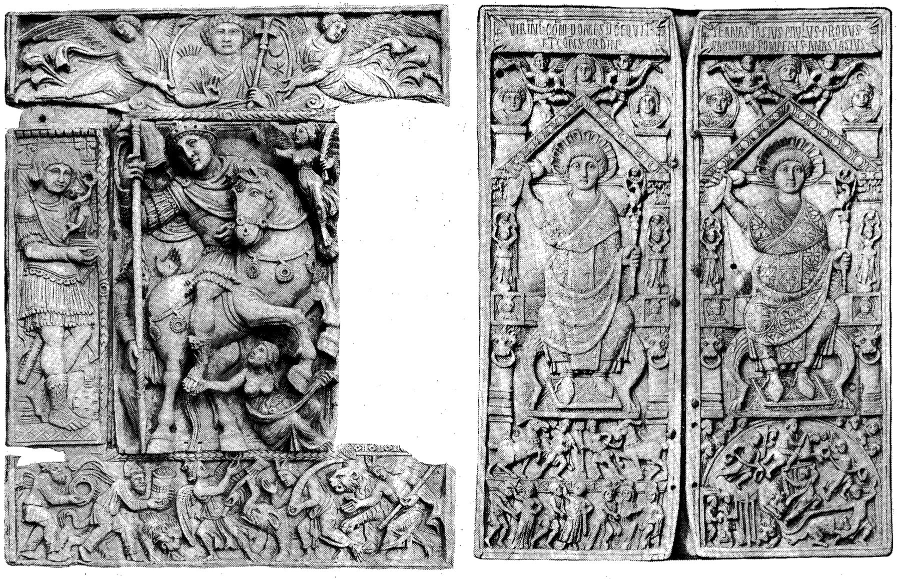

VIR INL COM DOMEST EQ ET CONS ORDIN
FL ANASTASIUS PAULUS PROBUS SABINIAN POMPEIUS ANASTASIUS

Figure 38: *Early Byzantine ivories.*
(Left) Leaf of the Barberini diptych showing mounted emperor with tribute bearers, *c.* 500. In the Louvre, Paris. 35.0 × 26.7 cm. (Right) Diptych of Flavius Anastasius with the consul enthroned and circus scenes below, 517. In the Cabinet des Médailles, Bibliothèque Nationale, Paris, 35.6 × 25.4 cm.

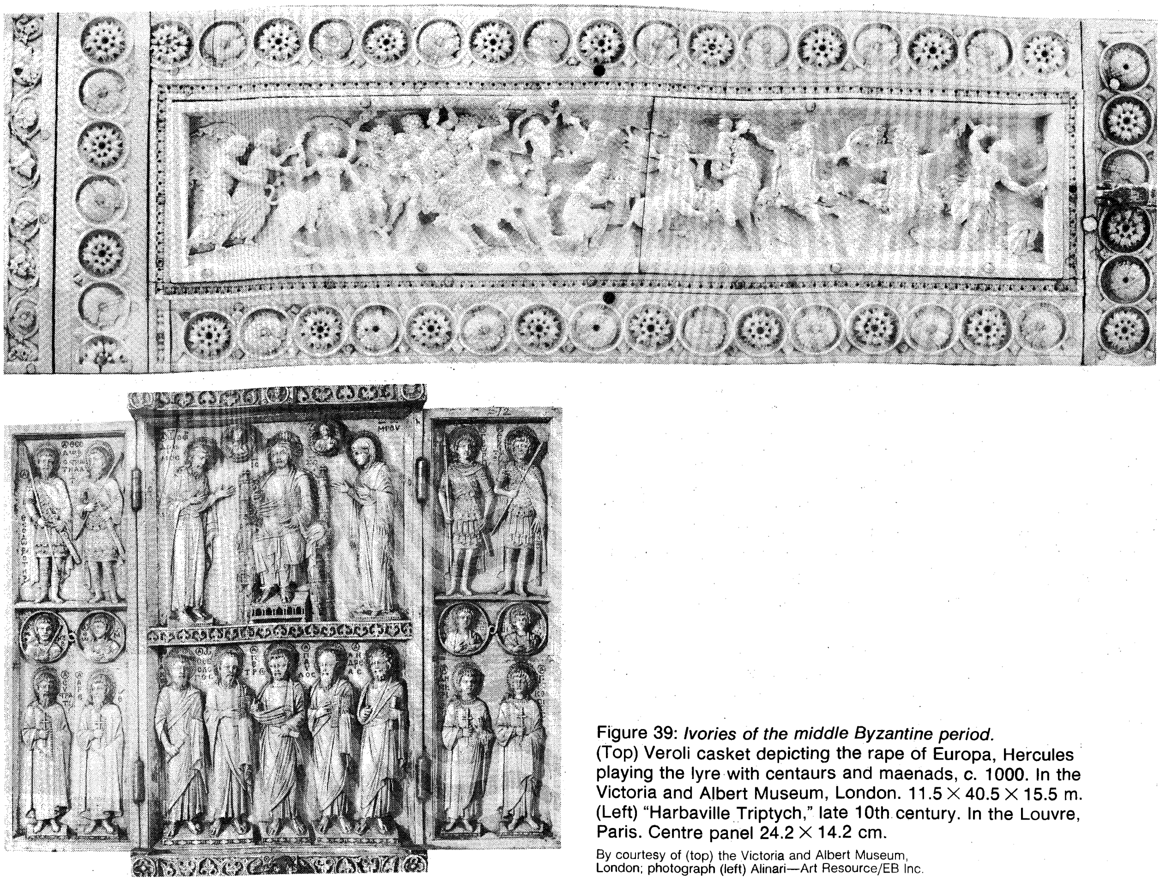


Figure 39: *Ivories of the middle Byzantine period.*
(Top) Veroli casket depicting the rape of Europa, Hercules playing the lyre with centaurs and maenads, c. 1000. In the Victoria and Albert Museum, London. 11.5 × 40.5 × 15.5 m. (Left) "Harbaville Triptych," late 10th century. In the Louvre, Paris. Centre panel 24.2 × 14.2 cm.

By courtesy of (top) the Victoria and Albert Museum, London; photograph (left) Alinari—Art Resource/EB Inc.

Sforzesco at Milan, where the decoration is confined to rosettes at the four corners and a medallion with a Latin inscription at the centre.

Most of the official ivories were probably carved at Constantinople, but it seems likely that others, which were intended for more general use or for the church, may well have been done elsewhere. Rome, Milan, Alexandria, and Antioch in Syria were all important centres, and there has been a good deal of dispute among experts as to where many of the ivories were made. Maximian's throne, the most elaborate of them all, has been assigned to Alexandria, Constantinople, and even to Ravenna itself; and there has been argument as to whether the consular diptychs were carved at Constantinople, Rome, or Alexandria. There is, however, unanimity with regard to certain types. Thus, a number of rather small plaques bearing decorations in a clumsy but expressive style can safely be assigned to Palestine, and probably to Jerusalem; another group, characterized by a similar search for realism but by greater technical proficiency, can perhaps be attributed to Antioch. A leaf in the British Museum, with the Adoration of the Magi above and the Nativity below, illustrates the first type; a composite diptych used as a book cover, now at Ravenna, represents the second. Each of its leaves is made up of five panels, like those of the imperial diptychs, but here Christ occupies the central one, and there are scenes from the Gospels and the Old Testament all around.

Alexandrian school

Work of a more polished type, where classical scenes, single figures, or, less often, events from the Bible are the subjects, has been associated with Alexandria. At one time this city was regarded as the primary centre of production, and numerous ivories of major importance were attributed to it, notably the throne of Maximian. The panels that compose the latter are in various styles and are certainly not all of the same school. Those on the sides, depicting scenes from the life of Joseph, are vivid and expressive, whereas those on the front, showing John the Baptist, prophets, and ornamental scrollwork, are grand and elegant. It is possible that the artist who did the Joseph scenes was trained in Alexandria, but most of the rest of the work is now generally regarded as Constantinopolitan, and it was probably there that the throne was carved, wherever the craftsmen had been trained. Also typical of Constantinople, especially during the rule of Justinian, is a large panel in the British Museum representing the archangel Michael. The treatment of this youthful figure and his drapery is in a style reminiscent of classic Greek art, but this is happily combined with ornate decoration and a hieractic composition.

A few ivories bearing secular scenes may also be assigned to the capital; one of the most important is a diptych in the Hermitage at Leningrad with depictions of animal combats in the circus.

Classification of mid-Byzantine ivories

A fragment of a sceptre in the name of Leo VI (886–912) at the Staatliche Museen Preussischer Kulturbesitz, Berlin, a panel showing the crowning of Constantine VII Porphyrogenitus by Christ (944) in Moscow, and one with the crowning of Romanus II (945) in the Cabinet des Médailles, Paris, can be dated exactly. But in most cases, dates can only be suggested on the basis of style. The ivories have been classified under a number of headings in a monumental survey made by A. Goldschmidt and K. Weitzmann. They term their first group that of Romanus and associate a number of ivories with that showing his crowning, mentioned above; they include triptychs with the deesis on the central panel in the Vatican, the Palazzo Venezia at Rome, and the Louvre, the last known as the "Harbaville Triptych" (Figure 39, left), as well as panels at Dresden, Venice, Vienna, and elsewhere.

Goldschmidt and Weitzmann's second group is built up around an ivory in the church of Sta. Francesca at Cortona, Italy, which bears the name of Nicephorus II Phocas (963–969). It includes among others a fine triptych with the Virgin on the central panel, at Luton Hoo, Bedfordshire, England. The faces are broader and heavier than those on ivories of the Romanus group. Other groups are distinguished not so much on the basis of date as by form or style, such as groups termed the "painterly" and the "framed," while a more obvious group is composed of caskets. The majority of examples are dated to the later 10th or earlier 11th centuries, but manufacture of objects in this group apparently continued at least until the early 12th century, the later ones being either more linear in style, like a panel with the Baptist and four Apostles in the Victoria and Albert Museum, London, or the figures being very much elongated, as in a St. John at Liverpool. High relief and deep undercutting were apparently in special favour early in the 11th century.

Though the caskets were no doubt often carved by the same people who carved the plaques, they constitute an independent group not only because of their form but also because they are nearly all adorned with secular motifs that have been drawn from Classical literature. The panels bearing the scenes are framed in bands adorned with rosettes or sometimes human heads in profile; because of this, the caskets are often termed rosette caskets. The most exquisite in execution, if also mannered in style, is one in the Victoria and Albert Museum known as the Veroli casket (Figure 39, top). A few caskets of different type are also known; one at Florence has the rosette borders, but they frame panels bearing Christ, the Virgin, and saints; one at Troyes, France, has no rosette borders, while its side panels show horsemen of Persian type and, at the ends, phoenixes that are distinctly Chinese. During the later part of the 12th century, soapstone plaques became more common than ivories, probably for economic reasons, but they bore low-relief decorations in a very similar style. (D.T.R.)

Georgia. A distinct Georgian sculptural tradition did not emerge until the advent of Christianity, which stimulated a demand for a large number of carved stone reliefs. The earliest of these were based on Early Christian models. In the 8th and 9th centuries the high-relief figures of Early Christian art gave way to figures rendered in wholly linear fashion. In the 10th and 11th centuries the reliefs became gradually more plastic and expressive until they were again freed, to a considerable degree, from the background. At the same time there was an increasing interest in the disposition of figures in a harmonious design. By the 12th century, however, sculptors were beginning to look more to ornamentation than to figural representation. Repetition of themes characterized most of Georgian sculpture in subsequent centuries. Sculpture of all periods was always smaller than life-size.

Linear and plastic treatment of reliefs

Armenia. The stone construction of Armenian churches lent itself to carved decorations, and architectural sculpture was more extensively used in Armenia than in any other country of the Middle East, except Georgia. The reliefs of the 4th-century hypogeum (a subterranean structure hewn out of rock) at Aghts along with those on numerous funerary stelae (upright slabs of inscribed stone) antedating the Arab conquest exemplify the early stages of stone sculpture. Beginning with the 6th century, and perhaps even earlier, floral and geometric motifs as well as figure representations were carved around the windows of the churches, between the arches of the blind arcades, and on the lintels and the lunettes over the doors. Decorative ornaments became increasingly intricate during the later periods.

The outstanding example in Armenian art of the use of architectural sculpture is the Church of the Holy Cross, built in the early 10th century on the island of Aghthamar in Lake Van (Figure 40); this is the earliest medieval example, either in the East or in the West, of a stone building entirely covered with relief sculpture. Around the dome and on the four facades may be seen a variety of animals, vine and other floral scrolls, and large figures of saints and scenes from the Old Testament. A portrait of King Gagik I Artsruni, offering to Christ a model of the church he had erected, appears on the west facade. Such donor portraits, sometimes carved in the round as at Ani, were one of the characteristic features of the decoration of Armenian churches.

Donor portraits

Coptic Egypt. Strictly speaking, the adjective Coptic, when it is applied to art, should be confined to the Christian art of Egypt from the time when the Christian faith may be recognized as the established religion of the country among both the Greek-speaking and Egyptian-speaking elements of the population. In this sense Coptic art is essentially that reflected in the stone reliefs, wood carvings, and wall paintings of the monasteries of Egypt, the earliest foundations of which date from the 4th and 5th centuries AD. It is, however, common practice to include within Coptic art all forms of artistic expression that, like the so-called Coptic textiles, need have no religious intent or purpose. The term has also been further extended to denote stylistic characteristics that can be traced back to the 2nd and 3rd centuries AD and perhaps earlier.

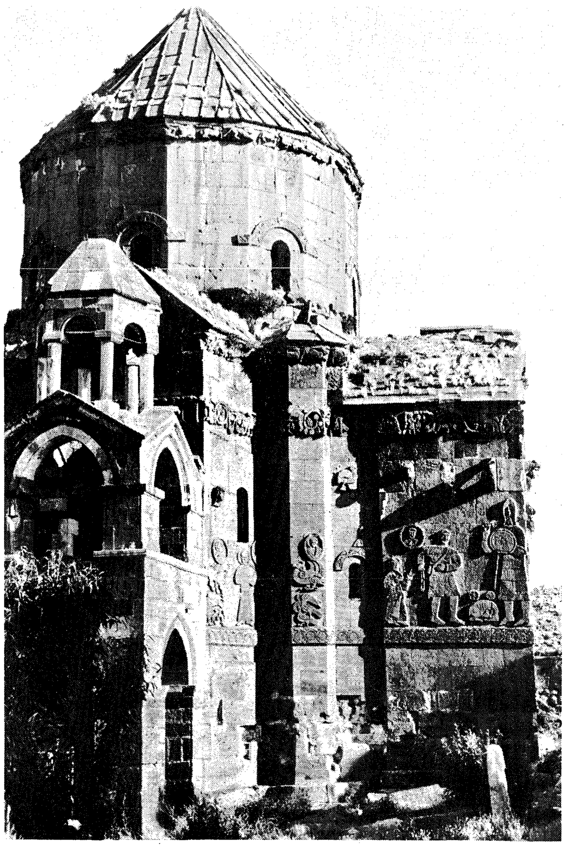


Figure 40: Church of the Holy Cross, Aghthamar, Lake Van, Armenia (now Turkey), early 10th century.
John Donat

A specifically Christian art was slow in developing: when it did emerge, it was not the product of a school of Christian artists inventing new forms of expression. It continued the style current in the country, evolving from the late antique art of Egypt, in which themes derived from Hellenistic and Roman art may or may not have been given new allegorical significance. There is little direct legacy from the art of pharaonic Egypt either in the style of execution or in the choice of decorative themes. The most obvious survival in Christian iconography is the peculiar looped form of cross derived from the ancient Egyptian writing of the word for life (*ankh*). Less convincing is the connection postulated between the concept of *Maria lactans* (representations of the Virgin nursing her child) and bronze and terra-cotta statues of the ancient Egyptian goddess Isis suckling the infant sun god Horus or between representations of saints on horseback and some late figures of the adult Horus in an identical pose.

Origins in ancient Egyptian and Greco-Roman art

The extent to which Egypt may have exerted a major creative influence on Christian art is uncertain in the absence of material remains of the Christian period from Alexandria, the great metropolis of Egypt from the time of the Ptolemies and a city that played an important and, at times, decisive role in the intellectual life of the early church. A series of Christian ivory carvings, of unrecorded provenance, is frequently referred to as Alexandrian on stylistic considerations and adduced as proof of a continuing artistic skill in the Hellenistic tradition.

Objects found in the hinterland depart from the Classical canons of proportion and mode of representation. Political and economic conditions in Egypt from the time of its incorporation in the Roman and, later, Byzantine empires doubtless account for much of the provincial appearance of Egyptian and Coptic art and the emergence of a freer, more popular folk style. Lack of the kind and degree of patronage that had been given by the pharaohs, Ptolemies, and, to some extent, Roman emperors to the old religion of Egypt meant an impoverishment of schools of skilled craftsmen, avoidance of costlier materials, and a decline in the high standard of finish. Particularly noticeable is the absence of carving in the round, of work of monumental scale, and of the use of the harder ornamental stones that had been characteristic of pharaonic art.

Characteristic Coptic stylistic features are to be observed in tombstones from the Delta site of Terenuthis. These depict the dead man frontally posed beneath a gabled pediment of mixed architectural style, hands extended at right angles from the body and bent upward from the elbow in the orans (praying) position, a pose that appeared frequently in the earliest Christian art in Rome. There is no firm evidence, however, that the community was Christian. Similarly, the series of architectural elements carved in relief from Oxyrhynchus and Heracleopolis may not all be from Christian buildings. The earlier material from Heracleopolis, dating probably from the 4th century, is notable for its figure subjects drawn from classical mythology, carved in a deep relief that leaves them almost freestanding, producing an effective play of light and shade. As such reliefs were painted, the absence of fine detail in the carving was less noticeable.

Much of the material available for a study of Coptic sculpture has not been found in context, and, in the absence of assured information concerning its provenance and of circumstantial evidence for dating (even in the cases of pieces from known sites), it is impossible to provide a detailed account of the development of Coptic sculpture. In general, the figures are stiff in pose and movement; there is a tendency for the carving to become flat, and there is little in the way of narrative scenes drawn from biblical stories. The most successful carvings are probably the impressive variety of decorated capitals, particularly from the monasteries of Apa Jeremias at Ṣaqqārah and of Apa Apollo at Bāwīṭ. Among them are basket-shaped examples decorated with plaitwork, vine and acanthus leaves, and animal heads. The form imitates a style introduced into Constantinople by the emperor Justinian I, and it is clear that, in the hinterland of Egypt, there was during the 6th century certain artistic influence on Coptic art from Byzantium, despite religious and political differences. Contemporary Byzantine influence seems to have been at work on other architectural elements at Bāwīṭ, as, for example, in the finely carved limestone pilaster depicting, on one side, a geometric and floral pattern surmounted by a saint and, on the other, vine scrolls and birds below an archangel. (A.F.Sh.)

Carved capitals

WESTERN CHRISTIAN

With the dissolution of the Roman Empire in the West, cultural hegemony passed to the Eastern Empire, but older traditions remained in western Europe and intermingled with several invaders—Germanic tribes arriving from the north and Christians arriving from Constantinople as well as from Rome. The Merovingian art of the Franks, which was culturally predominant throughout Europe in the 6th century, survives principally in grave relics, such as jewelry, hollowware, and the like.

In Italy the Lombards, who invaded the country in 568, propagated Germanic art, but there is a strong Mediterranean influence in the sculpture—stone plaques for choir screens, altars and altar canopies, sarcophagi, and details of architecture, for example; the abstract decorations, many of them interlaced motifs, were to be blended with more and more Byzantine elements (Figure 41). The creatures and vegetation become almost impossible to recognize—they aspire, as it were, to be ornamental stone writing rather than representation. Similar ornaments were also applied in stucco; for example, in S. Salvatore at Brescia and especially in the famous Tempietto at Cividale del Friuli (both 8th century). At Cividale del Friuli, standing figures of saints have been incorporated in decoration in which the Byzantine influence is obvious.

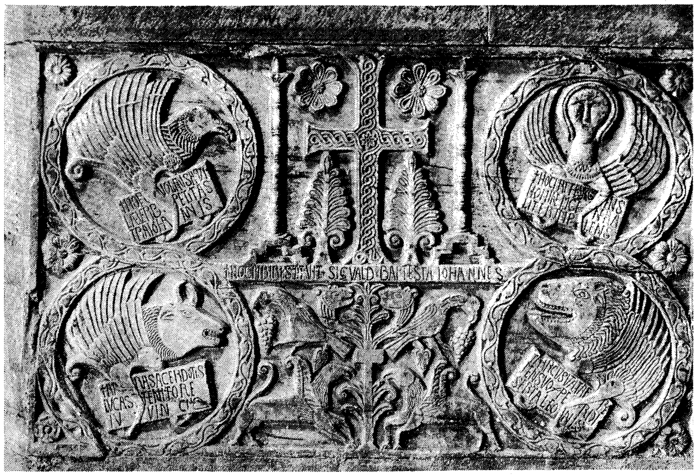


Figure 41: Marble relief inscribed by the Patriarch Sigwald, part of the canopy over the baptismal font in the cathedral of Cividale, Italy, 762–776. 0.91 m × 1.52 m.

Bildarchiv foto Marburg—Art Resource/EB Inc.

In Ireland, monumental crosses represented the Celtic Christian tradition, and similar Anglo-Saxon crosses may be found in England (Figure 42). The abstracted decoration recalls the relief style in Italy, but here the surface is not a flat plane but is packed with round, knoblike projections that create a plastic rather than a glyphic effect.

Carolingian and Ottonian periods. The cultural revival of the Carolingian period (768 to the late 9th century), stimulated by the *academia palatina* at Charlemagne's court, is the first phase of the pre-Romanesque culture, a phase in which late Classical and Byzantine elements amalgamated with ornamental designs brought from the East by the Germanic tribes. The German Ottonian and early Salian emperors (950–1050), who succeeded the Carolingians as rulers of the Holy Roman Empire, assumed initially the Carolingian artistic heritage, although Ottonian art later evolved into a distinct style.

By courtesy of the Irish Tourist Board

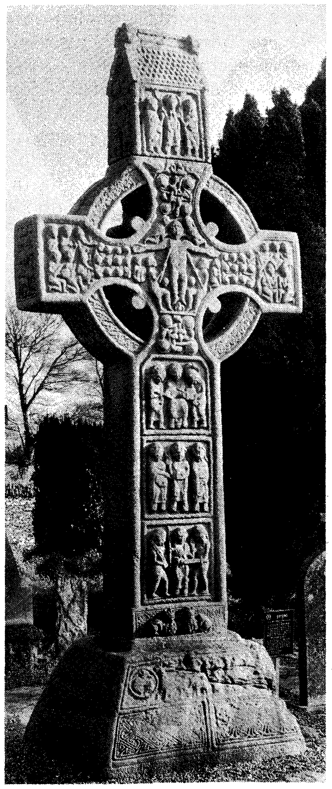


Figure 42: "Cross of Muirdach," Clonmacnoise, 923. Height 5.38 m.

Little Carolingian sculpture has survived, but in Ottonian days the sculpting of freestanding statues was taken up again, although the earliest specimens, serving as they did as reliquaries, were still closely related to the silversmith's and goldsmith's art; for example, the famous statue of "Sainte-Foy" at Conques (France) and the "Golden Madonna" at Essen. The wooden "Gero Crucifix" (about 73.6 inches [187 centimetres] high; cathedral of Cologne; Figure 43), which was carved before 986, already reveals a certain realism in the representation of the shape of the body, in contrast to the contemporary crucifix of Gerresheim (before 1000). The so-called Bernward Crucifix at Ringelheim (Germany) is between the two. The reliefs on the wooden doors of Sankt Maria im Kapitol at Cologne display an affinity with the mid-11th-century Romanesque ivories of the Meuse district. The Carolingian bronze doors in Aachen were imitated at Mainz, where Bishop Willigis had similar portal wings made for his cathedral. He was far surpassed, however, by Bernward at Hildesheim, who had the still extant door wings of the cathedral (1015) decorated with typological images in parallel, scenes from the Old and the New Testament; in theme, the images go back to early Christian examples Bernward had seen in Italy, but the force of the gestures and the use of unadorned surface as dramatic interval in the episode of Adam and Eve reproached by the Lord has no precedent in the history of art (Figure 44). The influence of Classical art manifests itself clearly in the so-called Christ's Column (12.8 feet [3.9 metres] high; *c.* 1020; St. Michael's, Hildesheim), which, with its figures spiralling around the shaft, reminds one of the triumphal columns of Trajan and Marcus Aurelius. Originally, it was crowned by a cross. As belonging to the art associated with Bernward, one must also reckon the seven-branched candlestick in the Minster of Essen (90.6 inches [230 centimetres] high; before 1011) and the bronze crucifix at Essen-Werden (42.5 inches [108 centimetres] high; *c.* 1060), a late product of the same school.

"Christ's Column"

Romanesque. The term Romanesque—coined in 1818—denotes in art the medieval synthesis of the widespread Roman architectural and artistic heritage and various regional influences, such as Teutonic, Scandinavian, Byzantine, and Muslim. Although derived primarily from the remains of a highly centralized imperial culture, the Romanesque flowered during a period of fragmented and unstable governments. It was the medieval monasteries, virtual islands of civilization scattered about the continent, that provided the impetus—and the patronage—for a major cultural revival.

The bronze "Christ's Column" is a modest prophecy of the monumental spirit that would distinguish the sculptural decoration of the new monastic buildings rising in much of western Europe. Developed in the abbey doorways and on the pillars and capitals of cloisters, where the sculptor had to learn anew the technique of stone carving and of rendering the human figure, this spirit gradually grew stronger.

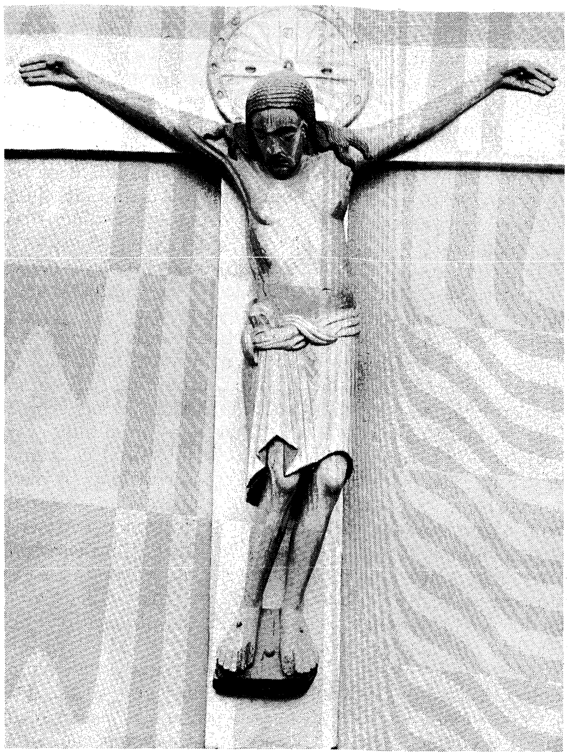


Figure 43: The "Gero Crucifix," carved oak corpus (with contemporary nimbus and stem), before 986. In the cathedral of Cologne. Height 187 cm.

Bildarchiv foto Marburg—Art Resource/EB Inc.

During the 11th century more and more churches were constructed in the Romanesque style, the massive forms of which are another indication of this sculptural instinct. Romanesque sculpture culminated in France in the great semicircular relief compositions over church portals, called tympanums. The example at Moissac (*c.* 1120–30), which represents the Apocalyptic vision with the 24 elders, is a particularly brilliant demonstration of how devices of style can so transform the objects of nature that they seem entirely purged of terrestriality (Figure 45). All the forms are suspended in a predominating plane that denies physical space. Differences in scale are masterfully exploited: the tiny figures of the elders are a foil to the looming image of Christ in the centre. With great consistency, every detail has been subjected to a process of stylization that produces rhythmic patterns in the drapery, hair, and feathers. The central figure is so flattened as to appear disembodied, while the two towering angels have been so attenuated that their bodies have lost all mass.

The astonishing variety that master sculptors such as Gislebertus, Benedetto Antelami, and Nicola Pisano achieved within the confining principles of Romanesque

Bildarchiv foto Marburg—Art Resource/EB Inc.

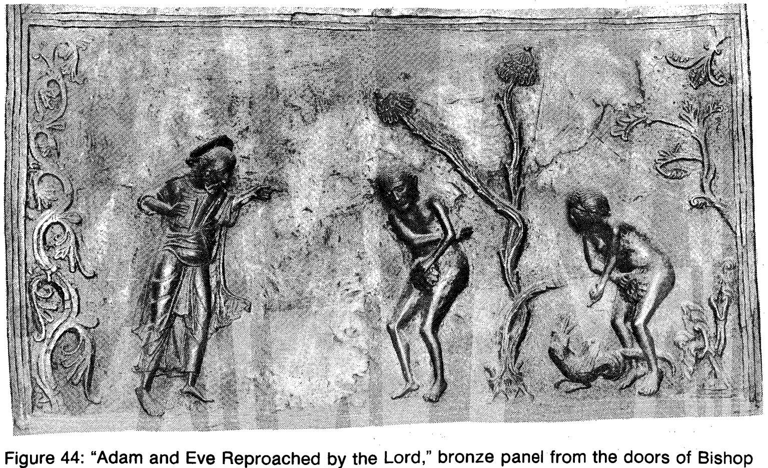


Figure 44: "Adam and Eve Reproached by the Lord," bronze panel from the doors of Bishop Bernward at the cathedral, Hildesheim, West Germany, 1015. Panel 58.6 cm × 196 cm.

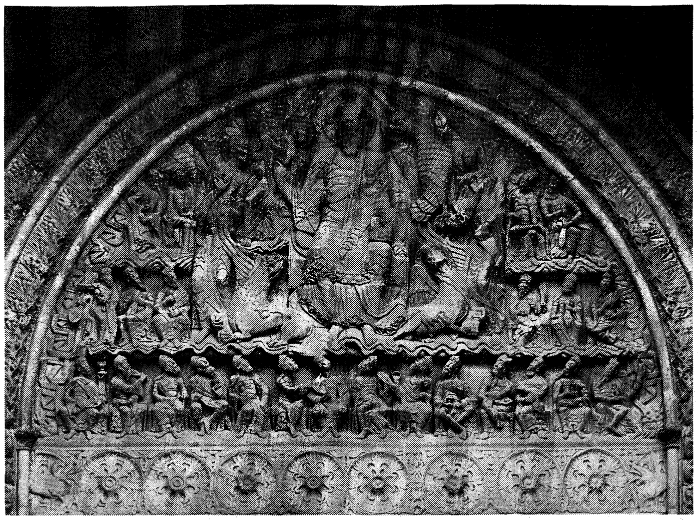


Figure 45: "Christ of the Apocalypse, with the 24 Elders," tympanum of the south portal of the abbey church of Saint-Pierre at Moissac, France, *c.* 1120–30.

Yan

style can be illustrated, on the one hand, by the tympanums of Burgundy, such as the spectral "Last Judgment" at Autun (Figure 46) or the "Pentecost" at Vézelay, and, on the other, by the less visionary sculpture of Provence, such as that of Saint-Trophime in Arles (Figure 47) or of the church in Saint-Gilles, which retain many of the forms and characteristics of Classical antiquity.

Another sculptural form that reappeared in Europe during the latter part of the Romanesque period was sepulchral sculpture, in which a sculptured figure of the deceased was cut or molded on top of a sarcophagus or on the sepulchral slab set into the floor of an abbey or cloister. (J.J.M.T./Ed.)

Gothic

Origin of the term Gothic

The difficulty with many anatomies of Gothic art is that they become involved in attributing a meaning to Gothic that it is incapable of sustaining. It is not, for one thing, a medieval word; instead, it is an invention of the 16th century attributed, as it were, posthumously, by historians after the Gothic style had been trampled into virtual insensibility by the Italian Renaissance. The word refers to the Teutonic tribes who were thought to have destroyed Classical Roman art and were thus considered barbarians. But nobody in the 13th century thought of himself as Gothic. The fact is that the literature of art criticism is virtually nonexistent in the Middle Ages. Certainly people talked about art, patrons valued it, connoisseurs appraised it. But the terms in which this was done must now, for the most part, be a matter of speculation or imagination. There was not necessarily anything mysterious about this. It is common to suppose that medieval discussions on art were infused with a degree of spirituality. This is probably mistaken. There is, for instance, little that is spiritual about financing the building of a gigantic cathedral. It is certain that clergymen preached sermons about art, giving it a spiritual and symbolic interpretation. It is also true that, since a large proportion of art served a religious function, artists were, in some sense, "servants of God." But they were also the servants of far more worldly considerations, such as earning a living or achieving a reputation, and these should never be discounted in any imaginative re-creation of the medieval artist's existence.

Bildarchiv foto Marburg—Art Resource/EB Inc.

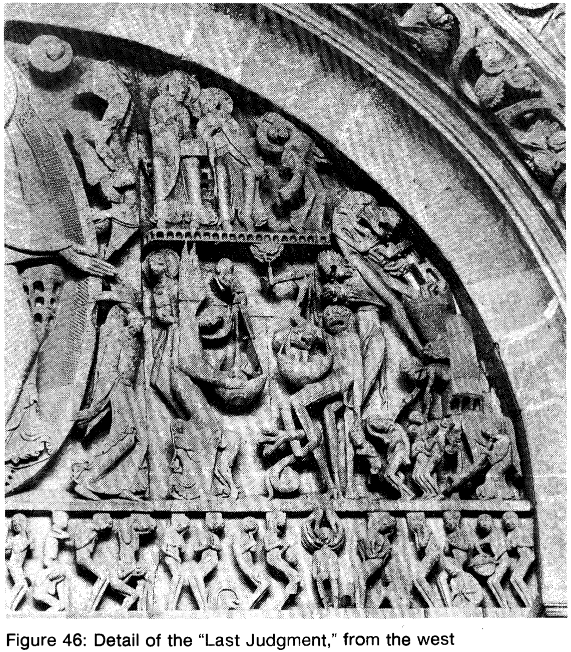


Figure 46: Detail of the "Last Judgment," from the west tympanum of the cathedral of Saint-Lazare, Autun, France, carved by Gislebertus before 1135.

Giraudon—Art Resource

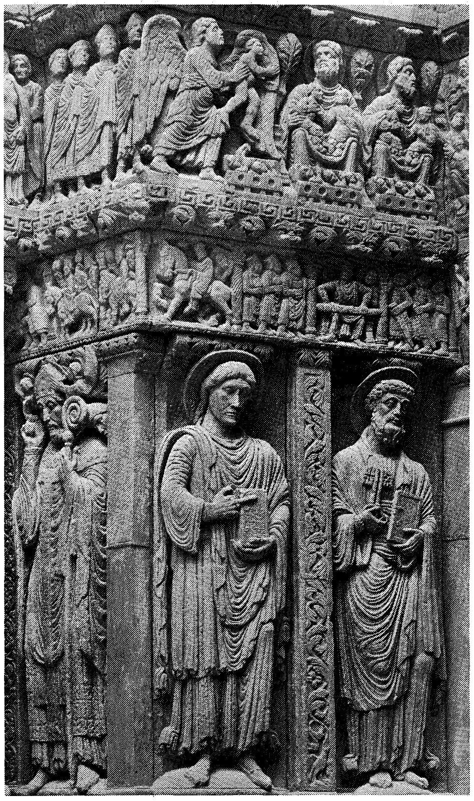


Figure 47: Detail from the main portal of the church of Saint-Trophime, Arles, France, 12th century.

EARLY GOTHIC

Portal sculpture

Throughout this period, as in the Romanesque period, the best sculptors were extensively employed on architectural decoration. The most important agglomerations of figure work to survive are on portals, and, in this, once again, the church of Saint-Denis assumes great significance. The western portals (built 1137–40), part of a total facade design, combined features that remained common throughout the Gothic period: a carved tympanum (the space within an arch and above a lintel or a subordinate arch); carved surrounding figures set in the voussoirs, or wedge-shaped pieces, of the arch; and more carved figures attached to the sides of the portal. As it survives, Saint-Denis is disappointing; the side figures have been destroyed and the remainder heavily restored. The general effect is now more easily appreciated on the west front of Chartres cathedral.

If one compares the portals here (*c.* 1140–50; Figure 48, left) with those of early 13th-century Reims, one can see that the general direction of the changes in this early

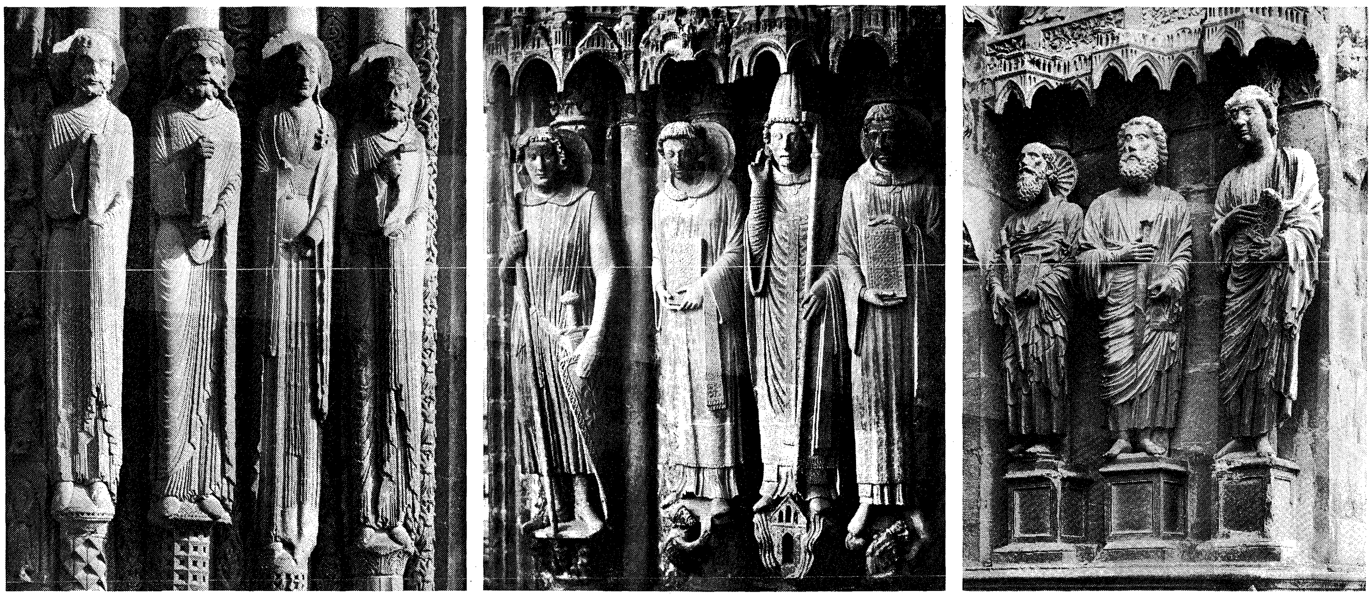


Figure 48: *French early Gothic architectural sculpture.*
(Left) Figures from the Old Testament, centre portal of the west front of Chartres cathedral, *c.* 1140–50. (Centre) Saints, south transept, Chartres cathedral, *c.* 1210–20. (Right) Apostles of the Judgment Portal, north transept, Reims cathedral, *c.* 1225.
(Left) Madame Simone Roubier, Paris, (centre) Giraudon—Art Resource/EB Inc., (right) Archives Photographiques

period of Gothic sculpture was toward increased realism. The movement toward realism is not manifest in a continuous evolution, however, but in a series of stylistic fashions, each starting from different artistic premises and achieving sometimes a greater degree of realism but sometimes merely a different sort of realism. The first of these fashions can be seen in the sculpture on the west front of Chartres. That the Christ and the Apostle figures are in some sense more human than the Romanesque apparitions at Vēzelay and Autun (*c.* 1130) need hardly be argued. That the figures, with their stylized gestures and minutely pleated garments, are at all "real" is doubtful. That their forms are closely locked to the architectural composition is clear. The features of the Chartres sculpture had a wide distribution; they are found, for example, at Angers, Le Mans, Bourges, and Senlis cathedrals. There are stylistic connections with Burgundy and also with Provence. The fashion lasted from *c.* 1140 to 1180.

The centre of development for the second style lay in the region of the Meuse. The activity of one of the chief artists, a goldsmith called Nicholas of Verdun, extends at least from the so-called Klosterneuburg altar (1181) into the early years of the 13th century. His style is characterized by graceful, curving figures and soft, looping drapery worked in a series of ridges and troughs. From these troughs is derived the commonly used German term for this style—*Muldenstil.* This drapery convention is essentially a Greek invention of the 4th century BC. It seems likely that Nicholas seized the whole figure style as a tool to be used in the general exploration of new forms of realism. It remained extremely popular well into the 13th century. A rather restrained version of the style decorated the main portals of the transepts (the transversal part of a cruciform church set between the nave and the apse or choir) of Chartres (*c.* 1200–10; Figure 48, centre). It is also found in the earliest sculpture (*c.* 1212–25) of Reims cathedral (Figure 48, right) and in the drawings of the *Sketchbook* of Villard de Honnecourt (*c.* 1220).

In the opening years of the 13th century yet another type of realism emerged. It seems to have originated at Notre-Dame, Paris (*c.* 1200), and to have been based on Byzantine prototypes, probably of the 10th century. The looping drapery and curving figures were abandoned; instead, the figures have a square, upright appearance and are extremely restrained in their gestures. Figures in this style are found at Reims, but the major monument is the west front (*c.* 1220–30) of Amiens cathedral (Figure 49, left).

The Joseph Master

Once again, the style changed. On the west front of Reims worked a man called after his most famous figure, the Joseph Master (Figure 49, right). Working in a style that probably originated in Paris *c.* 1230, he ignored the restraint of Amiens and the drapery convolutions of the *Muldenstil* and produced (*c.* 1240) figures possessing many of the characteristics retained by sculpture for the next 150 years: dainty poses and faces and rather thick drapery hanging in long V-shaped folds that envelop and mask the figure.

Another aspect of this quest for realism was the spasmodic fashion throughout the 13th century for realistic architectural foliage decoration. This resulted in some astonishingly good botanical studies—at Reims cathedral, for example.

The effects elsewhere in Europe of this intense period of French experiment were as piecemeal and disjointed as the effects of the architectural changes. In England, the concept of the Great Portal, with its carved tympanum, voussoirs, and side figures, was virtually ignored. The remains of a portal the style of which may be connected with Sens cathedral survive from St. Mary's Abbey, York, England (*c.* 1210). Rochester cathedral (*c.* 1150) has carved side figures, and Lincoln cathedral (*c.* 1140) once had them. The major displays of English early Gothic sculpture, however, took quite a different form. The chief surviving monument is the west front of Wells cathedral (*c.* 1225–40), where the sculpture, while comparing reasonably well in style with near-contemporary French developments, is spread across the upper facade and hardly related at all to the portal.

In Germany, the story is similar. On the border between France and Germany stands Strasbourg, the cathedral of which contains on its south front some of the finest sculpture of the period (*c.* 1230). A very fine and delicate version of the *Muldenstil,* it comes reasonably close to the best transept sculpture of Chartres. But it differs in two important respects. Predictably, its architectural framework is entirely different; and it has the slightly shrill emotional character, common in German art, that represents an effort to involve and move the spectator. Shrill emotionalism is again found at Magdeburg cathedral in a series of "Wise and Foolish Virgins" (*c.* 1245) left over from some abandoned sculptural scheme. Influenced by Reims rather than Chartres, the sculpture of Bamberg cathedral (*c.* 1230–35; Figure 50, left) is a heavier version of the *Muldenstil* than that at Strasbourg. But of all this German work, by far the most interesting complex is in the west choir (*c.* 1250) of Naumburg cathedral (Figure 50, right). Here, the desire for dramatic tension is exploited to good effect, since the figures—a series of lay founders in

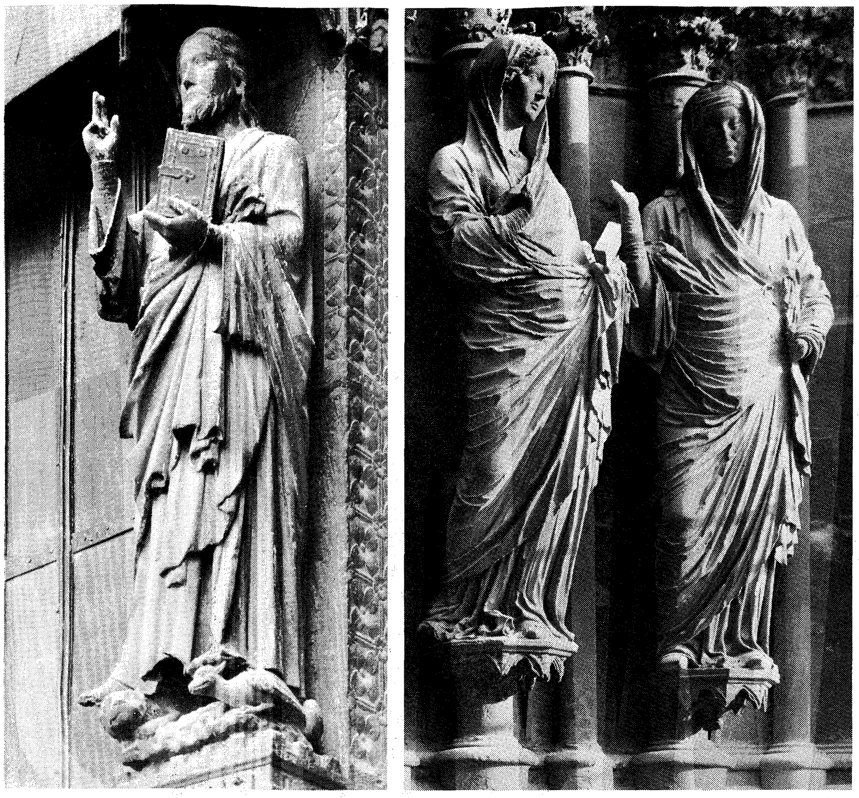


Figure 49: *Styles of realism in portal sculpture in France.*
(Left) Statue of Christ ("Le Beau Dieu"), centre portal of the west facade, Amiens cathedral, *c.* 1220–30. (Right) Visitation, detail of the Virgin's Portal, west facade, Reims cathedral, 1225–45.
Jean Roubier

contemporary costume—are given a realistic place in the architecture, alongside a triforium gallery. Naumburg also has a notable amount of extremely realistic foliage carving.

It is hard to say what a French mason would have made of this English and German work. With the major Spanish work of the period, however, he would have felt instantly at home. Burgos cathedral has a portal (1230s) that is very close to the general style of Amiens, and its layout is also, by French standards, reasonably conventional.

HIGH GOTHIC

Late sculptural developments of the early Gothic period were of great importance for the High Gothic period. The Joseph Master at Reims and the Master of the Vierge Dorée at Amiens both adopted a drapery style that, in various forms, became extremely common for the next century or more; both introduced into their figures a sort of mannered daintiness that became popular. These features appear in an exaggerated form in some of the sculpture for the Sainte-Chapelle, Paris.

On the whole, this period saw the decline of architectural sculpture. Given the emphasis placed on geometric patterning by the Rayonnant style, perhaps this is not surprising. A few portals, such as those on the west front of Bourges cathedral, were completed, but they have a very limited interest. The field of sculpture that expanded with great rapidity was the more private one, represented by tombs and other monuments.

For this, the family feeling of Louis IX was partly responsible. By making sure that both his remote ancestors and his next of kin got a decent burial—or reburial—he was responsible for an impressive series of monuments (the remnants of which are now chiefly in Saint-Denis) executed mainly in the years following 1260. Although earlier examples and precedents may be found, Louis IX had a large share in popularizing the idea of the dynastic mausoleum, and many other important people followed suit.

The monuments executed for St. Louis have come down in such a battered state (almost entirely as a result of the destruction wrought during the French Revolution) that it is difficult to generalize about them. One can say, however, that Louis's masons popularized two important ideas. One was the tomb chest decorated with small figures in niches—figures generally known as weepers, since they often represented members of the family who might be presumed to be in mourning. Later, in the early 14th century, the first representations appear of the heavily cloaked and cowled professional mourners who were normally employed to follow the coffin in a funeral procession. The second innovation introduced by Louis's masons lay in the emphasis given to the effigy. Around 1260 the first attempts were made to endow the effigy with a particular character. This may not have involved portraiture (it is obviously hard to be sure), but it did involve a study of different types of physiognomy, just as the botanical carving of the early Gothic period had involved a study of different kinds of leaves.

A somewhat similar story may be told of English sculpture during this period. The architectural carving found at Westminster Abbey (mainly of the 1250s) has much of the daintiness of contemporary French work, although the drapery is still more like that of the early Chartres or Wells sculpture than that of the Joseph Master. The baggy fold forms of the Joseph Master rarely appear in England before the sculptured angels of the Lincoln Angel Choir (after 1256).

Architectural sculpture in England probably remained more interesting than the continental equivalent because first-rate masons continued to work in this field in England until the end of the 13th century. Hence, around 1295 one can still find a work such as the botanical carving of Southwell Chapter House. Even in the 14th century, there are such architectural and sculptural curiosities as the west front of Exeter cathedral. Sculptural interest, however, in buildings such as Gloucester Cathedral Choir (begun soon after 1330), where the effect depends on traceried panels, is virtually nonexistent; and the "leaves of Southwell" were succeeded almost at once by an extremely dull form of foliage commonly known as "bubbleleaf,"

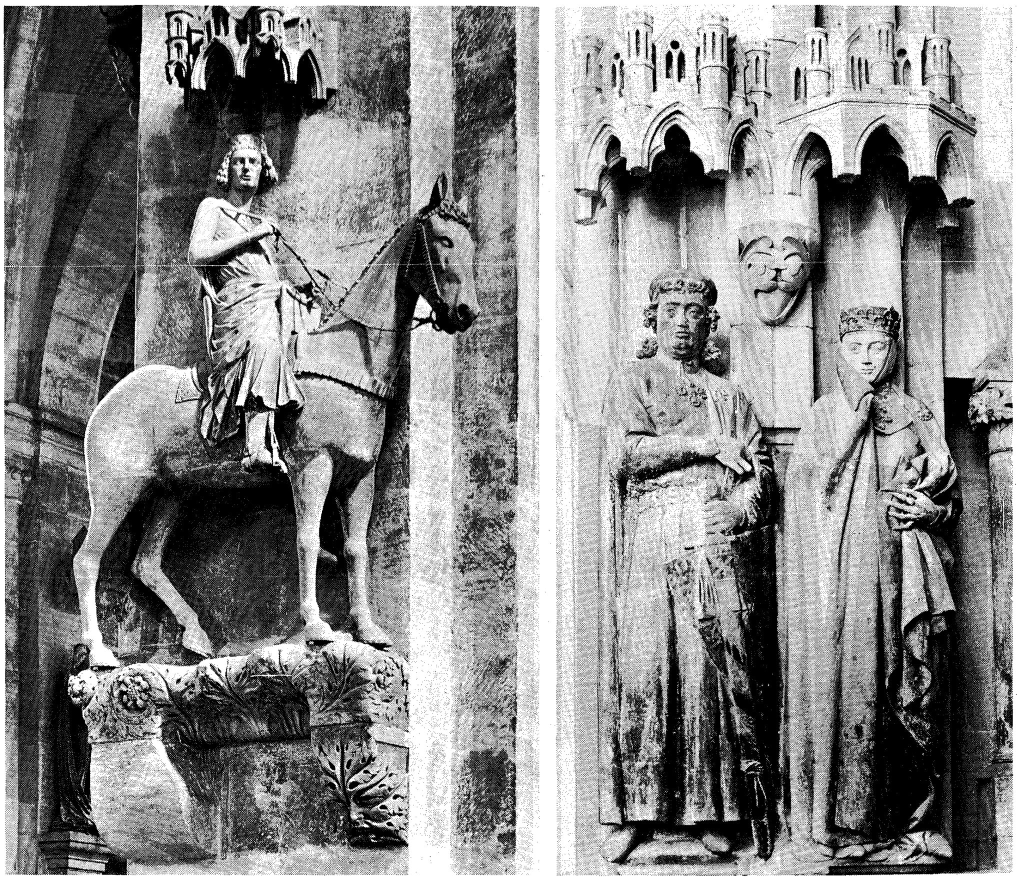


Figure 50: *German Gothic sculpture.*
(Left) "Bamberg Horseman," possibly a king or emperor, Bamberg cathedral, Germany, *c.* 1230–35. (Right) "Ekkehard and Uta," Naumburg cathedral, Germany, *c.* 1250.
(Left) Foto Marburg—Art Resource/EB Inc., (right) H. Roger-Viollet

which remained more or less standard for the 14th and 15th centuries.

Monuments of Westminster Abbey

As in France, much of the virtuosity in carving went into private tombs and monuments. The best surviving medieval mausoleum is Westminster Abbey, where a large number of monuments in a variety of mediums (especially purbeck, bronze, alabaster, and freestone) is further enhanced by some of the floors and tombs executed by Italian mosaic workers introduced by Henry III. Especially well preserved is the tomb of Edmund Crouchback, earl of Lancaster (died 1296), which has a splendid canopy and retains some of its original colouring.

As in the early Gothic period, the west of England produced some highly original work that appears to stand outside the normal canon of European development. The earliest monument in this series is the tomb of Edward II (*c.* 1330–35), which is notable for one of the most elaborate surviving medieval canopies. It is preceded stylistically by the wooden canopies of stalls in Exeter cathedral and thus is likely to be a translation into stone of carpenters' work. It was followed by a series of monuments, in Tewkesbury and elsewhere, extending into the 15th century and then dying out.

German High Gothic sculpture is represented by some rather dainty, elegant figures, enveloped in curving and bulky drapery, around the choir of Cologne cathedral (consecrated in 1322). There is also some impressive figure sculpture on the west front of Strasbourg cathedral (begun after 1277). It is strongly influenced by the Joseph Master of Reims but also by the earlier Gothic sculpture of Strasbourg itself. Although it varies in style, much of it is far more expressive than the related French work. The sculptors seem to have been trying to capture an emotive mood.

Spanish High Gothic architectural sculpture is probably less interesting but, by French standards, is more conventional than the German. Major portals exist at León (13th century) and Toledo (14th century) cathedrals, which conform more or less to the rather elegant and mannered French style. Spain also possesses a considerable number of interesting tombs from this period.

Italian Gothic. The figurative arts in Italy during the period 1250–1350 have a strong line of development. The most important 13th-century sculptors were Nicola Pisano (1210/20–1278/84) and his son Giovanni (*c.* 1245–after 1314). Both worked mainly in Tuscany, and both executed pulpits that rank as their major completed works. Nicola's style, as seen in the Pisa Baptistery (1259–60) and Siena cathedral (1265–68) pulpits, was heavily influenced by Classical sculpture—especially by the facial types and the methods of constructing pictorial relief compositions (Figure 51). Nevertheless, his reliefs resemble 13th-century sculpture, particularly in the handling of the drapery. Moreover, in moving from Pisa to Siena, one is conscious of a transition from a strongly antique style to something much closer to northern Gothic sculpture. Nicola's use of Classical ideas was in some way linked with a search for a more realistic style. It forms, in this respect, an interesting parallel to the *Muldenstil* work of Nicholas of Verdun, who was active in the Mosan region from the late 12th to the early 13th century.

The sculptural style of Giovanni does not develop from that of his father. His pulpit in S. Andrea Pistoia (completed 1301), for instance, is technically less detailed and refined but emotionally much more dramatic. While it is possible that the emotionalism of his work was inspired by Hellenistic sculpture, it is also possible that Giovanni had travelled in and been influenced by the north, especially Germany.

Giovanni's first major independent work was a facade for Siena cathedral (*c.* 1285–95). The lower half alone was completed, and it survives in the present building along with a large proportion of Giovanni's imposing figure sculpture. It is quite dissimilar to French facades,

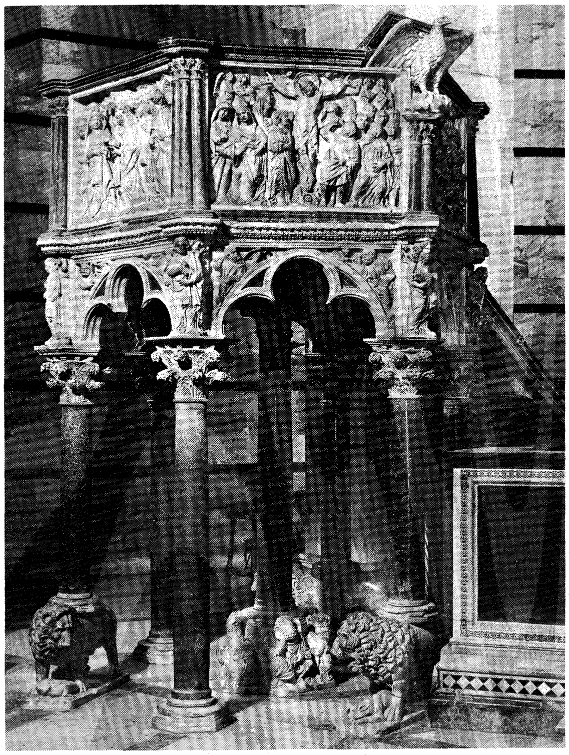


Figure 51: Marble pulpit by Nicola Pisano, 1259–60. In the Pisa Baptistery.

Alinari—Art Resource/EB Inc.

although the placing of the main sculpture above the portals finds an elusive parallel in Wells cathedral, in England (*c.* 1225–40).

Influence of the Pisano workshop

The fame of Nicola's workshop spread to other areas of Italy. For S. Domenico in Bologna, his workshop made a shrine for the body of St. Dominic (1260s). And in Milan, a shrine for the body of St. Peter Martyr was made for S. Eustorgio (1335–39) by Giovanni di Balduccio in a style derived from the Pisano workshop. The most famous Pisano "exports," however, were Arnolfo di Cambio, who worked for the papal court in Rome *c.* 1275–1300, and Tino di Camaino, who worked at the Neapolitan court *c.* 1323–37.

Arnolfo's style is the more difficult to understand. Although he worked alongside Giovanni Pisano during the 1260s, their works have little in common. Arnolfo's sculpture is very solid and impassive. He excelled at formal, static compositions, such as were required for church furniture and for tombs. He designed the funerary chapel as well as the tomb of Pope Boniface VIII and like the Pisanos was architect as well as sculptor; indeed, he was the first architect of the new cathedral of Florence (founded 1296).

Tino di Camaino went south after a training in Siena and a successful career in Tuscany. Sometimes his style approaches the elegance and sweetness of northern 14th-century sculpture, but there is generally a residual heaviness, especially in the faces, that reminds one of his origins in the Pisano circle. He was famous as a tomb sculptor, and the largest collection of his monuments is in Naples (much of the sculpture, however, was executed by his workshop). The tombs make an interesting comparison with those of the French and English royal houses. At another mausoleum (of the Scaliger family), at Verona, the figure sculpture is reminiscent of the Pisano style, but the decorative canopy work is more elaborate and closer to northern art.

The workshop of the facade of Orvieto cathedral and the work of the sculptor and architect Andrea Pisano (no relation to Nicola and Giovanni) are less clearly connected with the Pisano tradition. The facade of Orvieto was designed by the Sienese Lorenzo Maitani *c.* 1310. The sculptural decoration is in varying styles, the best of which is an extraordinarily low and delicate relief that gives an almost pictorial quality.

Andrea Pisano is known chiefly through the bronze doors completed for the Baptistery of Florence cathedral during the 1330s. The scenes of the life of St. John the Baptist are set in quatrefoils (a four-lobed foliation), a common High Gothic decorative motif. Within this awkward shape, the episodes are composed with masterly skill. Although nothing certain has been established about the training of Andrea Pisano, his background is likely to have been similar to that of some of the Orvieto sculptors. The main difference is the evident impact of Giotto's painting, which led Andrea to make his figures rather stocky and solid.

Andrea had a son, Nino Pisano, about whom little is known but from whose hand a group of Madonnas survives. They are interesting in that they veer strongly in the direction of daintiness and sweetness and, to this extent, look more northern than almost any other group of Italian sculpture before the early work of Lorenzo Ghiberti.

International Gothic. The plastic arts are harder to understand in this period, because they have been far more frequently the subject of wanton destruction. Enormous quantities, for example, of goldsmiths' work owned by the French royal family have almost entirely vanished. A few of the remaining pieces testify to the quality of the work, which is beautifully finished and gaily coloured in the technique of *en ronde bosse* enamelling—for example, the "Thorn Reliquary" (*c.* 1400–10; British Museum, London), and the "Goldenes Rössel" at the Stiftskirche, Altötting, Germany (1403).

More seriously, large quantities of private monumental sculpture have been lost in France and the Low Countries. The main sculptor of the French royal family in the second half of the 14th century was a native of Valenciennes, André Beauneveu. His reputation was so widespread that he rather surprisingly earned a mention in the chronicles of Jean Froissart. He produced a large number of monuments, especially for King Charles V, of which several effigies survive (Figure 52). This sculpture, while technically good, is somewhat pedestrian and hardly serves as a prelude to the work of Claus Sluter, who worked for Charles V's brother Philip the Bold, duke of Burgundy.

Claus Sluter

Sluter's surviving work is mainly at Dijon, France, where he was active from about 1390 to about 1406. His figure style is very strongly characterized and detailed and, at times, emotional. This suggests that his origins are German and that he may have come from the region of Westphalia. The intrusive realism of Sluter's work, however, is

Giraudon—Art Resource/EB Inc.

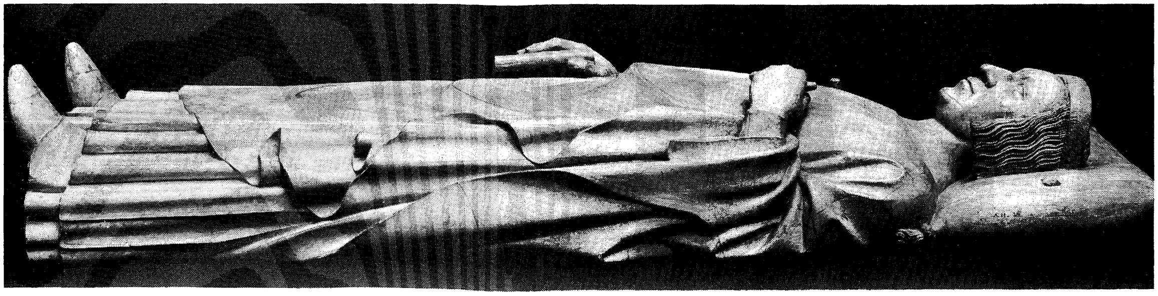


Figure 52: Tomb of Charles V in the abbey of Saint-Denis, France, by André Beauneveu, 14th century.

also symptomatic of a gradual change in sculptural style during this period. The strong characterization of the faces of his figures finds parallels in the near-contemporary triforium busts and Přemyslid tombs in St. Vitus' Cathedral in Prague. Sluter's drapery style, which veers dramatically away from the somewhat reticent elegance of previous court sculpture, also has parallels in the east. Bohemia and Austria possess a series of Madonna figures (*Schöne Madonnen*) swathed in extremely elaborate and artificial drapery arrangements.

The International Gothic sculptural style forms an interesting prelude to developments in Italy, especially to the early work of Donatello and the gradual introduction of Classical ideas into sculpture, for these ideas can be seen as part of a search for an alternative to the elegance of International Gothic. How far Florentines had any knowledge of northern developments is not clear. Ghiberti certainly knew a little about them; moreover, the task of rebuilding Milan cathedral during this period (*c.* 1400) brought large numbers of northern masons across the Alps. As yet, however, the extent to which the sculpture on Milan cathedral was influenced by northern ideas has not been determined.

England stands apart from much of the development represented by Sluter's style. The royal tombs in Westminster Abbey, which extend up to Richard II (died 1400), do not reflect changes subsequent to the phase of André Beauneveu. Further, a fashion for bronze effigies, going back to the effigy of Henry III (1291–93), persisted in England. But whatever the regional idiosyncracies, Westminster tombs, existing as a group in situ, provide a somewhat faded and battered impression of what these great collections of medieval family monuments looked like.

LATE GOTHIC

In the years around 1400, when International Gothic flourished, Italian and northern artists had achieved some sort of rapprochement. Under the renewed influence of antique art, Italy drew away again, and it was not until the 16th century that the north showed any real disposition to follow suit in the imitation of Classical models. While painting and architecture of the 15th century have a reasonably well-defined development, sculptural development is harder to trace—partly because much crucial work (especially in the Low Countries) has been destroyed. It is clear, however, that elaboration rather than restraint was the rule—indeed, the exceptions to the rule (mainly found in France) stand out. This taste for the highly complicated and elaborate—especially in Spain and Germany—was encouraged by the dual influences of painting and architecture. Like the painters, the sculptors enjoyed giving extremely realistic detail and expression to their figures; and, like the architects, they enjoyed complicated tracery work, often encasing their compositions in tabernacle-like enclosures of brilliantly fantastic architecture. To 20th-century eyes, the result may seem overloaded and the total impression exhausting; but in its time the work of, for example, Michael Pacher or Veit Stoss must have been admired precisely for the way in which the sculptor used every conceivable opportunity to display his virtuosity.

The taste for the elaborate

One interesting characteristic of the late Gothic period deserves comment: the increase in the amount of art produced by foreign artists for countries such as Hungary, Poland, and Scotland. Competition between countries for the work of the best artists was not new. Throughout the Middle Ages artists travelled widely. In the 13th century Villard de Honnecourt went from northern France to Hungary, and Roman marble workers journeyed to Westminster. In the period *c.* 1400 there was much interchange between northern and southern Europe. In the 15th century, this general pattern was confirmed; the Netherlandish sculptor Gerhaert Nikolaus von Leyden, for instance, became court sculptor in Vienna, and the Italian sculptor and architect Andrea Sansovino served the Portuguese court in the 1490s. There is also the work of the Franconian sculptor Veit Stoss for the Polish court at Cracow (*c.* 1480) and the work of Bernt Notke of Lübeck for Aarhus (Denmark), Tallinn (Estonia), and Stockholm (*c.* 1470–90). Numerous other objects could be added. More specifically, there is the altar executed by Meister Francke of Hamburg for Helsingfors (1420s) and Hugo van der Goes' panels for the Palace of Holyrood, near Edinburgh (1470s).

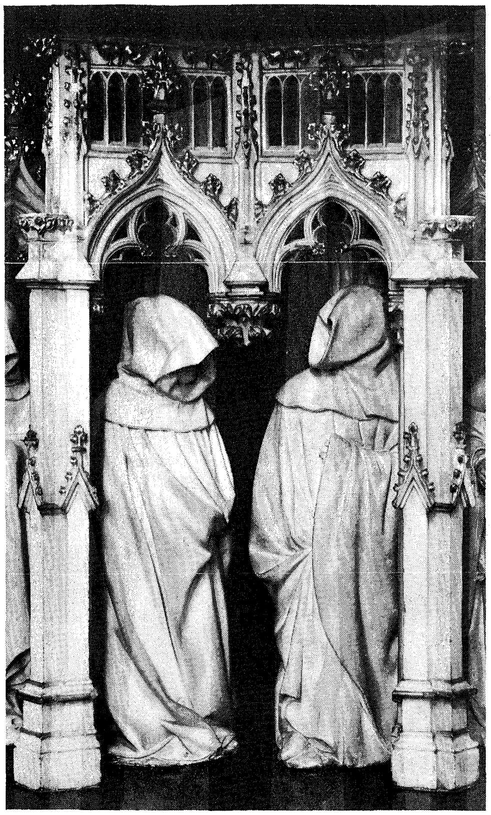


Figure 53: Weepers from the tomb of Philip the Bold, by Claus Sluter and Claus de Werve, completed 1411. In the Musée des Beaux-Arts, Dijon, France. Height 40.6 cm.
R. Remy

Sluter's work for the court of Burgundy lasted about 15 years. During this time, he worked on three major items: the main portal of the chapel of the Charterhouse near Dijon; inside the chapel, the tomb of his patron, Philip the Bold (Figure 53); and a large Calvary group for the Charterhouse cloisters. When he died in 1406, the continuance of his work was assured by the employment of his nephew and heir, Claus de Werve, until his death in 1439. Further, the pattern of the finally completed tomb of Philip the Bold became famous immediately and was frequently imitated all over Europe.

The forcefulness and boldness of Sluter's sculpted figures is combined with elaborate decorative work—on the canopy of the tomb of Philip the Bold, for example. A similar decorativeness is found in the contemporary carved Dijon altarpieces of Jacques de Baerze. The combination remained more or less constant for the rest of the Gothic period.

The spread of this style is hard to trace. In Germany, the most interesting artists worked in the second half of the century. Two of the more important sculptors were Gerhaert Nikolaus von Leyden and Michael Pacher of Brunico. They were followed by a number of virtuoso southern German artists: Veit Stoss of Nürnberg, Tilman Riemenschneider of Würzburg (Figure 54), and Adam Kraft of Nürnberg. In northern Germany, the most original figure was Bernt Notke of Lübeck. Much of the fantastic decorative involvement of his work may now seem overwhelming. The love of realistic detail is well illustrated by Notke's monumental group of St. George and the Dragon (St. Nicholas' Church, Stockholm; Figure 55), where the dragon's spines are made from real antlers. The group as a whole is, of course, of wood, a medium that could be employed to create intricate, open, thin, and spiky forms impossible in stone.

Bernt Notke of Lübeck

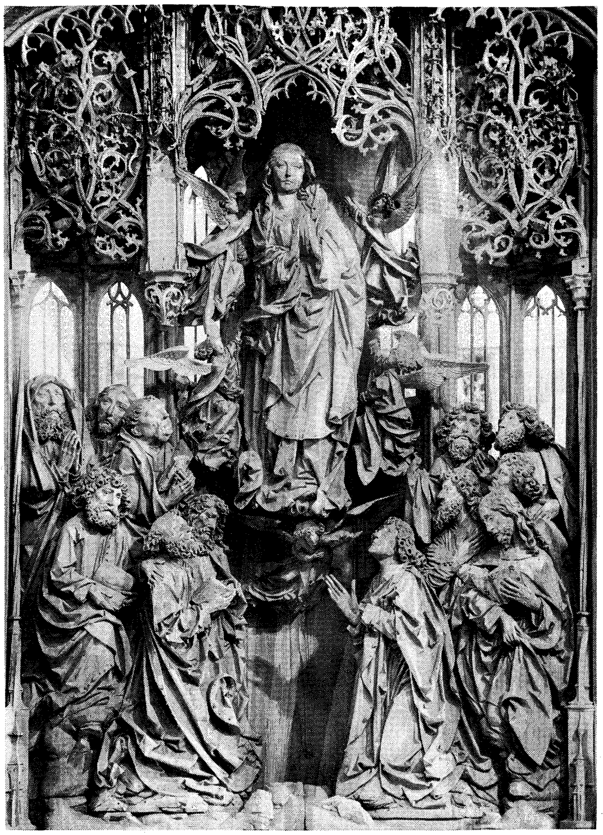


Figure 54: "The Assumption of the Virgin," part of the Altar of the Virgin by Tilman Riemenschneider, 1505–10. In the Herrgottskirche, near Creglingen, Germany.

From Herman Flesche, Gunther and Klaus Beyer, *Tilman Riemenschneider* (1957); published by Veb Verlag der Kunst, Dresden

On the whole, the sculpture produced in France seems to show more decorative restraint. Certainly, the chief French works surviving take the form of large groups, as in the Tonnerre "Entombment" (1450s), or of architectural schemes in which the decoration is clearly subordinate to the figures, as in Châteaudun, Castle Chapel (*c.* 1425).

Restraint is also notable in the chantry chapel of Richard Beauchamp, earl of Warwick (*c.* 1450; Warwick), which has some obvious motifs taken over from the workshop of Sluter. But many of the chantry chapels so common in 15th-century England—for instance, the Henry V Chantry, Westminster Abbey (1440s), or the chantries of John Alcock (*c.* 1488) and Nicolas West (*c.* 1534) at Ely cathedral—show an extraordinary mixture of sculpture and tracery work more reminiscent, as an expression of taste, of Germany or Spain.

The full impression of such profusion can now best be judged from the Chapel of Henry VII (*c.* 1503–*c.* 1515; Westminster Abbey), which is unique in England for the amount of sculpture that has been preserved.

Spanish 15th-century sculpture also tended to be extremely ornate. A number of huge, carved high altarpieces survive—for instance, in the cathedrals of Burgos (1486–88) and Toledo (begun 1498). Some of the altar pieces, like that at Toledo, were designed and executed under the direction of German or Netherlandish artists (Figure 56).

The change from late Gothic to Renaissance was superficially far less cataclysmic than the change from Romanesque to Gothic. In the figurative arts, it was not the great shift from symbolism to realistic representation but a change from one sort of realism to another.

Architecturally, as well, the initial changes involved decorative material. For this reason, the early stages of Renaissance art outside Italy are hard to disentangle from late Gothic. Monuments like the huge Franche-Comté chantry chapel at Brou (1513–32) may have intermittent Italian motifs, but the general effect intended was not very different from that of Henry VII's Chapel at Westminster. The Shrine of St. Sebaldus at Nürnberg (1508–19) has the general shape of a Gothic tomb with canopy, although much of the detail is Italianate. In fact, throughout Europe the "Italian Renaissance" meant, for artists between about 1500 to 1530, the *enjolivement,* or embellishment, of an already rich decorative repertoire with shapes, motifs, and figures adapted from another canon of taste. The history of the northern artistic Renaissance is in part the story of the process by which artists gradually realized that Classicism represented another canon of taste and treated it accordingly.

But it is possible to suggest a more profound character to the change. Late Gothic has a peculiar aura of finality about it. From about 1470 to 1520, one gets the impression that the combination of decorative richness and realistic detail was being worked virtually to death. Classical antiquity at least provided an alternative form of art. It is arguable that change would have come in the north anyway and that adoption of Renaissance forms was a matter of coincidence and convenience. They were there at hand, for experiment.

Refot, Stockholm

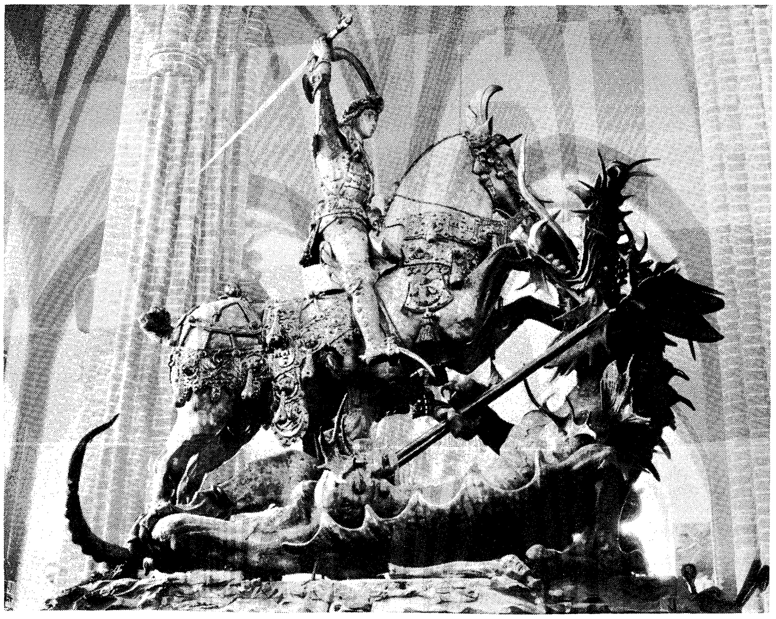


Figure 55: "St. George and the Dragon," wood sculpture by Bernt Notke, 1483–89. In St. Nicholas' Church, Stockholm. Approximately 3.05 × 4.19 m.

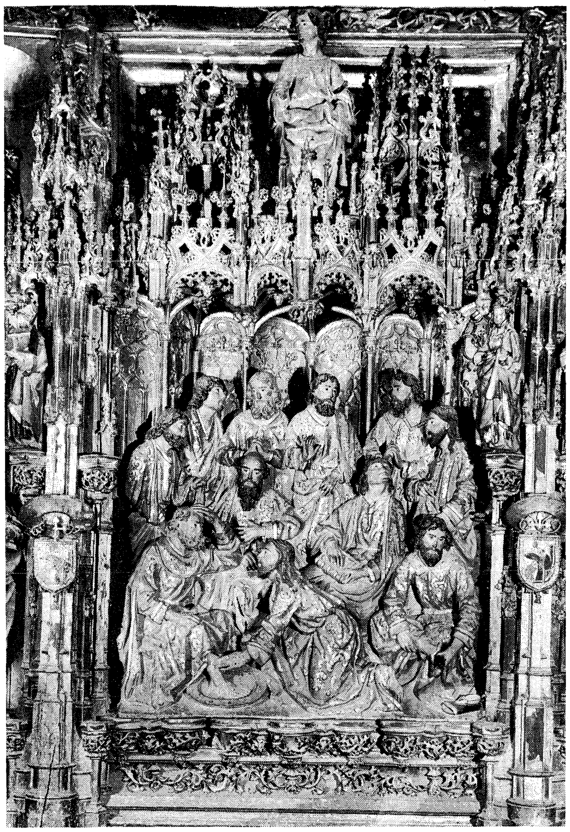


Figure 56: "The Washing of the Feet," polychromed wood, detail from the retable of the High Altar, 1498. In Toledo cathedral, Spain.

Archivo Mas, Barcelona

The "rightness" of Classical antiquity

Their use was certainly encouraged, however, by the general admiration for Classical antiquity. They had a claim to "rightness" that led ultimately to the abandonment of all Gothic forms as being barbarous. This development belongs to the history of the Italian Renaissance, but the phenomenon emphasizes one aspect of medieval art. Through all the changes of Romanesque and Gothic, no body of critical literature appeared in which people tried to evaluate the art and distinguish old from new, good from bad. The development of such a literature was part of the Renaissance and, as such, was intimately related to the defense of Classical art. This meant that Gothic art was left in an intellectually defenseless state. All the praise went to ancient art, most of the blame to the art of the more recent past. Insofar as Gothic art had no critical literature by which a part of it, at least, could be justified, it was, to that extent, inarticulate. (A.Ma./Ed.)

The Renaissance

ITALY

The revival of Classical learning in Italy, which was so marked a feature of Italian culture during the 15th century, was paralleled by an equal passion for the beauty of Classical design in all the artistic fields; and when this eager delight in the then fresh and sensuous graciousness that is the mark of much Classical work—to the Italians of that time, seemingly the expression of a golden age—became universal, complete domination of the Classical ideal in art was inevitable.

This turning to Classical models was less sudden and revolutionary than it seemed. Throughout the history of Romanesque and Gothic Italian art, the tradition of Classical structure and ornament still remained alive; again and again, in the 12th and 13th centuries Classical forms—the acanthus leaf, moulding ornaments, the treatment of drapery in a relief—are imitated, often with crudeness, to be sure, but with a basic sympathy for the old imperial Roman methods of design. Nicola Pisano, at work in the mid-13th century, was but the first of many Italian artists, particularly sculptors, to turn definitely to Roman antecedents for inspiration.

Early Renaissance. Sculpture was the first of the arts in Florence to develop the Renaissance style. Some would date the beginning of the Renaissance to the sculptural competition in 1401 for the bronze doors of the Baptistery of the cathedral of Florence; others would propose the commission to Donatello and Nanni di Banco in 1408 for four seated saints for the facade of the cathedral. The competition reliefs for the bronze doors, submitted in 1402, reveal a change in attitude toward sculpture, and the figures of the Evangelists are the manifestation of that change. The development of Florentine sculpture roughly parallels the development in painting from a dignified monumental style to a relaxed sweetness, although there is no one in painting to approach the rich inventive genius of Donatello.

The genius of Donatello

Donatello, like his friends the architect Brunelleschi and the painter Masaccio, was one of the most outstandingly original artists in Western history. He undoubtedly was influenced by the concepts of antiquity current in Florence, but there was relatively little antique sculpture visible for him to study in his formative years. He first appears as a mature genius working on two of the major projects of the 15th century, the sculptural decoration of the cathedral of Florence and of the guild church of Or San Michele.

His "St. George," begun *c.* 1415 for the niche of the Armourer's Guild at Or San Michele, indicates the new direction in sculpture (Figure 57). Here he reveals such a deep knowledge of the human figure at rest and in movement that he may already have begun his investigation into proportion and the statics and dynamics of the human figure. But the tension between repose and action—the representation, in fact, of pause—also is a psychological achievement, hardly to be matched in earlier sculpture. It is noteworthy, too, that the monumental simplicity and power of the piece is achieved by such a subtle manipulation of the planes and such a technical virtuosity in

Alinari—Art Resource/EB Inc.

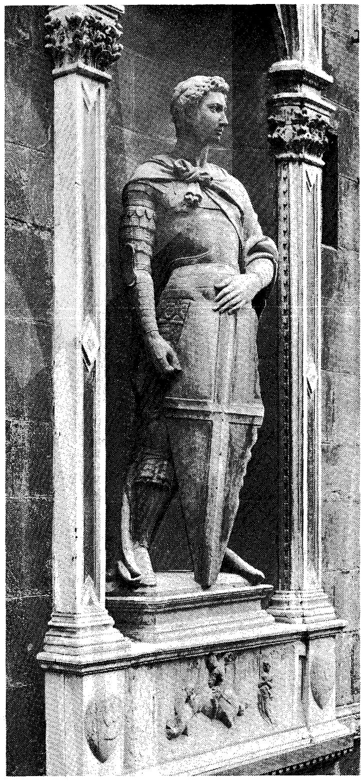


Figure 57: "St. George," bronze copy of a marble statue by Donatello, begun *c.* 1415. In Or San Michele, Florence. (The original statue has been transferred to the Bargello, Florence.) Height 2.08 m.

carving the marble that the observer is rarely concerned with the material. The figure is neither flesh nor stone; it simply is.

In the relief under the niche occupied by "St. George," Donatello introduced another great innovation that was to have unlimited repercussions in Florentine art. Relief has always been a problem for sculptors because it must follow a narrow path between the two-dimensionality of painting and the three-dimensionality of full-round sculpture. Donatello conceived of a very low relief in which the subtle modelling of planes suggests the illusion of depth and figures moving in space while still respecting the integrity of the plane. He continued to develop the potentialities of this relief style throughout his long career and strongly determined the kind of relief sculpture executed in Florence.

In his brief career Nanni di Banco was as prolific and inventive as Donatello. In his earliest works, such as the "Isaiah," he approached more closely the Classic ideal than did Donatello, and in his late work at the Porta della Mandorla he began to evolve a relaxed style that was to have its greatest impact after mid-century. About 1411–13 he executed the "Quattro Santi Coronati" ("Four Crowned Saints") for the niche of the woodworkers and stoneworkers guild at Or San Michele. In this commission he solved one of the most difficult problems facing the sculptor, that of the group conceived in the round. Although some of the figures still retain certain Gothicizing elements in the draperies and in the heads, the major impression is of a group of Roman senators born again in the Renaissance. The group is bound together by the spatial relation of one to the other and by a kind of mute conversation in which they are all engaged.

Ghiberti's "Gates of Paradise"

Lorenzo Ghiberti won the competition for the bronze doors of the Baptistery. He began work in 1403 and set the doors in place in 1424. Ghiberti's fame rests upon his second set of doors, the "Gates of Paradise" (1425–52). The gilded bronze reliefs are treated almost like paintings, for they are rectangular in format and contained within a frame. Unlike the earlier doors, in which the ground plane is simply a neutral backdrop, it is here treated in such a way that it suggests sky and space. Figures are placed in landscape or in perspectivally rendered architecture to suggest a greater depth to the relief than actually exists. At the time that he was executing his first set of bronze doors, Ghiberti undertook to cast the first life-sized bronze statue since antiquity, his "St. John the Baptist" (1412–16) for Or San Michele. Although the figure and its draperies reveal Ghiberti's strong adherence to a late Gothic style, with this work he moved technically into the Renaissance. The influence of Donatello and Nanni di Banco liberates the "St. Matthew" of 1419–22, for Or San Michele, from the older traditions. Ghiberti achieved fame in his own time as a bronze founder and as the master of the shop in which many sculptors and painters of the early Renaissance were trained.

The Sienese sculptor Jacopo della Quercia was the most important sculptor of 15th-century Siena. He executed the Fonte Gaia (1414–19), a public fountain for the Piazza del Campo, the main square of Siena, and was awarded the commission for a baptismal font in the baptistery of Siena cathedral. Always a procrastinating artist, he postponed work on the font to such a degree that the reliefs were finally awarded to other sculptors, including Donatello and Ghiberti. Jacopo's major work is the relief sculpture around the main portal of S. Petronio, Bologna (1425–38). The sculptural treatment of the low relief figures and the suggestion of a space adequate to contain them parallels the painting of Masaccio. The dramatic vigour and powerfully conceived forms had a great influence on the young Michelangelo.

Donatello dominated Florentine sculpture of the second quarter of the 15th century. He executed a series of prophets and a "Cantoria," or singing balcony, for the cathedral, saints for Or San Michele, decorative reliefs and bronze doors for the Old Sacristy of S. Lorenzo, and a bronze "David" (now in the Bargello, Florence) that comes closer to recapturing the spirit of antiquity than any other work of the early Renaissance—indeed, the very idea of a free-standing sculpture of a nude hero was without precedent since antiquity. During the decade 1443–53 Donatello was in Padua executing the equestrian statue of Gattamelata to stand in front of the church. Erasmo da Narni, called Gattamelata, was a condottiere, or leader of mercenary troops, who rose to a position of importance. The statue is an idealization of nature in both horse and rider and a reinterpretation of antiquity. Donatello certainly knew the antique statue of Marcus Aurelius in Rome during his stay there (1431–33). He uses the concept of antiquity, the pose of the antique bronze horses at St. Mark's in Venice, and the forms of the war-horse of his own time. The rider is clothed in quasi-antique armour and bears little or no resemblance to the effigy on Gattamelata's tomb inside the church. Donatello is not concerned with particulars but with the idealized and generalized aspects of man that reveal his potential nobility. The "Gattamelata" states the basic concept that almost all equestrian statues have followed since that time. Donatello's presence in Padua gave rise to a productive local school of bronze sculptors and workers, and his reliefs on the high altar there influenced painters and sculptors of northern Italy.

Donatello's "Gattamelata"

One of his first works upon his return to Florence was a wooden statue of Mary Magdalene for the baptistery of the cathedral. The nervous energy and conscious distortion of forms that may be detected in all his work becomes explicit in the emaciated figure clothed in her own hair. This same emotionalism and distortion is even more pronounced in his last work, the pulpits for the church of S. Lorenzo in Florence.

Antonio Pollaiuolo expresses in his sculpture the same sort of muscular activity and linear movement as in his painting—he has the energy but not the interest in emotion found in Donatello. His small bronze "Hercules and Antaeus" (*c.* 1475; Bargello, Florence) is a forceful depiction of the struggle between these two powerful men from classical mythology. The angular contours of the limbs and the jagged voids between the figures are all directed toward expressing tautness and muscular strain, and the work is one of the earliest examples of the statuette in modern times.

The popularity of small bronzes, usually of secular, often of pagan, subjects and sometimes objects of utility (inkwells, candleholders, and so on), increased in popularity toward the end of the century. The elegant, polished antique gods made by Antico in Mantua and the brilliantly modelled satyrs made by Riccio in Padua set a standard in such works that has never been excelled. Bronze statuettes were made by almost all the major sculptors of the 16th century in Italy.

The "sweet style" of the Florentine school

In complete contrast with Pollaiuolo, Desiderio da Settignano is perhaps best known for his portraits of women and children, although he also executed two public monuments of major importance in Florence—the tomb of Carlo Marsuppini in Sta. Croce (*c.* 1453–55) and the "Tabernacle of the Sacrament" in S. Lorenzo (1461). The tabernacle, which was probably assembled and completed by assistants after Desiderio's death, indicates the new trends taking shape in Florentine sculpture. The central panel employs a perspectivized space. The figures moving into that space are defined in a linear manner that emphasizes contours and billowing draperies to suggest movement. The lateral, full round figures of angels are modelled with a delicacy and subtlety of surface to create relaxed and sweet figures very different from Donatello's strong, virile early saints.

Antonio Rossellino collaborated with his older brother Bernardo on the tomb of Leonardo Bruni (*c.* 1445–49) in Sta. Croce but soon became the dominant personality in the family business. The great sculptural complex of the Cardinal of Portugal tomb (1461–66) in S. Miniato al Monte at Florence reveals the same general tendencies as Desiderio's contemporary work. The tomb is decorated with soft and relaxed angels and a tender Madonna and Christ Child in the roundel (Figure 58). Similar tendencies can be found in such artists as Agostino di Duccio, Mino da Fiesole, and Luca della Robbia.

Andrea del Verrocchio was more interested than these sculptors were in movement, which he expressed in a

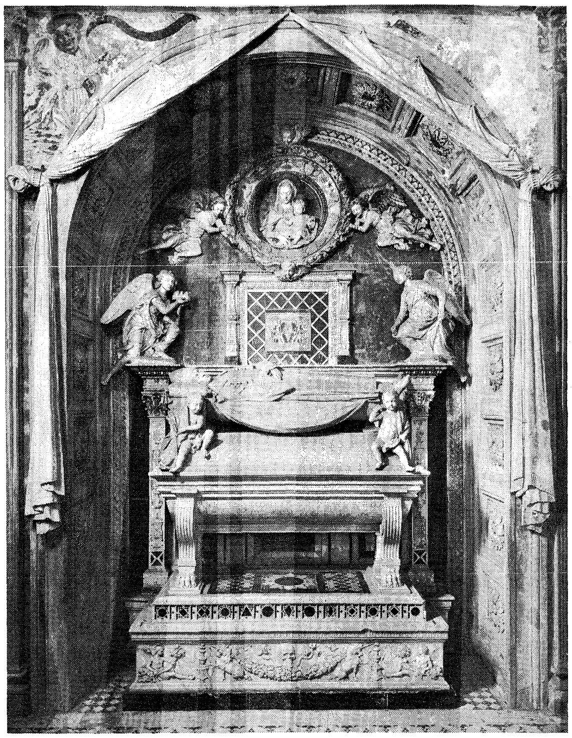


Figure 58: The Cardinal of Portugal tomb, marble sculptural complex by Antonio Rossellino, 1461–66. In the church of S. Miniato al Monte, Florence.

Alinari—Art Resource/EB Inc.

somewhat restrained manner. His group of "Christ and St. Thomas" for Or San Michele (*c.* 1467–83) solves the problem of a crowded niche by placing St. Thomas partly outside the niche and causing him to turn inward toward the figure of Christ. His large equestrian statue of Bartolomeo Colleoni (1483–88) in Venice descends from Donatello's "Gattamelata," but a comparison of the two works reveals Verrocchio's evidence of greater interest in movement. The "Putto with Dolphin" (*c.* 1479; formerly in the Palazzo Vecchio, Florence, but now replaced by a copy) is at once an exquisite fountain decoration, an antique motif restated in Renaissance terms, and the clearest statement of Verrocchio's interest in suggested movement. The child in the piece is seen to be turning; the movement is reinforced by the fish, and the suggestion of motion culminates in the actual movement of the water spouting from the dolphin's mouth. Verrocchio also reveals his indebtedness to Desiderio in the way he treats the surfaces.

High Renaissance and Mannerism. Sixteenth-century sculpture is dominated by the figure of Michelangelo. Although he was born and trained in the 15th century, his style and the bulk of his creations place him firmly in the 16th century. Michelangelo's example was so powerful that Mannerist Florentine artists such as Bartolommeo Ammannati and Baccio Bandinelli could only struggle feebly against it. Others, such as Vincenzo Danti, found it easier to succumb and to follow docilely. Jacopo Sansovino effectively escaped the influence of Michelangelo by transferring his activities to Venice. In Padua a group of bronze workers continued to develop the tradition of fantastic and often beautiful small bronzes that had its origins in Donatello's shop. It was only toward mid-century with artists such as Benvenuto Cellini or at the end of the century with Giambologna that Florentine sculpture found individuals who were able to assimilate Michelangelo's pervasive influence.

Michelangelo Buonarroti is said to have learned sculpture from the minor Florentine sculptor Bertoldo di Giovanni, who provided a link with the tradition of Donatello. An early work, the "Madonna of the Stairs" (*c.* 1492; Casa Buonarroti, Florence), reflects a type of Donatello Madonna and Donatello's very low relief. After the expulsion of the Medici from Florence, Michelangelo fled to Bologna; there he executed three figures for the tomb of S. Domenico and saw the powerful reliefs of Jacopo della Quercia. By 1496 he was in Rome, where he carved a "Bacchus," now in the Bargello, Florence. Michelangelo recaptures the antique treatment of the young male figure by the soft modulation of contours. The figure seems to be slightly off-balance, and the parted lips and hazy eyes suggest that he is under the influence of wine. The little faun also joins in the Bacchic revel by slyly stealing some grapes. In his first major sculptural work the 21-year-old artist succeeded in capturing the spirit of the antique as no artist before him had done. The "Pietà" (today in St. Peter's), commissioned by a French cardinal, was begun immediately upon the completion of the "Bacchus." The motif of the pietà is German in origin, but it is so completely transformed by Michelangelo that the work is one of the harbingers of the High Renaissance. The robes of the Madonna are exaggerated to create a solid base for the pyramidal composition. The figure of Christ is bent and twisted, in part to express the suffering of the crucifixion and in part to make it conform to the contours of the pyramid. All is directed toward creating a calm, dignified, and stable composition that expresses emotion and religious fervour by implication rather than by overstatement. The work is carried to a higher degree of finish than any of the succeeding works, and it is one of the few that Michelangelo signed.

Michelangelo's mastery of the antique

In 1501 Michelangelo was recalled to his native city of Florence to execute an over-life-size figure of "David." When the piece was completed, Michelangelo's contemporaries judged it too important to place out of sight high up on the cathedral, as had been originally proposed, and a committee voted to place it in front of the Palazzo Vecchio, the seat of Florentine civic government. Michelangelo's technical virtuosity is dramatically demonstrated by the fact that he extracted a figure about 14 feet (four metres) tall from a spoiled block. The youthful David was one of the symbols of Florence. Michelangelo sees him as a slightly awkward adolescent with large hands

Michelangelo's "David"

Alinari—Art Resource/EB Inc.

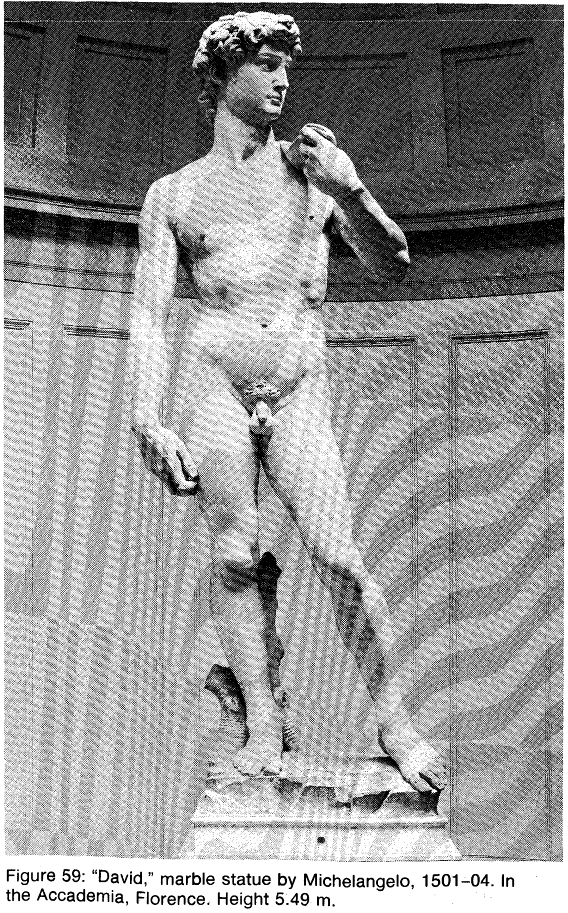


Figure 59: "David," marble statue by Michelangelo, 1501–04. In the Accademia, Florence. Height 5.49 m.

and feet, the body of a boy, and the head of a young man—a powerful figure who has not yet realized his full potential. The balance of the figure is subtly arranged to keep the bearing leg under the head while permitting the apparently nonbearing leg to be relaxed. The positions are reversed in the arms, giving the cross-axis balance of working and relaxed members. The head turns to the left to meet Goliath and the stone of the sling is concealed in the right hand. It is this subtle balance and adjustment of parts to create a unified and harmonious whole that places this work firmly in the High Renaissance style that was appearing simultaneously in painting and architecture (Figure 59).

While in Florence from 1501 to 1505, Michelangelo carved "Madonna and Child" for Notre-Dame in Brugge. He began but did not finish a "St. Matthew" for the cathedral, and he painted the "Holy Family" (*c.* 1503–05; Uffizi, Florence), his reply to Leonardo's eminently popular "The Virgin and Child with St. Anne." In competition with Leonardo he began but did not finish the "Battle of Cascina" for the Palazzo Vecchio. On command of Julius II he returned to Rome.

The Roman years (1506–16) are characterized by what Michelangelo later called the tragedy of the tomb. He had been called to Rome to execute a monumental sepulchre for Pope Julius II. The Pope's financial difficulties and the jealousies of the papal court diverted the artist from the tomb to the painting of the Sistine ceiling. The death of Julius in 1513 caused the heirs to press for a smaller tomb and rapid completion. After many years of negotiations, in 1545 a much-reduced version was set in place in S. Pietro in Vincoli, instead of in St. Peter's as originally planned. The figures by Michelangelo for the tomb are now widely scattered. Only the "Moses" remains in place from the original projects. This figure, which recalls so strongly Donatello's "St. John the Evangelist," was intended to be placed well above the observer's head and is so adjusted. The "Dying Slave" and the "Bound Slave" are now in the Louvre. The "Victory," also intended for the tomb, was executed *c.* 1532–34 in Florence, where it has remained. Four unfinished figures of slaves were carved before 1534 and remained in Florence, where they once formed part of the grotto decoration at the Pitti Palace.

With the election of Pope Leo X in 1513, Michelangelo was diverted from his projects and sent to Florence to design a facade for S. Lorenzo, a church under Medici patronage. Although Michelangelo promised that the facade would become the showplace of Italian sculpture, nothing came of the project. He was assigned instead to construct a tomb chapel as a pendant to Brunelleschi's Old Sacristy, and later to provide suitable housing for the Medici library in S. Lorenzo. While engaged in these projects Michelangelo was also put in charge of the fortifications of Florence prior to and during the siege of 1529. He complained, justly, that no one can plan and execute three projects simultaneously.

The Medici tombs (1520–34) gave the artist the opportunity to plan the architectural setting of his sculpture and to control both the light cast on the work and the position of the observer. Since the chapel was originally planned to contain the tombs of the Medici popes Leo X and Clement VII, it is best seen from behind the altar, where the papal celebrant of the mass for the dead would have stood. On the left is the tomb of Giuliano, on the right the tomb of Lorenzo, and before the observer the Madonna and Christ Child with the Medici patron saints, Cosmas and Damian; and beneath the two sarcophagi respectively lie the recumbent figures of "Night" and "Day," and "Dawn" and "Dusk."

The "Pietà," or "Deposition," in the museum of the cathedral of Florence dates from around 1550 and may have been intended by Michelangelo for use in his own tomb. The figure of Nicodemus is a self-portrait and indicates Michelangelo's deep religious convictions and his growing concern with religion. His final work, the "Rondanini Pietà" (1552–64), now in the Castello Sforzesco, Milan, is certainly his most personal and most deeply felt expression in sculpture. The artist had almost completely carved the piece when he changed his mind, returned to the block, and drastically reduced the breadth of the figures. He was working on the stone 10 days before he died, and the piece remains unfinished. In its rough state the "Rondanini Pietà" clearly shows that Michelangelo had turned from the rather muscular figure of Christ of his earlier works (as can be seen from the partially detached original right arm) to a more elongated and more dematerialized form.

Followers of Michelangelo

Whether in Rome or Florence, Michelangelo had a strong influence on sculptors of the 16th century. Vincenzo Danti followed closely in Michelangelo's footsteps. His bronze "Julius III" of 1553–56 in Perugia is derived from Michelangelo's lost bronze statue of Julius II for Bologna. Many of his figures in marble are only free variations on themes by Michelangelo. In much the same way, Baccio Bandinelli attempted to rival the monumentality of Michelangelo's "David" and the complexity of his "Victory" in the statue of "Hercules and Cacus" (1534), which was placed as a companion to the "David" in front of the Palazzo Vecchio. Bartolommeo Ammannati should be best known for his design of the bridge of Sta. Trinità in Florence, but his most visible work is the Neptune Fountain (1560–75) in the Piazza della Signoria, with its gigantic figure of Neptune turned toward the "David" in presumptuous rivalry.

Benvenuto Cellini through his celebrated autobiography has left a fuller account of his picturesque life than that of any other artist of the 16th century. He was in Rome from 1519 to 1540 and was one of the defenders of the pope during the siege of the Castel Sant'Angelo. In France from 1540 to 1545, he executed there the celebrated saltcellar for Francis I and the "Nymph of Fontainebleau" (Louvre). The saltcellar is at once an example of 16th-century conspicuous consumption and of Mannerist conceits in art. It is of solid gold, which is covered in part by enamels as though it were a base metal. It was designed for use as a functional object upon the King's table to hold nothing more than common table salt. On his return to Florence in 1545 Cellini received the commission to cast the bronze "Perseus," now in the Loggia dei Lanzi, Florence, which he describes in some detail in his *Autobiography.* The youthful figure of Perseus seems to retain some of the airiness from his flight on the winged sandals of Hermes. He holds aloft the head of the Medusa in an outstretched arm, thus creating an open composition that exploits to the full the potential of the bronze medium. Void is almost as important as solid in this light and airy composition that would have been unthinkable and impossible in marble. Cellini intended the figure to be seen from a variety of viewing points, a relatively new idea in sculpture of this sort, and he leads the observer around by the position of the arms and the legs.

Giambologna

Florentine sculpture at the end of the 16th century was dominated by the Fleming Giambologna and by his shop assistants. Giambologna went to Italy for study shortly after mid-century and settled in Florence in 1557. His earlier major work in Italy is the Fountain of Neptune (1563–66) in Bologna. By early 1565 he had also cast the earliest of his many versions of the bronze "Flying Mercury" that is his most famous creation. The ideas of Cellini's "Perseus" are here carried to their logical conclusion. The god borne along on the air by his winged sandals touches earth only on the slenderest base possible, which is, in fact, represented as a jet of air from the mouth of a wind god. The statue is perfectly balanced according to principles discovered early in the 15th century, yet the outthrust arms and legs give it a feeling of movement and of lightness. Giambologna understood Michelangelo's *figura serpentinata,* the upward spiralling composition, better than any sculptor of the 16th century. His marble group of the "Rape of the Sabines" (1579–83), in the Loggia dei Lanzi, Florence, interweaves three figures in an upward spiralling composition that prefigures the Baroque. Outside Florence, at the present Villa Demidoff in Pratolino, he carved a figure of the Apennines (1581) that seems to be a part of the living rock; it is an excellent example of late Mannerism, in which a paradoxical relationship between art and nature is often cultivated. As the favourite sculptor of the Medici, Giambologna and his prolific shop dominated Florentine sculpture at the end of the 16th century, training artists who were to carry late 16th-century ideas into the rest of Europe and prepare the way for the nascent Baroque. (J.R.Sp./Ed.)

Venetian sculpture

In sculpture, Venice was less independent of Florence and Rome than in painting. The major 16th-century impetus came from Jacopo Sansovino, a central Italian who arrived in Venice in 1527. Sansovino never adopted the full-scale Mannerism of Florence, and his style retained a High Renaissance flavour, but his pupils Danese Cattaneo and Alessandro Vittoria were selectively able to develop the more mannered aspects of Sansovino's style into a Venetian species of Mannerism.

Vittoria stands closer to Florentine style than his contemporaries in painting, particularly in his decorative work, and his small bronzes display a serpentine grace surpassed only by Giambologna in Florence. His marble figures are, however, often more directly expressive than those of Florentine sculptors. His altarpiece for S. Francesco della Vigna (1561–63) conforms with the attenuated canons of Mannerist elegance. In sculpture as in painting, the narrative Venetian style proved to be more easily adaptable to the demands of the Counter-Reformation than the abstract artiness of central Italian Mannerism. The work of Vittoria and of the painter with whom he was most closely associated, Palma il Giovane, seems to anticipate many of the characteristics of Baroque art.

MANNERIST SCULPTURE OUTSIDE ITALY

In the north of Europe, Giambologna's influence was paramount. Both Hubert Gerhart and Adriaan de Vries, the leading exponents of northern Mannerist sculpture, can be considered as followers of the expatriate Fleming. Gerhart worked (1583–94) for Hans Fugger at Kirchheim, Augsburg, and at Amsterdam under de Sustris, and for the archduke Maximilian I of Bavaria, at whose court he produced bronze figures of considerable accomplishment (1598–1613). De Vries joined Bartholomaeus Spranger in 1601 at Rudolf's court in Prague. His "Psyche with Three Cupids" (Nationalmuseum, Stockholm) is a characteristic example of his stylishness—a wonderful satin finish, spiralling complexity, and a soaring grace reminiscent of Giambologna's "Mercury" (Figure 60).

By courtesy of the Nationalmuseum, Stockholm

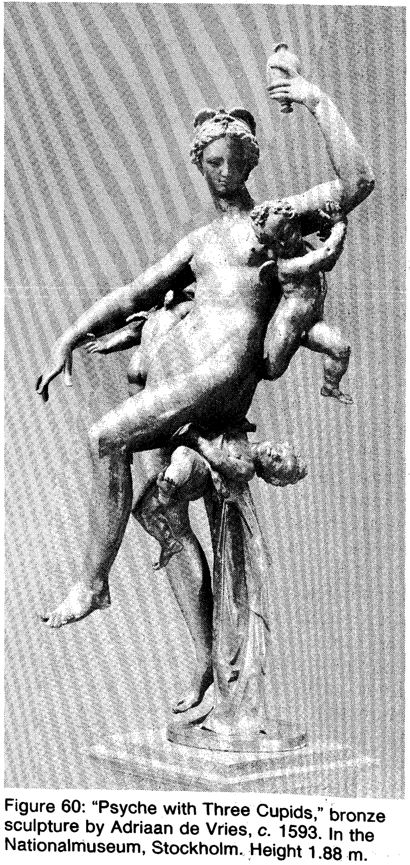


Figure 60: "Psyche with Three Cupids," bronze sculpture by Adriaan de Vries, c. 1593. In the Nationalmuseum, Stockholm. Height 1.88 m.

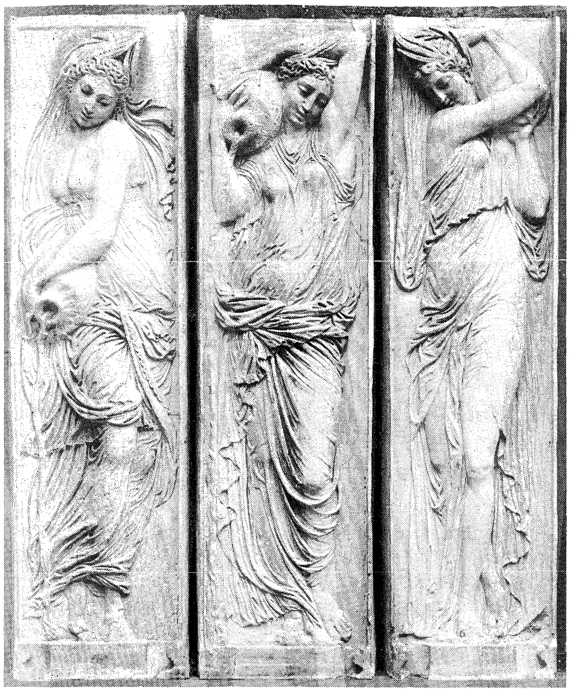


Figure 61: Nymphs from the "Fountain of the Innocents," Paris, relief panels (after plaster casts), by Jean Goujon, 1547–49. About life-size.

J.E. Bulloz

As in painting, France owed its early acquisition of Mannerist sculptural style to Italian artists at Fontainebleau, to Primaticcio's stucco style, and to Cellini. Jean Goujon began from this point of inspiration, and his decorations for the "Fountain of the Innocents" at the Louvre (1547–49) possess a sophisticated refinement *all'antica* unequalled by any non-Italian artist of the period (Figure 61).

The influence of Primaticcio's suave stucco decorations is even more apparent in the early work of the other great French sculptor of the century, Germain Pilon. This is not surprising since his elegant "Monument for the Heart of Henry II" was probably completed under Primaticcio's supervision. His statues for Primaticcio's Tomb of Henry II, however, show him moving toward greater naturalism and expressiveness. In his later works Pilon achieved a freedom of plasticity and feeling for texture that anticipated Baroque developments.

Spanish Renaissance sculpture at first relied heavily upon visiting Italians, led by Andrea Sansovino, but with the advent of Ordóñez, Diego de Siloé, and the painter-sculptors Machuca and Beruguete, a native Spanish school of Mannerism was formed. Like his father (the painter Pedro), Alonso Beruguete studied in Italy. On his return to Spain about 1517, he began to develop an elaborately pictorial style in sculptural complexes of great originality. The fluid quality of his designs reaches its peak in the surging motions of the "Transfiguration Altar" (1543–48) for Toledo cathedral. Beruguete's greatest successor at Valladolid was Pompeo Leoni, who collaborated with his father, Leone, on portraits of Charles V, composed in a disciplined and sternly Roman style, quite different from the expressive fluency of native Spanish sculpture that reemerged at the turn of the century in the few sculptures of polychromed wood by El Greco. (M.J.Ke.)

The Baroque period

ITALY

Early and High Baroque. At the beginning of the 17th century, sculpture in all of Italy, with the exception of Florence, was at a low ebb; and the dry, frankly propagandist nature of the decoration of the Borghese and Sistine chapels in Sta. Maria Maggiore, Rome, reveals this only

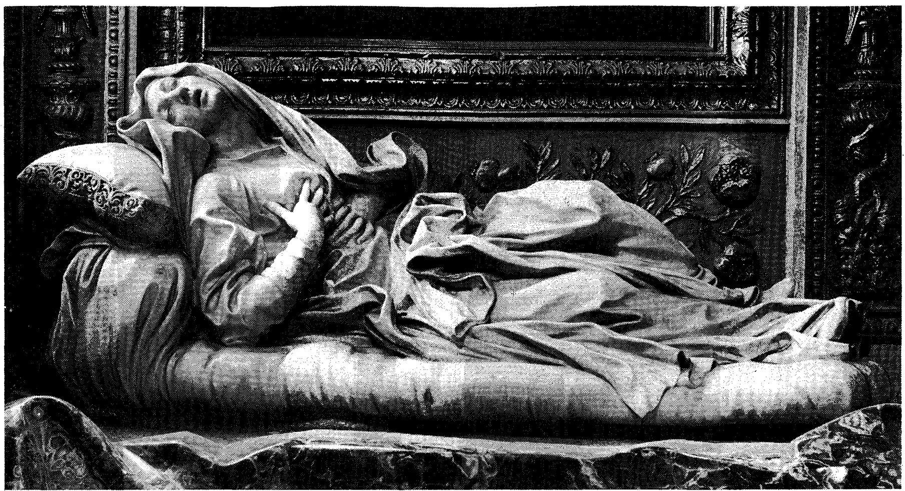


Figure 62: "Ludovica Albertoni," marble effigy by Gian Lorenzo Bernini, *c.* 1674. In the Altieri Chapel, S. Francesco a Ripa, Rome.
Anderson—Alinari from Art Resource/EB Inc.

too clearly. With Stefano Maderno and Camillo Mariani a slightly more imaginative interpretation of the demands of the Council of Trent is to be found, while certain aspects of the work of Pietro Bernini (1562–1629) were to have considerable influence on his son Gian Lorenzo. The first breath of the new Baroque spirit, however, is to be found in the immense vitality of the equestrian monuments in Piacenza (1612–25) by Francesco Mochi; and a comparable fiery vigour is the keynote of the fresco "Aurora" by Guercino in the Casino Ludovisi, Rome (1621–23). The forms are pierced and opened up, and the momentary, unstable poses, with draperies fluttering and tails lashing, give a vivid movement that releases the figures from the Mannerist spell.

No field was more congenial to the spirit of Baroque art than sculpture carried out on a conspicuous scale. The Baroque artist achieved dramatic pictorial unity by abolishing the traditional limits separating painting, sculpture, and architecture. The solid masses of sculpture and even of architecture were made to move in space by means of such motive forms as undulations; sculpture was transformed by such painter's devices as richly varied illusionistic textures, coloured materials, and irregularly dappling light effects.

The genius of Bernini

Gian Lorenzo Bernini, the greatest sculptor of the 17th and 18th centuries, established the sculptural principles for those two centuries in a series of youthful works of unrivalled virtuosity, as the "Apollo and Daphne." Stone was now completely emancipated from stoniness by open form and by an astonishing illusion of flesh, hair, cloth, and other textures, pictorial effects that had earlier been attempted only in painting. These qualities made what his contemporaries called his "speaking portraits" seem unprecedentedly alive; portrait sculpture for two centuries was a variation of these innovations. In the statue of St. Longinus in St. Peter's in Rome, Bernini created the characteristic formula of Baroque sculpture by throwing the draperies into a violent turmoil, the complicated and broken involutions of which are not rationally explained by the figure's real bodily movement but seem paroxysmally informed by the miracle itself. The passion with which he imbued his sculptured figures, capturing the most transitory states of mind, reached its apogee in the representation of the ecstasy of St. Teresa in the Cornaro Chapel, Sta. Maria della Vittoria, Rome (1645–52) and in the figure of the expiring Ludovica Albertoni (Figure 62) in the Altieri Chapel, S. Francesco a Ripa, Rome (*c.* 1674). The former is generally considered the masterpiece of Baroque religious sculpture and shows how Bernini could organize the arts of architecture, painting, and sculpture in an overwhelming assault on the senses that dispels the resistance of the intellect. This ambitious plan was typical of the mature Bernini, whose spiritual and artistic aspirations exceeded the scope of his early secular salon statues. His later works were largely religious and unprecedentedly vast in scale, as in the dazzling "Cathedra Petri," which covers the whole end of St. Peter's in Rome with a teeming multitude of figures.

The tombs of Bernini are magnificent spectacles in which symbolic figures, clothed in sweeping draperies, with rhetorical gesture and expressive features, share in some emotional experience, theatrically depicted. An example is the tomb of Alexander VII in St. Peter's, Rome. The pontiff, set in a great apse, kneels on a high pedestal about which Charity, Truth, Justice, and Wisdom weep disconsolately while Death, a skeleton, raises the great draperies of polychrome and gold that veil a darkened doorway. Another work, the fountain of the Triton in the Piazza Barberini, Rome, from which all clarity of profile or of shadow, all definiteness of plane, are removed, is also characteristic of Bernini's style, widely imitated throughout Europe.

Bernini's art was the basis of all Baroque sculpture, but his example was not always followed, and the work of his more restrained contemporaries, such as Alessandro Algardi (relief of "Meeting of Attila and Pope Leo," 1646–53, St. Peter's, Rome) and the Fleming François Duquesnoy, attracted more approval from theorists of art. The

Giraudon—Art Resource/EB Inc.

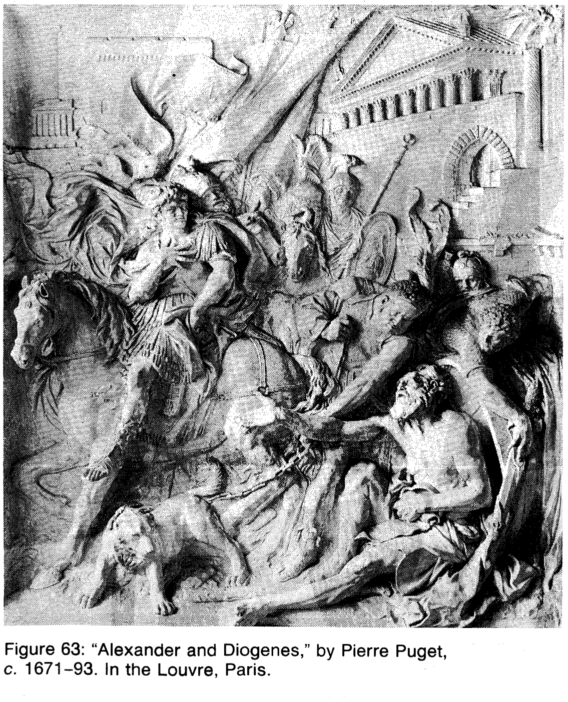


Figure 63: "Alexander and Diogenes," by Pierre Puget, *c.* 1671–93. In the Louvre, Paris.

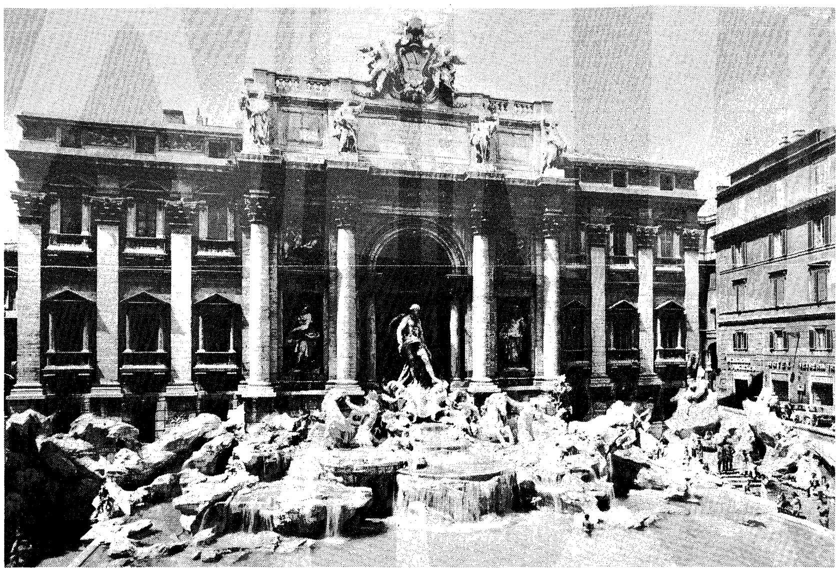


Figure 64: Fontana di Trevi, Rome, designed and begun by Niccolò Salvi (1732) and completed by Giuseppe Pannini, 1762.

Shostal Inc.—EB Inc.

latter's "St. Susanna" in Sta. Maria di Loreto in Rome, a figure after the antique but enlivened with Berninian textures, was originally made to look toward the observer and, with a gesture, to direct his attention to the altar. The distinction between art and life that the Mannerists had cultivated was banished by this active participation of the statue in the viewer's space and activities, another important innovation of Bernini.

Late Baroque. In late 17th-century painting, composition became increasingly decorative rather than structural, and there was a loosening of design in the individual figures as well. This dissolution is also to be found in sculpture of the period, such as in the proto-Rococo figures of Filippo Carcani (active 1670–90) in Rome and, to a lesser extent, in those of Filippo Parodi (1630–1702) in Genoa, Venice, and Naples. Outside Venice and Sicily the true Rococo made little headway in Italy.

A more or less classical late Baroque style, best exemplified by the heroic works of Camillo Rusconi in Rome, was dominant in central Italy through the middle of the 18th century. Rusconi's work had considerable influence outside Italy as well.

The latter half of the century saw the emergence of a much lighter and more theatrical manner in the works of Agostino Cornacchini and of Pietro Bracci, whose allegorical figure "Ocean" on the Fontana di Trevi by Niccolò Salvi (completed 1762) is almost a parody of Bernini's sculpture. Filippo della Valle worked in a classicizing style of almost French sensibility, but the majority of Italian sculpture of the mid-18th century became increasingly picturesque with a strong tendency toward technical virtuosity. Complex sculptured groups designed by Luigi Vanvitelli for the park of the palace at Caserta (*c.* 1770) are almost *tableaux vivants* ("living pictures") in a landscape setting, while the Cappella Sansevero de' Sangri in nearby Naples (decorated 1749–66) is one of the most important sculptured complexes of the time. Allegorical groups by Antonio Corradini and Francesco Queirolo vie with each other in virtuosity and include such conceits as fishnets cut from solid marble and the all-revealing shrouds developed by Giuseppe Sammartino. Florentine sculpture of the 18th century is less spectacular, and Giovanni Battista Foggini took back from Rome the compromise style of Ferrarza, while Massimiliano Soldani-Benzi seems to have been instrumental in the brilliant revival there of small-scale bronze statuettes. Giovanni Marchiori worked in Venice with an attractive painterly style, in part based on the wood carvings of Andrea Brustolon; and Giovanni

Archivo Mas, Barcelona

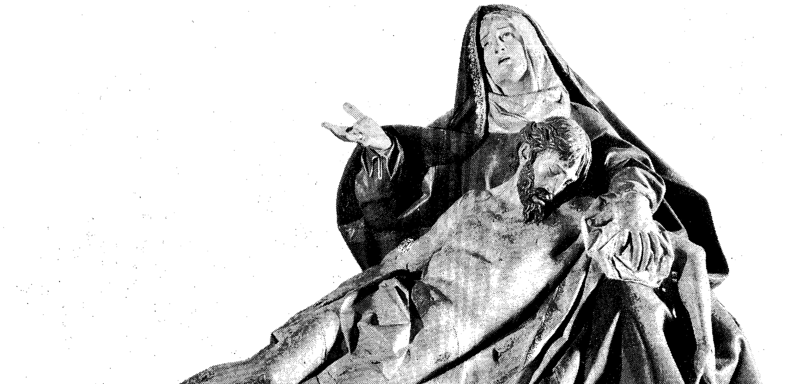


Figure 65: "Pieta," polychromed wood sculpture by Gregorio Hernández, 1617. In the Museo Nacional de Esculturas, Valladolid, Spain. Height 1.8 m.

Maria Morlaiter ran the full gamut to a late 18th-century classicism close to the early works of the great Neoclassical sculptor Antonio Canova.

BAROQUE AND ROCOCO OUTSIDE ITALY

Spain. Spanish sculpture of the 17th and 18th centuries exhibits a greater continuity with late Gothic art than does the painting; and the Counter-Reformation demands for realism and an emotional stimulus to piety led to sculpture with glass eyes, human hair, and even real fabric costumes. Italian Renaissance sculpture had made a very limited impact in Spain, and with few exceptions this was in the court ambience only, while Spanish Baroque sculpture is almost entirely religious and of a fundamentally popular nature. Gregorio Hernández in sculptures like the "Pieta" (1617; Museo Nacional de Esculturas, Valladolid, Spain) revealed an emotional realism more Gothic than Baroque (Figure 65); but in the figures of Manuel Pereira there is a clear-cut monumentality and intense concentration comparable to that of Zurbarán. Both were active in Castile, though the main centre of sculptural activity was Seville and Granada, with Juan Martínez Montañés as the dominant personality. The intense realism and deep spirituality of his figures were followed by his pupil Alonso Cano; but in the figures of Cano's pupil Pedro de Mena, his simple monumentality is replaced by a more picturesque and theatrical gracefulness. José de Mora, also a pupil of Cano, took this process even further. But in general the 18th century saw a sad decline in Spanish sculpture.

Realism and emotionalism of religious sculpture

Flanders. In comparison with painting, the sculpture of the 17th century in the southern provinces is extremely disappointing. The Flemish sculptor François Duquesnoy spent almost all of his career in Rome, while those who remained in Flanders, such as his brother Hieronymus Duquesnoy the Younger, were mostly secondary artists influenced by Rubens. Artus Quellinus the Elder reveals a much more individual style, particularly in his decorations for the Town Hall in Amsterdam, and the tendency toward a painterly style is more pronounced in the work of his son Artus Quellinus the Younger, Rombout Verhulst, and Lucas Faydherbe.

Decline of Antwerp in the late 17th century

The end of the Twelve Years' Truce in 1621 had brought back Antwerp's old troubles, and the control of the Scheldt by the United Provinces was confirmed by the Peace of Westphalia (1648). Economic depression and French aggression in the second half of the 17th century combined to make the southern provinces increasingly provincial, while under the provisions of the Treaty of Utrecht (1713) and the Treaty of Rastatt (1714) the territories passed to Austria. Eighteenth-century painting and sculpture became increasingly weak and provincial, though fantastic pulpits carved by Hendrik Frans Verbruggen, Michel Vervoort, and Theodor Verhaegen provide a remarkable parallel to those in central Europe.

France. Duquesnoy was much admired in France, where the sculptors of Louis XIV (the "Sun King"), such as François Girardon, continued his tradition of setting correct and charming allusions to the antique in a pictorial and spatial context that is wholly Baroque. Girardon's tomb of the Cardinal de Richelieu, in the church of the Sorbonne, Paris, is illustrative of the Baroque monuments of France, calmer and more conservative than those of Italy. The dying cardinal, lying on his sarcophagus and originally gesturing in supplication toward the altar, is upheld by Religion and mourned by Science. The three figures, united by the lines of skillfully arranged draperies, are informed by a solemn and touching sentiment. The academic discipline imposed by the Sun King's ministers, especially Colbert, discouraged less tractable spirits, such as the passionate genius Pierre Puget. His unique expressions of anguish are couched in the physical terms of highly original works like the "Milo of Crotona"; here the composition of a figure rigid with pain is given an almost unbearable tension.

Antoine Coysevox, another of the sculptors of Louis XIV, had begun in the official "academic Baroque" style, but his later works, undertaken after the death of Colbert, are witnesses of the gradual acceptance of the Baroque in France, which now acquired the artistic leadership that Italy had long held over the rest of Europe. At the same time, the style was made lighter, gayer, and more ornamental, in accordance with 18th-century taste, as seen in the famous "Chevaux de Marly" by Guillaume Coustou now marking the entrance to the Champs-Élysées in Paris but designed for Marly, as part of the most innovative outdoor display of sculpture since the 16th-century gardens of Italy. Coustou's bust of his brother Nicolas has a characteristic freshness and informality whereby 18th-century artists avoided the grandeur they found pompous in the Berninian tradition.

This 18th-century style that reduced the Baroque to exquisite refinement was the art of the aristocratic salon and boudoir. The little marble "Mercure" (1741) of Jean-Baptiste Pigalle is almost wholly Berninian, except in its intimacy and deliberate unpretentiousness; even in Pigalle's most ambitious undertakings, the relative scale of the figures is much reduced and the whole composition opened up, in contrast to Bernini's tombs. Nevertheless, the narrative and indeed the allegory of his masterpiece, the tomb of the Maréchal de Saxe (1753; Saint-Thomas, Strasbourg), is as enthralling and memorable as any 17th-century sculpture, although the theme, significantly, no longer seems to be inspired by the Christian faith. At the same time, the more classical current of French sculpture continued and gained importance as the 18th century advanced. The clarified form and continuous, unbroken contours of Étienne-Maurice Falconet's marble "Bather" (1757) adapt the Classic tradition to a pretty and intimate Rococo ideal that is the quintessence of 18th-century taste. This Classicism was purified by Jean-Antoine Houdon, who avoided the playful air of the Rococo boudoir in his "Diana" (*c.* 1777) and his marble nude in the Metropolitan Museum of Art, New York City (1782). His portrait sculptures are the ultimate in the 18th-century refinement of Bernini's tradition.

Jean-Antoine Houdon

In the context of the rather restrained French sculpture of the 18th century, the blatant sensuality of Clodion (byname of Claude Michel) is the exception rather than the rule (Figure 66). Portrait busts by Jean-Baptiste Lemoyne and Pigalle follow the direction taken by Coysevox in his "Robert de Cotte," but Augustin Pajou and Houdon

By courtesy of the Metropolitan Museum of Art, New York, bequest of Benjamin Altman, 1913

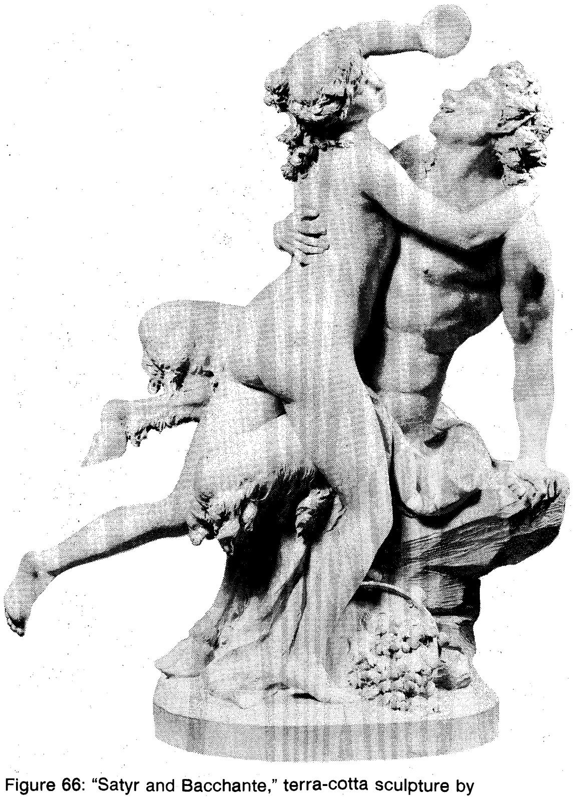


Figure 66: "Satyr and Bacchante," terra-cotta sculpture by Clodion, *c.* 1775. In the Metropolitan Museum of Art, New York City. Height 58.4 cm.

soon abandoned the Rococo in favour of a Neoclassical approach. Edme Bouchardon, however, flirted only briefly with the Rococo and otherwise remained firmly attached to the classicizing tradition of French sculpture.

England. English sculpture of the early 17th century was very provincial, with Nicholas Stone and Edward Marshall the only English-born sculptors to rise above the general level of mediocrity. Their styles were based on contemporary Netherlandish sculpture with small admixtures of Italian influence; and after 1660 the uncomprehending borrowings of John Bushnell from Bernini serve only to make his figures look ludicrous. The most distinguished English-born sculptor of the second half of the 17th century was Edward Pierce, in whose rare busts is to be found something of Bernini's vigour and intensity. But the general run of English sculpture as represented by Francis Bird, Edward Stanton, and even the internationally renowned woodcarver Grinling Gibbons remained unexceptional. It was not until John Michael Rysbrack from Antwerp settled in England in *c.* 1720, followed by the Frenchman Louis-François Roubillac in *c.* 1732, that two sculptors of European stature were active in England. The busts and tombs of Rysbrack and Roubillac have a power and vitality previously unknown in English sculpture; they were responsible for the revival that took place in the 18th century.

Central Europe. While the influence of Giambologna persisted in some quarters, Hans Krumper and Hans Reichle produced bronze figures less indebted to the Classical tradition but with stronger individuality. Jörg Zürn, whose finest wood carvings are to be seen at Überlingen, and Ludwig Münsterman, in Oldenburg, continued in the Mannerist style, whereas Georg Petel, who came under the influence of Rubens, is almost the only sculptor to reveal the impact of the Baroque. Petel's importance lies mainly in his ivories, and Leonard Kern in Franconia developed a similar Rubensian style for his small statuettes.

Influence of Bernini

Painting and sculpture recovered slowly from the ravages of the Thirty Years' War, and some of the earliest reflections of the high Baroque of Bernini are to be found in the sculpture of Matthias Rauchmiller at Trier (1675) and Legnica (Liegnitz) in Silesia (1677).

Among sculptors in Austria the forces of Classicism were stronger; and the weak north Italian late Baroque styles of Giovanni Giuliani and Lorenzo Mattielli were supplanted by the cool elegance and classical refinement of Georg Raphael Donner. His preference for the soft sheen of lead gave Austrian Baroque sculpture one of its most distinctive features.

Bildarchiv Foto Marburg—Art Resource/EB Inc.

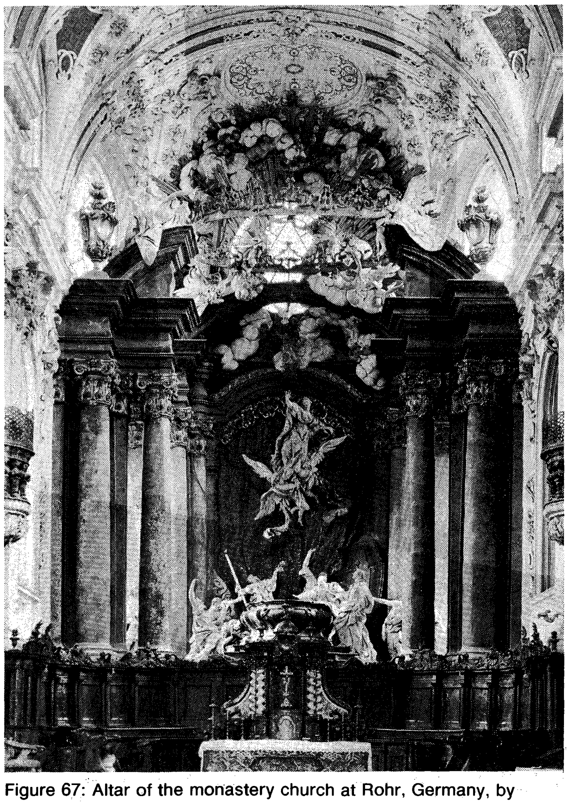


Figure 67: Altar of the monastery church at Rohr, Germany, by Cosmas Damian Asam and Egid Quirin Asam, 1718–25.

During the first four decades of the 18th century, Bohemian Baroque art developed almost independently of Vienna. The brilliant rugged stone sculptures of Matyás Bernard Braun and Ferdinand Maximilián Brokoff, with their dynamism and expressive gestures, were truly Bohemian in spirit.

Bavarian Baroque art in the hands of the brothers Egid Quirin Asam and Cosmas Damian Asam was almost entirely confined to churches, and their brilliant development of the theatrical illusionism of Bernini is achieved in the high altar (Figure 67) of the monastery church at Rohr, in Germany (1718–25), and in St. John Nepomuk in Munich (begun 1733). Cosmas Damian's style as a painter was influenced by Rottmayr as well as by the Italian masters whom he studied during his stay in Italy (1711–14), while the sculptural style of Egid Quirin was formed on the south German tradition of wood carving, as well as on Bernini.

The Saxon school

In Upper Saxony there was also a native tradition before the arrival of Permoser, represented by the heavy figures of Georg Heermann and Konrad Max Süssner, both of whom had been active in Prague in the 1680s. Balthasar Permoser was trained in Florence under Foggini, whence he was summoned to Dresden in 1689. His painterly conception of sculpture, derived from Bernini, is revealed in the complex "Apotheosis of Prince Eugene" (1721; Österreichische Galerie, Vienna) and above all in the sculptural decoration of the Zwinger in Dresden initiated during the second decade. Paul Egell was a pupil of Permoser in Dresden at the time of the Zwinger decorations, and in 1721 he was appointed court sculptor at Mannheim. Egell's elongated and refined Baroque figures were an effective counter to the Classicism of Donner, and his personality was decisive in Franconia and the Palatinate during the first half of the century.

Berlin under the Great Elector of Brandenburg had become an increasingly important centre, both politically and artistically; and the full-bodied Baroque style of Andreas Schlüter, as revealed by his equestrian monument to the Great Elector (1696–1708), now at Charlottenburg, was fully in sympathy with the time.

The Rococo style in Bavaria

No hard and fast division can be made between the Baroque and the Rococo in central and eastern Europe, either chronologically or stylistically. The first Rococo decorative ensembles in Germany, the Reiche Zimmer of the Residenz in Munich, were built by the Frenchman François de Cuvilliés in 1730–37, but in painting and sculpture the situation is more complicated. Ignaz Günther, the greatest south German sculptor of the 18th century, was trained under Johann Baptist Straub; the elongated forms of Egell's sculpture at Mannheim, however, deeply impressed him, and his development was toward an almost Mannerist grace and refinement. Günther was capable of the most extraordinarily sensitive characterization of surfaces, even when painted white; and this he combined with an interpretation of character comparable to the late Gothic sculptors, thus giving his figures a realism and immediacy that is almost uncanny. Apart from their lightness and vivacity, however, it is the figures' relationship to the altars on which they are placed that reveals their Rococo quality. Gone are the great coordinated ensembles of the Asams, and instead each figure has a totally separate existence of its own and a balance is only to be found when the church interior is taken as a whole.

Johann Michael Feichtmayr

Swabian sculpture of the period is characterized by the extremely successful partnerships between the sculptors and stucco artists. For Zwiefalten and Ottobeuren Joseph Christian provided the models from which Johann Michael Feichtmayr created the superb series of larger than life-size saints and angels that are the glory of these Rococo interiors. Feichtmayr was a member of the group of families from Wessobrunn in southern Bavaria that specialized in stucco work and produced a long series

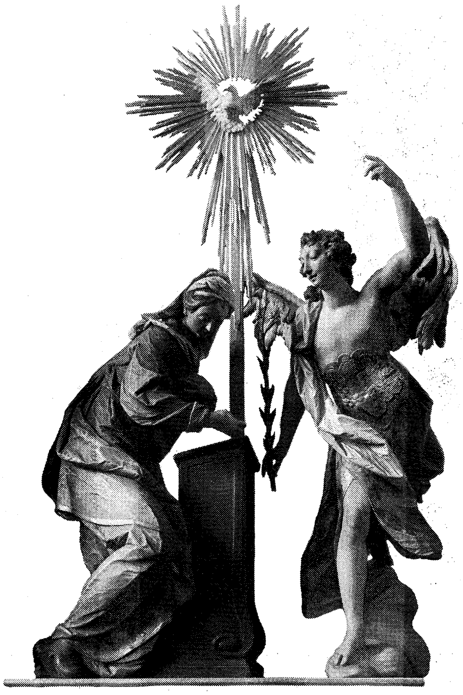


Figure 68: "The Annunciation," painted wood sculpture by Ignaz Günther, 1764. In the abbey church at Weyarn, Bavaria.

Bildarchiv Foto Marburg/Art Resource, NY

of masters, including Johann Georg Übelherr and Joseph Anton Feuchtmayer, whose masterpieces are the Rococo figures at Birnau on Lake Constance. The sculptor Christian Wenzinger worked at Freiburg im Breisgau in relative isolation, but his softly modelled figures have a delicacy that recalls the paintings of Boucher.

Until his death Johann Wolfgang van der Auvera was the most powerful personality in the field of sculpture in the area, but later Ferdinand Dietz at Bamberg pursued an increasingly individual Rococo style that often parodied the growing taste for Neoclassicism. Prussian Rococo sculpture was less distinguished, though the decorations of Johann August Nahl are among the most imaginative in Germany.

Austrian sculpture of the later 18th century, as represented by Balthasar Ferdinand Moll, inclined more toward a realistic Rococo style than to the Classicism of Donner; and, although the strange, neurotic genius Franz Xavier Messerschmidt began in this style, at the end of his career he produced a startling series of grimacing heads when he lived as a recluse in Bratislava.

Russia. The Baroque style as it was imported to Russia from western Europe by the imperial court never amounted to what might properly be termed a Russian Baroque period. A great influx of Western influence during this period, especially under the sponsorship of Peter the Great, did, however, dispel the predominance of Byzantine ideas and forms. The brilliant Baroque busts of Bartolomeo Carlo Rastrelli the Younger established during the early 18th century a distinguished tradition of Russian portrait sculpture that was maintained by Fedot Shubin. The parks and gardens of the Rococo palaces of the empress Elizabeth were adorned with sculpture, but the work was done almost exclusively by Italians and Frenchmen commissioned for the task.

Latin America. With the coming of Europeans to Central and South America, indigenous symbolism and sculptural forms blended with Renaissance realism, Baroque elegance, and subsequent stylistic currents. Indian traits appeared in such European-introduced sculptural forms as the stone crosses that were erected in churchyards; statues, whether by European sculpture or aboriginal pupils, depicted Jesus, the Virgin Mary, saints, and occasionally an earthly benefactor of the church. Materials were of wood, plant fibre pulp coated with canvas and gesso, or plaster. The statues often had real costumes and hair, glass eyes and teeth, and extremely realistic flesh—bloody, bruised, and torn—with taut muscles and distended veins. Gold halos or crowns were added and costume textures were imitated by the gold-leaf-and-paint estofado technique. Many of these were undoubtedly inspired by paintings brought from Europe.

Few sculptors are known by name from the colonial period and fewer attributions are possible. At least a dozen individuals can be identified in Mexico in the 16th century, however, and twice that number in the 17th; the best known are José Cora of Puebla and his nephew Zacarias, and Gudiño of Querétaro. Many were both sculptors and architects, a necessity of the times. In the 18th century considerable artistic stimulus was provided by the Spanish-born Neoclassicist Manuel Tolsa, first director of the Academy in Mexico City, first to produce an equestrian statue in the New World (of Charles IV), and teacher of many sculptors of subsequent fame. The second most important artistic centre of the colonial era was Quito, Ecuador, which was known particularly for its decorative sculpture.

The sculpture is marginally less provincial than the paintings, and, for example, the choir stalls carved by Pedro de Noguera and his assistants for Lima cathedral (1624–26) are of distinguished quality. The Baroque tradition tended to last until well into the 19th century in sculptures such as the robust figures of Antônio Francisco Lisboa (*e.g.*, "O Aleijadinho," or "The Little Cripple"), the greatest sculptor that Brazil has produced.

(P.C.-B./J.Hud./J.Hm./A.Vo./C.I.C./Ed.)

Neoclassical and Romantic sculpture

NEOCLASSICISM

The 18th-century arts movement known as Neoclassicism represents both a reaction against the last phase of the Baroque and, perhaps more importantly, a reflection of the burgeoning scientific interest in classical antiquity. Archaeological investigations of the classical Mediterranean world offered to the 18th-century cognoscenti compelling witness to the order and serenity of Classical art and provided a fitting backdrop to the Enlightenment and the Age of Reason. Newly discovered antique forms and themes were quick to find new expression.

The successful excavations contributed to the rapid growth of collections of antique sculptures. Foreign visitors to Italy exported countless marbles to all parts of Europe or employed agents to build up their collections. The accessibility of the sculpture of antiquity, in museums and private houses and also through engravings and plaster casts, had a far-reaching formative influence on 18th-century painting and sculpture. The great majority of ancient sculptures collected were Roman, although many of them were copied from Greek originals and were believed to be Greek.

In the writing of Johann Joachim Winckelmann, the great German historian of ancient art, Greek art had been considered immeasurably superior to Roman. It is curious, however, how little positive influence the marbles that Lord Elgin took to England from the Parthenon in Athens had on sculpture in western Europe, although they had a great influence on scholars. The ideals of Neoclassical sculpture—its emphasis on clarity of contour, on the plain ground, on not rivalling painting either in the imitation of aerial or linear perspective in relief or of flying hair and fluttering drapery in freestanding figures—were chiefly inspired by theory and by Roman neo-Attic works, or indeed by Roman pseudo-Archaic art. The latter class of art exerted an influence on John Flaxman, who was enormously admired for the severe style of his engravings and relief carvings.

"Decorum" and idealization. Academic theorists, especially those of France and Italy during the 17th century, argued that the costume, details, and setting of a work be as accurate as possible when representing a period and place in the historical past. The 18th century and, in particular, the Neoclassicists inherited this theory of "decorum" and, enabled by all the newly available archae-

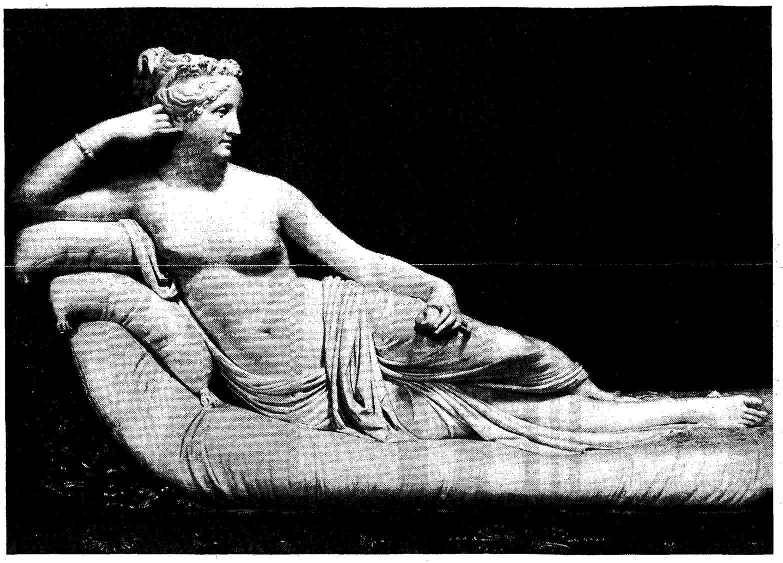


Figure 69: "Paolina Borghese as Venus Victrix," marble sculpture by Antonio Canova, 1805–07. In the Borghese Gallery, Rome. 1.60 m × 2.00 m.

Alinari—Art Resource/EB Inc.

ological evidence, implemented it more fully than had any of their precursors.

Question of the use of Classical or contemporary dress

A series of monuments to 18th- and early 19th-century generals and admirals of the Napoleonic Wars in St. Paul's Cathedral and Westminster Abbey demonstrate an important Neoclassical problem: whether a hero or famous person should be portrayed in Classical or contemporary costume. Many sculptors varied between showing the figures in uniform and showing them completely naked. The concept of the modern hero in antique dress belongs to the tradition of academic theory, exemplified by the English painter Sir Joshua Reynolds in one of his Royal Academy *Discourses:* "The desire for transmitting to posterity the shape of modern dress must be acknowledged to be purchased at a prodigious price, even the price of everything that is valuable in art." Even the living hero could be idealized completely naked, as in two colossal standing figures of Napoleon (1808–11; Apsley House, London, and Brera, Milan) by the Italian sculptor Antonio Canova. One of the most famous of Neoclassical sculptures is Canova's "Paolina Borghese as Venus Victrix" (1805–07; Borghese Gallery, Rome). She is shown naked, lightly draped, and reclining sensuously on a couch, both a charming contemporary portrait and an idealized antique Venus (Figure 69).

By courtesy of the Thorvaldsen Museum, Copenhagen

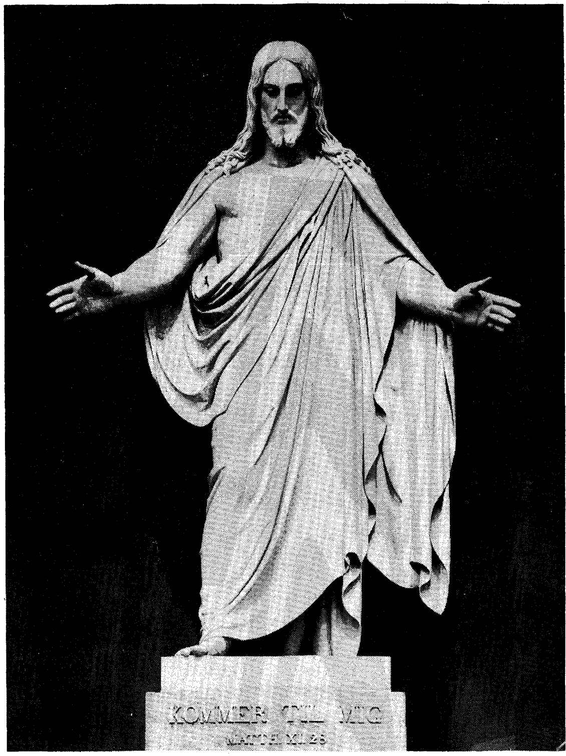


Figure 70: "Christ" from "Christ, John the Baptist and the Apostles," marble statue by Bertel Thorvaldsen, 1821. In the Church of Our Lady, Copenhagen. Height 3.36 m.

Relation to the Baroque and the Rococo. Classical academic theories circulating in the Renaissance, and especially in the 17th century, favoured the antique and those artists who followed in this tradition. The artists praised included Raphael, Michelangelo, Giulio Romano, and Annibale Carracci. The slightly later generation of writers added the name of the French painter Nicolas Poussin to the list. The exuberance and "fury" of the Baroque must be avoided, it was argued, because they led to "barbarous" and "wicked" works. Continuing in this tradition, Winckelmann, for example, argued that the Italian Baroque sculptor and architect Bernini had been "misled" by following nature.

Anti-Baroque feeling

Such hostility to Baroque works, however, did not immediately eradicate their influence on 18th-century artists, as can be seen, for example, in an early work by Canova, "Daedalus and Icarus" (1779; Museo Civico Correr, Venice), executed before he had been to Rome. In Canova's tomb of Pope Clement XIV (1784–87; SS. Apostoli, Rome), the Pope, seated on a throne above a sarcophagus, is treated in a dramatically realistic style with hand raised in a forceful gesture reminiscent of papal tombs of the 17th century.

Rococo influence

Although the Neoclassical artists and writers expressed contempt for what they regarded as the frivolous aspect of the Rococo, there is a strong influence of French Rococo on the early style of some of the Neoclassical sculptors. Étienne-Maurice Falconet, Flaxman, and Canova all started to carve and model with Rococo tendencies, which were then gradually transformed into more Classical elements.

Hostile critics of Neoclassical sculpture have tended to compare such works to "a valley of dry bones." Some artists misunderstood the advocacy of Winckelmann and his school to imitate ancient art. Winckelmann meant—as did 17th-century theorists before him, and writers such as Shaftesbury and Jonathan Richardson, who influenced him considerably—imitation to be a means of discovering ideal beauty and conveying the spirit of the original. He did not advocate servile copying of the antique. Unfortunately, spiritless copies were made, and these led to a proliferation of the Neoclassical style and its classification as "frigid." In sculpture some of the important commissions

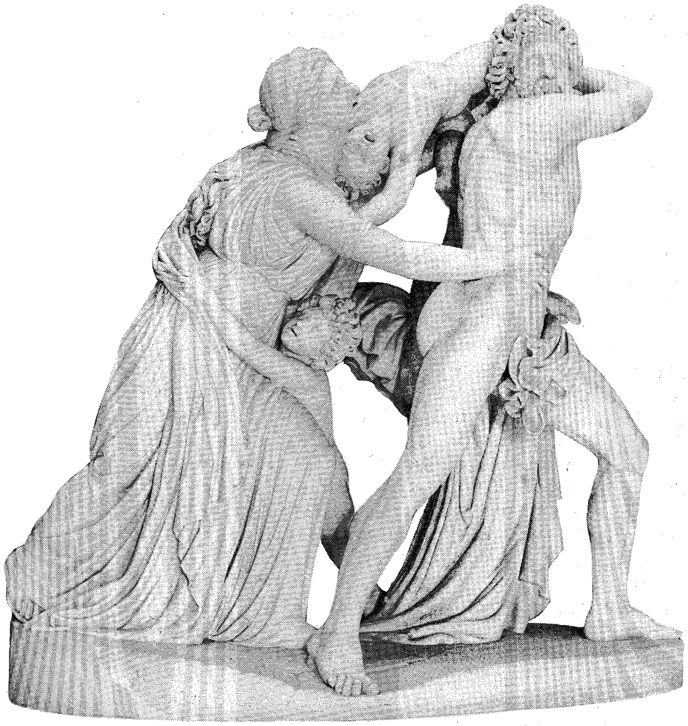


Figure 71: "Fury of Athamas," marble sculpture by John Flaxman, 1790–92. In Ickworth, Suffolk.

A.F. Kersting

regrettably resulted in this lifeless concept of Neoclassicism. Among the examples are large marbles of Christ (Figure 70), John the Baptist, and the Apostles by the Danish sculptor Bertel Thorvaldsen in the Church of Our Lady, Copenhagen (1821–27 and 1842). Thorvaldsen's marbles, unlike Canova's, lose little when seen only as plaster casts, and indeed the surface of the sculpture was deliberately left neutral and the act of carving left to others.

Classical serenity and restraint

Gestures and emotions in Neoclassical works are usually restrained. In bacchanalian scenes the gaiety is held in check, never bursting into exuberance. In a tragic scene, Andromache does not shed a tear as she mourns the death of Hector. When Flaxman did attempt terror, as in the marble "Fury of Athamas" (1790–92; Ickworth, Suffolk), the violence is forced and unconvincing (Figure 71). Indeed, there hardly exists in any Neoclassical sculptor's work a convincing image of rage. The concept of antique calmness permeated European art. Canova with his "Hercules and Lichas" (1796; Galleria Nazionale d'Arte Moderna, Rome) produced a large marble of exaggerated expression beyond his normal range and to some extent beyond his abilities. Like Flaxman, he was far more successful when carving images of delicacy, bordering on charm.

Prominent early British Neoclassicist sculptors included John Wilton, Joseph Nollekens, John Bacon the Elder, John Deare, and Christopher Hewetson, the last two working mostly in Rome. The leading artist of the younger generation was John Flaxman, professor of sculpture at the Royal Academy and one of the few British artists of the period with an international reputation. The last generation of Neoclassicists included the sculptors Sir Richard Westmacott, John Bacon the Younger, Sir Francis Chantrey, Edward Hodges Baily, John Gibson, and William Behnes.

While Neoclassicism in France was dominated by painting and architecture, the movement did find a number of notable exponents in sculpture. These included Claude Michel, called Clodion, creator of many small Classical figures, especially nymphs; Augustin Pajou; and Pierre Julien. Pigalle's pupil Jean-Antoine Houdon was the most famous 18th-century French sculptor, producing many Classical figures and contemporary portraits in the manner of antique busts. Other contemporary sculptors included Louis-Simon Boizot and Étienne-Maurice Falconet, who was director of sculpture at the Sèvres factory. The slightly younger generation included the sculptors Joseph Chinard, Joseph-Charles Marin, Antoine-Denis Chaudet, and Baron François-Joseph Bosio. The early sculpture of Ingres's well-known contemporary François Rude was Neoclassical.

Important among central European sculptors early in the period was Johann Heinrich von Dannecker. Subsequent Neoclassicists included Johann Gottfried Schadow, who was also a painter but is better known as a sculptor; his pupil, the sculptor Christian Friedrich Tieck; the painter and sculptor Martin von Wagner; and the sculptor Christian Daniel Rauch.

Canova

The most important Italian Neoclassicist was Antonio Canova, the leading sculptor, indeed by far the most famous artist of any sort, in Europe by the end of the 18th century. Canova's position in the following 20 years may be compared only with that enjoyed by Bernini in the 17th century. The differences between their careers, however, are of great importance. Only at the commencement of his career did Bernini carve gallery sculpture for princely collectors, but the majority of Canova's works belong to this category. Both artists remained resident for most of their life in Rome, but whereas Bernini was controlled by the popes and only rarely permitted to work for foreign potentates, Canova's principal patrons were foreigners, and he supplied sculpture to all the courts of Europe. A fine sculptor of varying styles, including austere, sentimental, and horrific, Canova produced an extensive body of work that includes Classical groups and friezes, tombs, and portraits, many in antique dress. He was also a painter, regrettably bad. His pupil and collaborator, Antonio d'Este, is one of the more interesting of the lesser Italian Neoclassical sculptors. Other Neoclassical sculptors in Rome included Giuseppe Angelini, best known for the tomb of the etcher and architect Giambattista Piranesi in the church of Sta. Maria del Priorato, Rome.

In Milan, Camillo Pacetti directed the sculptural decoration of the Arco della Pace. The work of Gaetano Monti, born in Ravenna, can be seen in many northern Italian churches. The Tuscan sculptor Lorenzo Bartolini executed some important Napoleonic commissions. The "Charity" (Pitti Palace, Florence) is one of the more famous examples of his later Neoclassicism. It should be noted, however, that he did not see himself as a Neoclassical artist and that he challenged the idealism that was favoured by Canova and his followers.

The Swede Johan Tobias Sergel, court sculptor to the Swedish king Gustav III, and the Dane Bertel Thorvaldsen, who lived most of his life in Rome, were among the best known Neoclassical sculptors in Europe. Thorvaldsen was the chief rival to Canova and eventually replaced him in critical favour. His work was more severe, sometimes even archaizing in character, and his religious sculpture, most notably his great figure of Christ in the Church of Our Lady in Copenhagen, exhibits a deliberately chilling, sublime style that still awaits sympathetic reassessment. Among his more notable pupils was the Swedish sculptor Johan Byström.

The principal Neoclassicists in Spain were the painter José de Madrazo y Agudo and the sculptor José Alvarez de Pereira y Cubero.

Both leading Russian Neoclassicists were sculptors. Ivan Petrovich Martos studied under Mengs, Thorvaldsen, and Batoni in Rome and became a director of the St. Petersburg Academy. His best works are tombs. Mikhail Kozlovskij contributed to the decoration of the throne room at Pavlovsk.

Apart from the painter Benjamin West, who worked almost entirely in London, the leading Neoclassicists among American artists were sculptors. William Rush produced standing Classical figures including those formerly decorating a waterworks in Philadelphia (now in the Pennsylvania Academy of Fine Arts). In the middle years of the 19th century there came into prominence four sculptors: Horatio Greenough, who executed several government commissions in Washington, D.C.; Hiram Powers, known particularly for his portrait busts; Thomas Crawford, who did monumental sculpture; and William Wetmore Story,

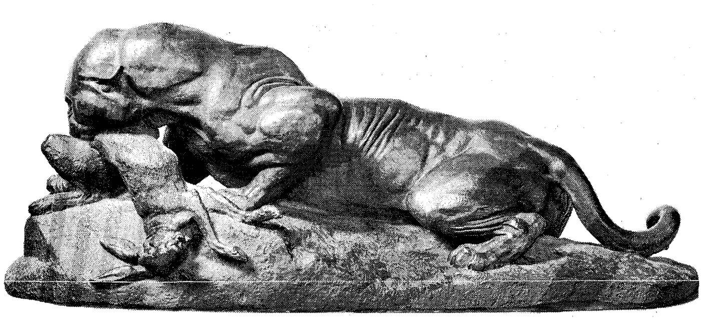


Figure 72: "Jaguar Devouring a Hare," bronze sculpture by Antoine-Louis Barye, 1850–51. In the Louvre, Paris. 41.9 cm × 95.2 cm.

Alinari—Art Resource/EB Inc.

who lived and worked in Rome, where he was associated with several other prominent 19th-century Americans.

(D.I.)

19TH-CENTURY SCULPTURE

In the 19th century sculptors throughout the Western world were affected in an unprecedented way by the great public annual exhibitions organized by the Academies. Great patrons at court or among the nobility could still play a very important part in making an artist's reputation, but publicity from these exhibitions was crucial. Among examples of sculptures that attracted sensational publicity of this sort are François Rude's "Neapolitan Fisherboy" (1834; Louvre), Hiram Powers' "Greek Slave" (1843), Auguste Clésinger's "Woman Bitten by a Snake" (1847; Louvre), and Randolf Rogers' "Nydia the Blind Girl" (1858).

In all these sculptures except the last the subject is more or less nude. In all except the first there is a strong narrative interest. In these respects they resemble the prize pieces set by the French Royal Academy of Painting and Sculpture and by its numerous imitators. Unlike those prize pieces, however, these works drew for their subjects not upon Greek or Roman mythology or history: Nydia is a Roman girl but taken from a modern novel about Pompeii, and the Greek slave is a contemporary Christian girl taken captive by the Turks. The old clichés about "academic" sculpture in the 19th century are hopelessly inadequate. The Academies in their educational program often encouraged a heroic but restrained Neoclassicism—their exhibitions, on the other hand, encouraged an appeal to novelty, to sentiment, and to sensationalism (often of an unfortunate kind), involving subjects from modern life and modern literature.

The exhibition piece was often a plaster cast of the original clay model. Several versions in marble or bronze were then made if there was the demand. These would be acquired for the sculpture galleries, conservatories, or gardens of great collectors, as well as for museums, which, for the first time, included collections of modern art. In reduced form they might also make an appearance amid the crowded furnishings of fashionable drawing rooms. Upon the chimneypiece perhaps some miniature scene of jungle violence modelled by Barye and cast in bronze might be displayed, while behind the ferns a marble nude would shrink in vain from male scrutiny.

The proliferation of domestic sculpture was made possible by a series of technical innovations chiefly associated with Paris. Improved reducing machines greatly facilitated the half-size replication of exhibition pieces, and the reproduction of such works on a still smaller scale as bronze statuettes; new methods of sand-casting meant that these bronzes were also available in larger editions and at a lower cost. The reproduction of terra-cotta sculpture also thrived in Paris as it had done in the late 18th century; busts of men of letters and women of fashion, together with groups of seductive nymphs, were always the most popular subjects. The miniature sculptures (often also reproductions of larger works) in biscuit porcelain, which had also been produced in 18th-century Paris, also continued to be popular in England for a while, as well as France.

Exalted notions of the artist's role, inculcated by the Academies and dramatized by Romantic literature, did little to encourage sculptors to involve themselves with what was often described as "mere" ornament. Mechanical methods—more and more sophisticated machinery for turning and pointing, as well as reducing machinery and novel techniques of casting—were often employed with great success. This resulted, however, not only in more bad sculpture than before but also in more badly carved and cast ornament in architecture, furniture, and metalwork. In Paris, however, the fertile genius of Albert Carrier-Belleuse particularly excelled in devising such objects as gasoliers supported by pretty girls in a luxurious style that combined elements from the art of the 16th, 17th, and 18th centuries. In England, Alfred Stevens, inspired by the versatility of the Italian Renaissance, was happy to devote himself to the design of cutlery and fire grates, and, at the end of the century, Alfred Gilbert, creator of the most remarkable metropolitan fountain since the Renaissance (the Eros in Piccadilly Circus), also became the first sculptor of the foremost rank since Cellini to devote himself wholeheartedly to the art of the goldsmith.

Perhaps the least successful aspect of 19th-century sculpture was the large-scale relief panels and pedimental ornaments and niche stances on churches and public buildings—the individual styles encouraged by the exhibition were inappropriate, and traditional styles tended to be artificially resurrected. The subject matter was often selected for negative purposes—to avoid offense, to seem impressive, to fill gaps. The unsuitability of this sort of task for the artist with a romantic sense of independence is obvious, and the situation did once arise, in the case of David d'Angers, of a sculptor's choosing his own program for one of the great public buildings in Paris (the Panthéon) against the wishes of his patrons. This same sense of independence also made for difficult relations between sculptors and architects. The quarrels between the architect of the Paris Opéra and Jean-Baptiste Carpeaux were typical; what was atypical was the success of Carpeaux's festive high relief of nymphs in abandoned dance (completed in 1869).

Another type of public sculpture—the portrait, typically in bronze, erected in a town square or other public space—

DEVANEY STOCK PHOTOS

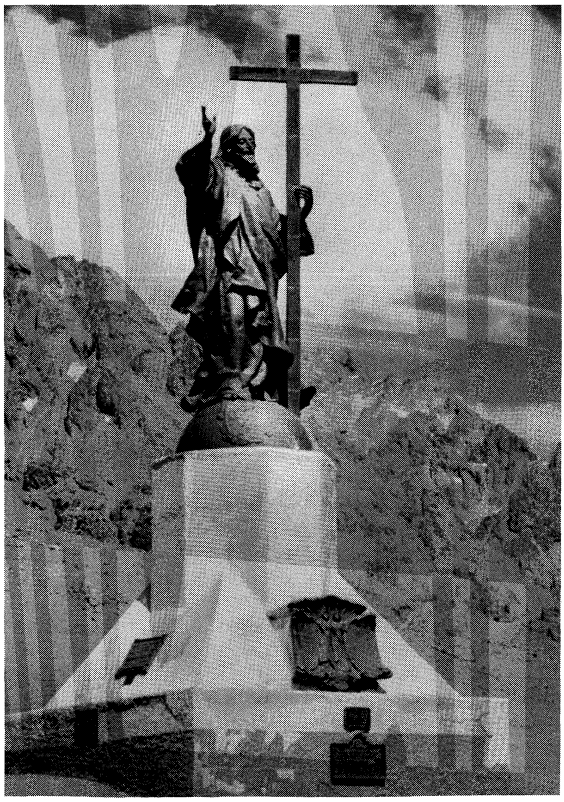


Figure 73: "Christ of the Andes" by Mateo Alonso, 1902. In the Uspallata Pass on the border between Argentina and Chile.

Commemorative portrait sculpture

flourished in the 19th century as it had not done since the first centuries AD. The first prominent sculptures of this sort commemorating nonroyal figures since antiquity seem to have appeared in Britain. The statues of Nelson by Sir Richard Westmacott erected in Liverpool and Birmingham soon after the subject's death were followed by statues of political heroes such as Fox and Pitt. By the end of the century, even relatively minor generals, philanthropists, or entrepreneurs were commemorated in this manner—almost invariably at the expense of public subscribers. The rest of Europe eventually followed this English example.

The young countries of the New World—the United States and later the republics of Latin America—commemorated with statues heroes whom they perceived as national saviours and founders. It may be that statues of Nelson excited as much patriotic sentiment as those of Washington or Bolívar, but Nelson could not embody the nation as the others did, nor certainly could any statue of a European monarch. For Europe national pride could best be promoted by an appeal to the past. Among the most remarkable public sculpture of the 19th century must certainly be counted Carlo Marochetti's "Duke Emmanuel Philibert" (1833, Turin) and Christian Daniel Rauch's "Frederick the Great" (1836–51, East Berlin) and the several statues of Joan of Arc in France. These were works of not simply historical but also topical and political significance, as indeed was the colossal "Christ of the Andes" by Mateo Alonso erected in 1902 on the border of Chile and Argentina. Abstractions were also endowed with a more urgent ideological content than in former centuries. In France, at least in the great "Triumph of the Republic" by Jules Dalou (unveiled in 1899 in the Place de la Nation), these could be animated with genuine passions. This is not true of the Statue of Liberty in New York City, which has nonetheless made an impact on the popular imagination.

Funeral sculpture

In the 19th century, funeral sculpture was as completely revolutionized as public sculpture. Whereas previously it had only really been in England that a large section of the wealthier classes had enjoyed the privilege of erecting substantial sculptured memorials, the opening up of large landscaped municipal cemeteries made this possible elsewhere. These cemeteries, of which the finest examples are in Paris and in Italy, were free from ecclesiastical censorship, and new themes quickly developed that were appropriate for an age of doubt and of desperate faith. The sentimentality and sensationalism of the annual exhibition were found here also, and so too was much exhibitionist virtuosity devoted to depicting the veiled faces and figures of ascending souls and their androgynous angelic escorts, as well as to recording bourgeois haberdashery.

This virtuosity is largely associated with Italian sculpture; and in a sense the Italians continued to dominate sculpture throughout the Western world after the death of Canova, by supplying the skilled carvers who were everywhere employed to translate into marble ideas worked out in clay. The greatest sculptors of the 19th century tended to play a smaller part than any of their predecessors in the actual carving, and the most vital sculpture of the period is preeminently plastic: when one thinks of the broken surfaces of the portrait busts by Carpeaux, for example, or of the precarious balances, open forms, and eloquent contours of Gilbert's statuettes, one thinks of wax and clay.

(N.B.P.)

Modern sculpture

19TH-CENTURY BEGINNINGS

The origins of modern art are usually traced to the mid-19th-century rejection of Academic tradition in subject matter and style by certain artists and critics. Painters of the Impressionist school that emerged in France in the late 1860s sought to free painting from the tyranny of the subject and to explore the intrinsic qualities of colour, brushwork, and form. This expansive notion of visual rendering had revolutionary effects on sculpture as well. The French sculptor Auguste Rodin found in it a new basis for life modelling and thus restored to the art a stylistic integrity that it had hardly possessed for more than two centuries.

Auguste Rodin

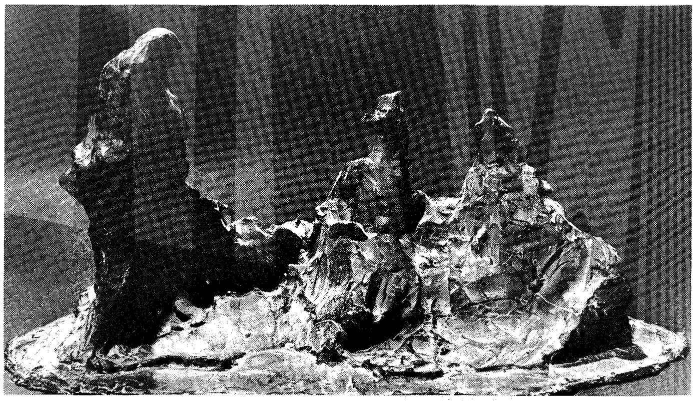


Figure 74: "Conversation in a Garden," wax-covered plaster sculpture by Medardo Rosso, 1893. In the Galleria Nazionale d'Arte Moderna, Rome. Height 43.2 cm.

By courtesy of the Galleria Nazionale d'Arte Moderna, Rome

Rodin's highly naturalistic early work, "The Age of Bronze" (1877), is effective because the banal studio pose of a man leaning on a staff produced an unconventional and expressive gesture when the staff was removed. From Honoré Daumier, Rodin had learned the bold modelling of surfaces that are emotive rather than literal; the statue is only a rough approximation that avoids the definitive finish of earlier sculpture and remains in a state of becoming. Eventually, Rodin even worked with mere fragments such as broken torsos, and he enormously enlarged the range of figure composition. The mass, until then the principal vehicle of sculptural composition, was explosively opened by these methods; in contrast to earlier sculpture, which depended on the interplay of solid and void, Rodin's works are fused with the surrounding space. These methods evolved in his many works, such as "Adam" (1880), "Eve" (1881), and others, originally conceived as a part of the masterpiece of modern sculpture, "The Gates of Hell," undertaken by Rodin in 1880 and never really completed. It was inevitable that the translucent nature of the marble surface should engage the attention of Rodin, and even though he always prepared the models in clay and left the execution in stone to assistants, such marbles as "The Kiss" (1885), when properly exhibited with light partly from the rear, appear to glow with the incandescence of their passionate intensity.

(J.Hud./J.Hm.)

Although the art of Rodin appears conservative in comparison to the painting of the time, in that he continued to use literary themes while painting did not, the new style that he evolved did much to revive sculpture's significance as an expressive medium, and his importance to 20th-century sculpture can hardly be overestimated. His fresh search and revelation of the basic movements of modern life had a profound influence on the generation of European sculptors who followed him.

Among Rodin's contemporaries, Edgar Degas, whose sculpture, begun in the 1880s, was an intimate study of movement and light, in several respects predicts 20th-century developments. Rodin's Italian counterpart, Medardo Rosso, lived in Paris during the 1880s; his work was known and owned by Rodin (Figure 74). Less gifted than Rodin but interested in the same problems, Rosso used wax in such a way that light was suffused through sensitively modelled portraits, and labile forms were created to express the flux that he felt was a condition of modern life. In Italy Rosso influenced Arturo Martini and through him Giacomo Manzù, Marino Marini, and Alberto Viani.

THE 20TH CENTURY

Rodin's successors

The ablest of Rodin's many pupils were Émile-Antoine Bourdelle and Charles Despiau. Bourdelle's "Hèraklès Archer" (1910) is an attempt to continue Rodin's active postures; but the results are melodramatic, and the forms are heavy and less sensitively modelled. Despiau, who

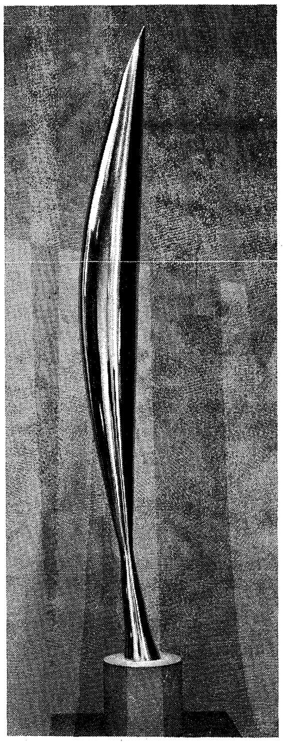


Figure 75: "Bird in Space," polished bronze sculpture by Constantin Brancusi, 1928? In the Museum of Modern Art, New York City. Height 1.37 m.

Collection, The Museum of Modern Art, New York, given anonymously

was director of Rodin's shop from 1907 to 1914, also responded to the interest in Classicism; his best work, "Girl from the Landes" (1904), was a balance of individual traits in the Rodin tradition, combined with graceful poses and well-rounded forms.

Two of the many other young sculptors attracted to Paris by Rodin's fame were Wilhelm Lehmbruck and Constantin Brancusi. Lehmbruck's early work has the soft modelling by touches of clay characteristic of the time, as in his "Mother and Child" (1907) and "Bust of a Woman" (1910). Brancusi's "Sleeping Muse" (1908) and the small "Bust of a Boy with Head Inclined" (1907) reflect Rodin's later interests in the expressiveness of modelling as opposed to strenuous gesture. Pablo Picasso and Henri Matisse were also early disciples of Rodin, as was Jacob Epstein, particularly in his naturalistic and psychologically incisive portraits.

Collection, The Museum of Modern Art, New York, Lillie P. Bliss Bequest

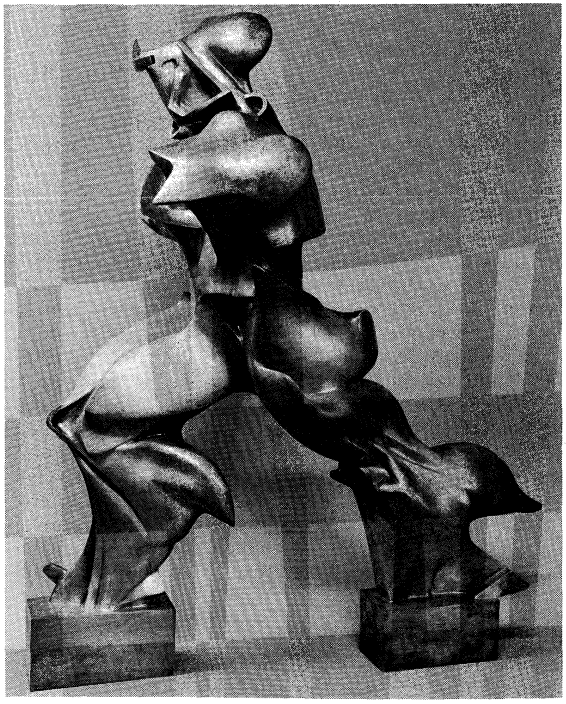


Figure 76: "Unique Forms of Continuity in Space," bronze sculpture by Umberto Boccioni, 1913. In the Museum of Modern Art, New York City. Height 1.1 m.

Avant-garde sculpture (1909–20). In the second decade of the 20th century the tradition of body rendering extending from the Renaissance to Rodin was shattered, and the Cubists, Brancusi, and the Constructivists emerged as the most influential forces. Cubism, with its compositions of imagined rather than observed forms and relationships, had a similarly marked influence.

One of the first examples of the revolutionary sculpture is Picasso's "Woman's Head" (1909). The sculptor no longer relied upon traditional methods of sculpture or upon his sensory experience of the body; what was given to his outward senses of sight and touch was dominated by strong conceptualizing. The changed and forceful appearance of the head derives from the use of angular planar volumes joined in a new syntax independent of anatomy. In contrast to traditional portraiture, the eyes and mouth are less expressive than the forehead, cheeks, nose, and hair. Matisse's head of "Jeanette" (1910–11) also partakes of a personal reproportioning that gives a new vitality to the less mobile areas of the face. Likewise influenced by the Cubists' manipulation of their subject matter, Alexander Archipenko in his "Woman Combing Her Hair" (1915) rendered the body by means of concavities rather than convexities and replaced the solid head by its silhouette within which there is only space.

Brancusi's "Bird in Space"

Brancusi also abandoned Rodin's rhetoric and reduced the body to its mystical inner core. His "Kiss" (1908), with its two blocklike figures joined in symbolic embrace, has a concentration of expression comparable to that of primitive art but lacking its spiritualistic power. In this and subsequent works Brancusi favoured hard materials and surfaces as well as self-enclosed volumes that often impart an introverted character to his subjects. His bronze "Bird in Space" became a *cause célèbre* in the 1920s when U.S. customs refused to admit it duty free as a work of art (Figure 75).

Raymond Duchamp-Villon began as a follower of Rodin, but his portrait head "Baudelaire" (1911) contrasts with that by his predecessor in its more radical departure from the flesh; the somewhat squared-off head is molded by clear, hard volumes. His famous "Horse" (1914), a coiled, vaguely mechanical form bearing little resemblance to the animal itself, suggests metaphorically the horsepower of locomotive drive shafts and, by extension, the mechanization of modern life. Duchamp-Villon may have been influenced by Umberto Boccioni, one of the major figures in the Italian Futurist movement and a sculptor who epitomized the Futurist love of force and energy deriving from the machine. In "Unique Forms of Continuity in Space" (Figure 76) and "Head + House + Light" (1911), he carried out his theories that the sculptor should model objects as they interact with their environment, thus revealing the dynamic essence of reality.

Jacques Lipchitz came to Cubism later than Archipenko and Duchamp-Villon, but after mastering its meaning he produced superior sculpture. In 1913, after several years of conservative training, he made a number of small bronzes experimenting with the compass curve and angular planes. They reveal an understanding of the Cubist reconstitution of the bodies in an impersonal quasi-geometric armature over which the artist exercised complete autonomy. Continuing to work in this fashion, he produced "Man with a Guitar" (Figure 77), and "Standing Figure" (1915), in which voids are introduced, while in the early 1920s he developed freer forms more consistently based on curves.

Lehmbruck's mature style emerged in the "Kneeling Woman" (1911) and "Standing Youth" (1913), in which his gothicized, elongated bodies with their angular posturings and appearance of growing from the earth give

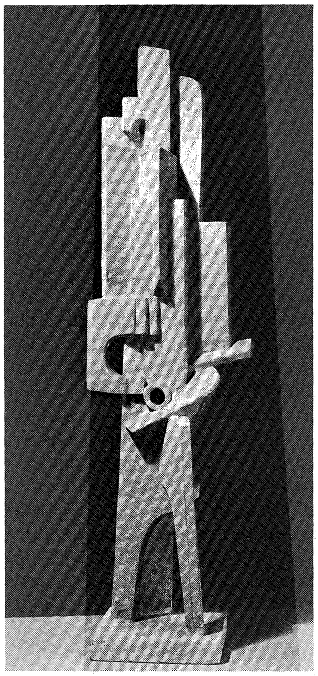


Figure 77: "Man with a Guitar," limestone sculpture by Jacques Lipchitz, 1916. In the Museum of Modern Art, New York City. Height 97.2 cm.

Collection, The Museum of Modern Art, New York, Mrs. Simon Guggenheim Fund (by exchange)

expression to his notions of modern heroism. In contrast to this spiritualized view is his "The Fallen" (1915–16), intended as a compassionate memorial for friends lost in the war.

Constructivism and Dada. Between 1912 and 1914 there emerged an antisculptural movement, called Constructivism, that attacked the false seriousness and hollow moral ideals of academic art. The movement began with the relief fabrications of Vladimir Tatlin in 1913. The Constructivists and their sympathizers preferred industrially manufactured materials, such as plastics, glass, iron, and steel, to marble and bronze. Their sculptures were not formed by carving, modelling, and casting but by twisting, cutting, welding, or literally constructing: thus the name Constructivism.

Sculpture as invention

Unlike traditional figural representation, the Constructivists' sculpture denied mass as a plastic element and volume as an expression of space; for these principles they substituted geometry and mechanics. In the machine, where the Futurists saw violence, the Constructivists saw beauty. Like their sculptures, it was something invented; it could be elegant, light, or complex, and it demanded the ultimate in precision and calculation.

Seeking to express pure reality, with the veneer of accidental appearance stripped away, the Constructivists fabricated objects totally devoid of sentiment or literary association; Naum Gabo's work frequently resembled mathematical models, and several Constructivist sculptures, such as those by Kazimir Malevich and Georges Vantongerloo, have the appearance of architectural models. The Constructivists created, in effect, sculptural metaphors for the new world of science, industry, and production; their aesthetic principles are reflected in much of the furniture, architecture, and typography of the Bauhaus.

The Dadaist assemblage

A second important offshoot of the Cubist collage was the fantastic object or Dadaist assemblage. The Dadaist movement, while sharing Constructivism's iconoclastic vigour, opposed its insistence upon rationality. Dadaist assemblages were, as the name suggests, "assembled" from materials lying about in the studio, such as wood, cardboard, nails, wire, and paper; examples are Kurt Schwitters' "Rubbish Construction" (1921) and Marcel Duchamp's "Disturbed Balance" (1918). This art generally exalted the accidental, the spontaneous, and the impulsive, giving free play to associations. Its paroxysmal and negativist tenor led its subscribers into other directions, but Dadaism formed the basis of the imaginative sculpture that emerged in the later 1920s.

Conservative reaction (1920s). In the 1920s modern art underwent a reaction comparable to the changes experienced by society as a whole. In the postwar search for security, permanence, and order, the earlier insurgent art seemed to many to be antithetical to these ends, and certain avant-garde artists radically changed their art and thought. Lipchitz' portraits of "Gertrude Stein" (1920) and "Berthe Lipchitz" (1922) return volume and features to the head but not an intimacy of contact with the viewer. Tatlin and Alexander Rodchenko broke with the Constructivists around 1920. Jacob Epstein developed some of his finest naturalistic portraiture in this decade. Rudolph Belling abandoned the mechanization that had characterized his "Head" (1925) in favour of musculature and individual identity in his statue of "Max Schmeling" of 1929. Matisse's reclining nudes and the "Back" series of 1929 show less violently worked surfaces and more massive and obvious structuring.

Aristide Maillol continued refining his relaxed and uncomplicated female forms with their untroubled, stolid surfaces. In Germany, Georg Kolbe's "Standing Man and Woman" of 1931 seems a prelude to the Nazi health cult, and the serene but vacuous figures of Arno Breker, Karl Albiker, and Ernesto de Fiori were simply variations on a studio theme in praise of youth and body culture. In the United States adherents of the countermovement included William Zorach, Chaim Gross, Adolph Block, Paul Manship, and Wheeler Williams.

Sculpture of fantasy (1920–45). One trend of Surrealist or Fantasist sculpture of the late 1920s and the 1930s consisted of compositions made up of found objects, such as Meret Oppenheim's "Object, Fur Covered Cup" (1936). As with Dadaist fabrications, the unfamiliar conjunction of familiar objects in these assemblies was dictated by impulse and irrationality and could be summarized by Isidore Ducasse's often-quoted statement, "Beautiful . . . as the chance meeting on a dissecting table of a sewing machine with an umbrella."

Of greater artistic importance was the sculpture of a second group that included Alberto Giacometti, Jean Arp, Lipchitz, Henry Moore, Barbara Hepworth, Picasso, Julio González, and Alexander Calder. Although these sculptors were sometimes in sympathy with Surrealist objectives, their aesthetic and intellectual concerns prohibited a more consistent attachment. Their art, derived from visions, hallucinations, reverie, and memory, might best be called the sculpture of fantasy. Giacometti's "Palace at 4 A.M." (Figure 78), for example, interprets the artist's vision not in terms of the external public world but in an enigmatic,

By courtesy of the Museum of Modern Art, New York

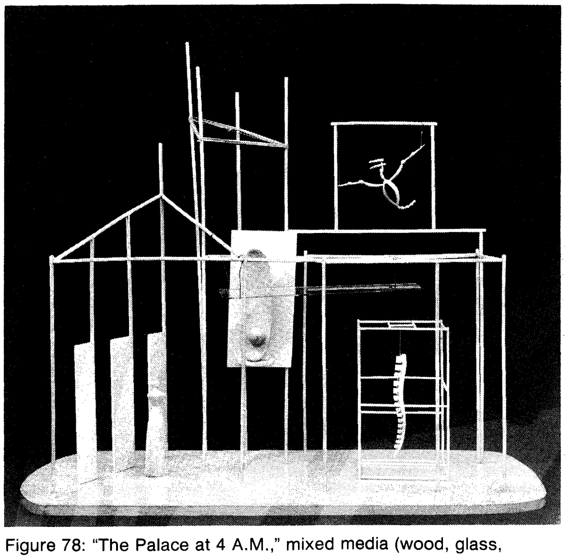


Figure 78: "The Palace at 4 A.M.," mixed media (wood, glass, wire, and string) by Alberto Giacometti, 1932–33. In the Museum of Modern Art, New York City. 63.5 × 71.8 × 40 cm.

private language. Moore's series of "Forms" suggest shapes in the process of forming under the influence of each other and the medium of space. The appeal of primitive and ancient ritual art to Moore, the element of surprise in children's toys for Calder, and the wellsprings of irrationality from which Arp and Giacometti drank were for these men the means by which wonder and the marvelous could be restored to sculpture. While their works are often violent transmutations of life, their objectives were peaceful, "... to inject into the vain and bestial world and its retinue, the machines, something peaceful and vegetative." ([Jean] Hans Arp, *On My Way,* Documents of Modern Art, vol. 6, p. 123, George Wittenborn, Inc., New York, 1948.)

Other sculpture (1920–45). The sculpture of Moore, Gaston Lachaise, and Henri Laurens during the 1920s and '30s included mature, ripe human bodies, erogenic images reminiscent of Hindu sculpture, appearing inflated with breath rather than supported by skeletal armatures. Lachaise's "Montagne" (1934–35) and Moore's reclining nudes of the '30s are identifications with earth, growth, vital rhythm, and silent power. Prior to Moore and the work of Archipenko, Boccioni, and Lipchitz, space had been a negative element in figure sculpture; in Moore's string sculptures and Lipchitz' transparencies of the 1920s, it became a prime element of design.

Moore's reclining nudes

Lipchitz' figure style of the late 1920s and '30s is inseparable from his emerging optimistic humanism. His concern with subject matter began with the ecstatic "Joy of Life" (1927). Thereafter his seminal themes were of love and security and assertive passionate acts that throw off the inertia of his Cubist figures. In the "Return of the Prodigal Son" (1931), for example, strong, facetted curvilinear volumes weave a pattern of emotional and aesthetic accord between parent and child.

The American sculptor John B. Flannagan rendered animal forms as well as the human figure in a simple, almost naive style. His interest in what he called the "profound subterranean urges of the human spirit in the whole dynamic life process, birth, growth, decay and death" (quoted in Carl Zigrosser, *Catalog for the Exhibition of the Sculpture of John B. Flannagan,* p. 8, The Museum of Modern Art, New York, 1942) resulted in "Head of a Child" (1935), "New One" (1935), "Not Yet" (1940), and "The Triumph of the Egg" (1941).

Somewhat more mystical are Brancusi's "Beginning of the World" (1924), "Fish" (1928–30), and "The Seal" (1936). As with Flannagan, the recurrent egg form in Brancusi's art symbolizes the mystery of life. Nature in motion is the subject of Alexander Calder's mobiles, such as "Lobster Trap and Fish Tail" (1939) and others suggesting the movement of leaves, trees, and snow. In the history of sculpture there is no more direct or poetic expression of nature's rhythm.

Developments after World War II. "The modern artist is the counterpart in our time of the alchemist-philosopher who once toiled over furnaces, alembics and crucibles, ostensibly to make gold, but who consciously entered the most profound levels of being, philosophizing over the melting and mixing of various ingredients" (Ibram Lassaw, quoted by Lawrence Campbell in *Art News,* p. 66, The Art Foundation Press, New York, March 1954). While work in the older mediums persisted, it was the welding, soldering, and cutting of metal that emerged after 1945 as an increasingly popular medium for sculpture. The technical and expressive potential of uncast metal sculpture was carried far beyond the earlier work of González and Picasso.

By courtesy of the trustees of the Tate Gallery, London

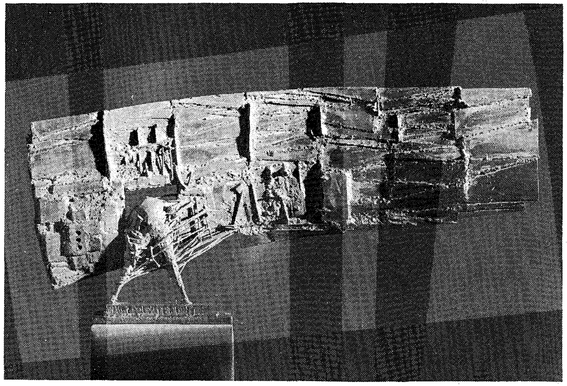


Figure 79: "The Man of Saint-Denis," by César (Baldaccini), 1958. In the Tate Gallery, London. Height 50.8 cm.

The appeal of metal is manifold. It is plentifully available from commercial supply houses; it is flexible and permanent; it allows the artist to work quickly; and it is relatively cheap compared to casting. Industrial metals also relate modern sculpture physically, aesthetically, and emotionally to its context in modern civilization. As the American sculptor David Smith has commented, "Possibly steel is so beautiful because of all the movement associated with it, its strength and functions. Yet it is also brutal, the rapist, the murderer and death-dealing giants are also its offspring" (quoted in Garola Giedion-Welcker, *Contemporary Sculpture,* Documents of Modern Art, vol. 12, p. 123, George Wittenborn, Inc., New York, 1955).

The technique of metal sculpture

The basic tool of the metal sculptor is the oxyacetylene torch, which achieves a maximum temperature of 6,500° F (3,600° C; the melting point of bronze is 2,000° F). The intensity and size of the flame can be varied by alternating torch tips. In the hands of a skilled artist the torch can cut or weld, harden or soften, colour and lighten or darken metal. Files, hammers, chisels, and jigs are also used in shaping the metal, worked either hot or cold. The sculptor may first construct a metal armature that he then proceeds to conceal or expose. He builds up his form with various metals and alloys, fusing or brazing them, and may expose parts or the whole to the chemical action of acids. This type of work requires constant control, and many sculptors work out and guard their own recipes.

Other sculptors such as Peter Agostini, George Spaventa, Peter Grippe, David Slivka, and Lipchitz, who were interested in bringing spontaneity, accident, and automatism into play, returned to the more labile media of wax and clay, with occasional cire-perdue casting, which permit a very direct projection of the artist's feelings. By the nature of the processes such work is usually on a small scale.

A number of artists brought new technique and content to the Dadaist form of the assemblage. Among the most important was the American Joseph Cornell, who combined printed matter and three-dimensional objects in his intimately sealed, often enigmatic "boxes."

Another modern phenomenon, seen particularly in Italy, France, and the United States, was the revival of relief sculpture and the execution of such works on a large scale, intended to stand alone rather than in conjunction with a building. Louise Nevelson, for example, typically employed boxes as container compartments in which she carefully disposed an assortment of forms and then painted them a uniform colour. In Europe the outstanding metal reliefs were those by Alberto Burri, Gio and Arnaldo Pomodoro, César (Figure 79), Zoltán Kemény, and Manuel Rivera.

New views of nature. Development of metal sculpture, particularly in the United States, led to fresh interpretations of the natural world. In the art of Richard Lippold and Ibram Lassaw, the search for essential structures took the form of qualitative analogies. Lippold's "Full Moon" (1949–50) and "Sun" (1953–56; commissioned by the Metropolitan Museum of Art, New York City, to hang in its room of Persian carpets) show an intuition of a basic regularity, precise order, and completeness that underlies the universe. Lassaw's comparable interest in astronomical phenomena inspired his "Planets" (1952) and "The Clouds of Magellan" (1953).

In contrast to the macrocosmic concern of these two artists were the interests of sculptors such as Raymond Jacobson, whose "Structure" (1955) derived from his study of honeycombs. Using three basic sizes, Jacobson constructed his sculpture of hollowed cubes emulating the modular, generally regular but slightly unpredictable formal quality of the honeycomb.

Isamu Noguchi's "Night Land" is one of the first pure landscapes in sculpture. David Smith's "Hudson River

Landscape" (1951), Theodore J. Roszak's "Recollections of the Southwest" (1948), Louise Bourgeois's "Night Garden" (1953), and Leo Amino's "Jungle" (1950) are later examples.

In the 1960s a number of sculptors, particularly in the United States, began to experiment with using the natural world as a kind of medium rather than a subject. Among the more notable examples were the American Robert Smithson, who frequently employed earth-moving equipment to alter natural sites, and the Bulgarian-born Christo, whose "wrappings" of both natural and man-made structures in synthetic cloth generated considerable controversy. The name environmental sculpture has come to denote such works, together with other sculptures that constitute self-contained environments.

The human figure. Since figural sculpture moved away from straightforward imitation, the human form has been subjected to an enormous variety of interpretations. The thin, vertical, Etruscan idol-like figures developed by Giacometti showed his repugnance toward rounded and smooth body surfaces or strong references to the flesh. His men and women do not exist in felicitous concert with others; each form is a secret sanctum, a maximum of being wrested from a minimum of material. Reg Butler's work (*e.g.,* "Woman Resting" [1951]) and that of David Hare ("Figure in a Window" [1955]) treat the body in terms of skeletal outlines. Butler's figures partake of nonhuman qualities and embody fantasies of an unsentimental and aggressive character; the difficulties and tensions of existence are measured out in taut wire armatures and constricting malleable bronze surfaces. Kenneth Armitage and Lynn Chadwick, two other British sculptors, make the clothing a direct extension of the figure, part of a total gesture. In his "Family Going for a Walk" (1953), for example, Armitage creates a fanciful screenlike figure recalling wind-whipped clothing on a wash line. Both Chadwick and Armitage transfer the burden of expression from human limbs and faces to the broad planes of the bulk of the sculpture. Chadwick's sculptures are often illusive hybrids suggesting alternately impotent De Chirico-like figures or animated geological forms.

Luciano Minguzzi admired the amply proportioned feminine form. Minguzzi's women (*e.g.,* "Woman Jumping Rope" [1954]) may exert themselves with a kind of playful abandon. Marini's women (*e.g.,* "Dancer" [1949]) enjoy a stately passivity, their quiescent postures permitting a contrapuntal focus on the graceful transition from the slender extremities to the large, compact, voluminous torso, with small, rich surface textures.

The segmented torso, popular with Arp, Laurens, and Picasso earlier, continued to be reinterpreted by Alberto Viani, Bernard Heiliger, Karl Hartung, and Raoul Hague. The emphasis of these sculptors was upon more subtle, sensuous joinings that created self-enclosing surfaces. Viani's work, for example, does not glorify body culture or suggest macrocosmic affinities as does an ideally proportioned Phidian figure; his torsos are seen in a private way, as in his "Nude" (1951), with its large body and golf ball-sized breasts.

A.C. Cooper Ltd.

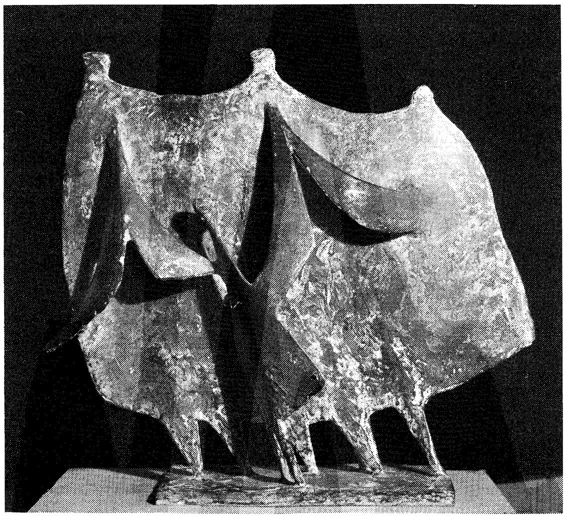


Figure 80: "Family Going for a Walk," bronze sculpture by Kenneth Armitage, 1953. In a private collection. Height 74 cm.

Among the most impressive figure sculptures made in the United States in the late 1950s were those by Seymour Lipton. Their large-scale, taut design and provocative interweaving of closed and open shapes restore qualities of mystery and the heroic to the human form.

The American George Segal emerged from the Pop movement of the 1950s and '60s as a major figurative sculptor. His plaster casts from live models, usually left white and indistinctly featured, are often situated in mundane settings of actual furniture or other objects.

The works of the French-born American artist Marisol contrast sharply with Segal's in their boxlike forms, onto which highly individualized features are usually painted. In the 1970s and '80s, Duane Hanson, another American, took Segal's live-model casting technique a step further with his startlingly naturalistic, fully pigmented cast fibreglass figures.

Archaizing, idol making, and religious sculpture. After World War II several sculptors became interested in the art of early Mediterranean civilizations. The result was a conscious archaizing of the human form with the intent of recapturing qualities of Cycladic idols, early Greek and Egyptian statuary, and some aspects of late Roman art.

Moore's admiration for archaic Greek sculpture produced "Draped Reclining Figure" (1952), which shows his return to the solid form and the suggestion of power and force by using drapery as a tense foil for the volumes that press against it. His "King and Queen" (1952–53) resulted from further excursions into the archaic Greek myth world.

The interest in recreating idols or totems was continued by Arp in his "Idol" (1950) and by Noguchi in his Stone Age-type sculptures for the Connecticut General Life Insurance Company (Hartford). By creating presences that elude rational definition, these artists restored to art its ancient aura of myth, mystery, and magic in an age that consistently disclaims their existence.

Modern religious sculpture

The argument that modern sculpture is inappropriate for religious requirements is disproved by works of Lipchitz, Lassaw, and Herbert Ferber. In keeping with the Jewish preference for nonfigural art, Ferber's ". . . and the bush was not consumed" (1951), commissioned by a synagogue in Millburn, New Jersey, comprises clusters of branches and boldly shaped weaving flames, invisibly suspended in a powerful and intimate vision that absorbs its viewers with its hypnotic rhythm. Lassaw's "Pillar of Fire," for the exterior of a synagogue in Springfield, Massachusetts, also has a mesmerizing pattern recalling the illusory images sometimes seen in flames. Lipchitz' sculpture of the "Virgin of Assy" (1948–54) was commissioned for the Catholic church at Assy, France.

Moreover, an increasing number of gifted sculptors are providing handsome liturgical objects and decorations, such as Harry Bertoia's shimmering reredos, Lipton's work for a synagogue in Tulsa, Oklahoma, and Roszak's sculptured spire for Kresge Chapel on the campus of the Massachusetts Institute of Technology, Cambridge.

Public and private memorials. After World War II there was a flood of public memorial sculpture, and in Europe especially many of the commissions were carried out by modern sculptors. A striking war memorial in Italy is Mirko Basaldella's gate for the monument to the Roman hostages killed in the Ardeatine Caves (1951). For its full effect the gate must be seen in connection with the rugged masonry wall to which it is attached. The gate was cast in metal and fashioned in a tangled, thicket-like pattern that suggests the painfully difficult passage from life to death for those who died in the caves.

Another imposing memorial is Ossip Zadkine's monument to the bombing of Rotterdam, a figure recoiling from the violence that descended from the sky. In Moore's "Warrior with a Shield" a soldier defiantly raises his shield and mutilated body toward the ill-starred heavens during the Battle of Britain. Epstein's public monument to "Social Consciousness" (1952–53), in Fairmount Park,

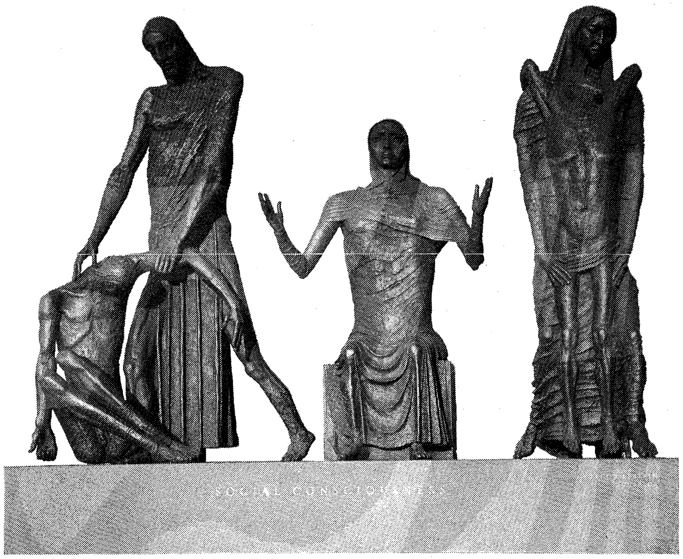


Figure 81: "Social Consciousness," bronze sculpture by Jacob Epstein, 1952–53. In Fairmount Park, Philadelphia. Height 3.9 m.

By courtesy of Philadelphia Museum of Art, Ellen Phillips Samuel Memorial, Fairmount Park Art Association

Philadelphia, treats the helplessness of those confronted with pressures over which they have no control. In contrast to the invulnerable champions of academic art, these sculptures image the hero in distress.

Modern sculpture and architecture

Other developments. Despite the rapid and exciting developments in both architecture and sculpture, the two have seldom been meaningfully and integrally united. The architecture of Le Corbusier, Frank Lloyd Wright, Pier Luigi Nervi, Ludwig Mies van der Rohe, and others occasionally shows strong sculptural qualities, but relatively rarely were their surfaces planned to receive sculpture. Freestanding sculptures such as those created by Gabo, Pevsner, De Rivera, Calder, and Noguchi have been used to provide intimacy and visual relief from the severity of the "cult of the cube" in architecture. The architectural firm of Skidmore, Owings & Merrill successfully used Bertoia's brilliant screens and Noguchi's sculptures and garden ideas; Roszak's "Eagle" for the American embassy in London and Moore's changeable reliefs on the London Time and Life Building held out hope for further thoughtful integration of the arts.

Also of great moment is the phenomenon of the sculptor-designer who has produced important changes in furniture and industrial design. Max Bill's school in Ulm, Germany, showed great promise. Playground facilities have been revolutionized by such designs as those made by Noguchi for Creative Playthings Inc. in the United States and the slides, hollowed forms, and climb apparatus of Egon Moeller-Nielson for parks in Stockholm. Noguchi, Moholy-Nagy, Bill, Bertoia, and many other modern artists contributed to the breakdown of the distinction between the object of utility and the work of art. Not since Gothic times has sculpture shown such promise of becoming an extensive and important part of human existence. (A.E.El./Ed.)

In Italy, traditional trends in sculpture are reflected in the brilliant accomplished modelling of Giacomo Manzù; Marino Marini, devoting himself almost entirely to the single theme of horse and rider, gave a bald realistic style an oddly apocalyptic force. The rough-hewn monumentality of the figures of the Austrian carver Fritz Wotruba is characteristic of this phase. Joannis Avramidis, also working in Vienna, turned figures into clusters of simplified formal echoes; the third sculptor of the Viennese group, Rudolf Hoflehner, who worked in iron, transformed them into symbolic presences. The segmental iron sculpture of the Spaniard Eduardo Chillida deals with a more limited and powerful range of forms.

Robert Rauschenberg in the United States sought to place his subtly calculated "combines" in the gap between reality and art, contrasting the significance of paint with the borrowed imagery and objects that are juxtaposed to it. Another American, Claes Oldenburg, began by reconstructing common things out of the random pictorial substance of Abstract Expressionism; his later reconstructions of the rigid furniture of life are tailored out of limp plastic sheeting, and the paradox oddly extends one's knowledge of the objective world.

In the reliefs of the Venezuelan Jesús Raphael Soto, the shifting paradoxes of vision are given a delicate order. Aside from this, the widespread work in kinetic mediums, such as that of Nicholus Takis, during the 1960s formed a separate genre, winking and shuddering on its own, most nearly linked to the Surrealist tradition.

Other sequels of the general rationalization and concentration of artistic means have been more fertile. In the hands of the U.S. painters Kenneth Noland and Frank Stella, painting discovered new shapes, both within the rectangular canvas and beyond it. The new value that was given to the painted plane did not benefit painting only. The British painter Richard Smith deployed it in three dimensions in painted constructions that re-create impressions of commercial packaging in terms of the spatial imagination of the arts. Sculpture, reequipped with colour, developed remarkably, and Anthony Caro led a group of British sculptors in exploration of spatial modulation and formal analogy. (La.G.)

BIBLIOGRAPHY

General: An excellent general history of world art is HUGH HONOUR and JOHN FLEMING, *A World History of Art* (1982; U.S. title, *The Visual Arts: A History*), which examines sculpture in relation to the other arts. H.W. JANSON, *History of Art* (1962; 2nd ed., 1977), is also recommended. Among books that discuss sculpture of many periods, RUTH BUTLER, *Western Sculpture: Definitions of Man* (1975), is unusually valuable. So, too, is F. DAVID MARTIN, *Sculpture and Enlivened Space* (1981). For the techniques of sculpture see W. VERHELST, *Sculpture: Tools, Materials, and Techniques* (1973); and RUDOLF WITTKOWER, *Sculpture* (1977). The making of bronze sculptures, omitted from the latter, is brilliantly elucidated by JENNIFER MONTAGU, *Bronzes* (1963, reissued 1972). ERWIN PANOFSKY, *Tomb Sculpture* (1964), traces from ancient Egypt to about 1800 some of the major themes of one very important class of Western sculpture.

Ancient Mediterranean: Sculpture in the early civilizations of southern Europe is seldom studied separately, but it is featured in the following general works: JOHN BOARDMAN, *Pre-Classical* (1967, reissued 1979); R.W. HUTCHINSON, *Prehistoric Crete* (1962); A. ARRIBAS, *The Iberians* (1964); N.K. SANDARS, *Prehistoric Art in Europe* (1968); and SPYRIDON MARINATOS, *Crete and Mycenae* (1960).

Greek, Hellenistic, Etruscan, and Roman art: An authoritative and comprehensive account of ancient Greek art (which, for the most part, means Greek sculpture) is MARTIN ROBERTSON, *A History of Greek Art* (1975). For a succinct introduction to sculpture only, see JOHN BARRON, *Introduction to Greek Sculpture* (1981, reissued 1984). For the Archaic period, G.M.A. RICHTER, *Archaic Greek Art Against Its Historical Background* (1949), is still valuable; for the so-called Classical period, BRUNILDE S. RIDGWAY, *Fifth Century Styles in Greek Sculpture* (1981), is a good detailed guide; and for the later periods, MARGARETE BIEBER, *The Sculpture of the Hellenistic Age*, 2nd rev. ed. (1981), is highly useful. For the ancient literature on art, see J.J. POLLITT, *The Art of Greece 1400–31 B.C.: Sources and Documents* (1965). Etruscan sculpture is best discussed in OTTO J. BRENDEL, *Etruscan Art* (1978). Sculpture features prominently in the most lively general books on Roman art: R. BIANCHI BANDINELLI, *Rome: The Centre of Power* (1970; originally published in Italian, 1969), and *Rome: The Late Empire* (1971); and RICHARD BRILLIANT, *Roman Art* (1974). Of more limited scope but great interest is JOCELYN M.C. TOYNBEE, *Art in Roman Britain* (1962). See also J.J. POLLITT, *The Art of Rome c. 753 BC–AD 337: Sources and Documents* (1966, reissued 1983).

Early Christian and early medieval: Good general surveys of the early Christian period that include some discussion of sculpture are ERNST KITZINGER, *Byzantine Art in the Making* (1977); JOHN BECKWITH, *The Art of Constantinople*, 2nd ed. (1968); ANDRÉ GRABAR, *The Beginnings of Christian Art: 200–395* (1967, originally published in French, 1966); STEVEN RUNCIMAN, *Byzantine Style and Civilization* (1975); and CYRIL A. MANGO, *The Art of the Byzantine Empire 312–1453: Sources and Documents* (1972). This last volume, together with ERNST KITZINGER, *Early Medieval Art* (1940; rev. ed., 1983), concerns also the early medieval period. Among more specialized studies

of sculpture in the early Christian period, JOHN BECKWITH, *Coptic Sculpture* (1963); and JOSEPH NATANSON, *Early Christian Ivories* (1953), should be mentioned. For general information on the early medieval period, see PETER LASKO, *Ars Sacra 800–1200* (1972); GEORGE HENDERSON, *Early Medieval* (1972); and GEORGE ZARNECKI, *Art of the Medieval World* (1975). Valuable studies specifically on sculpture include GEORGE H. CRICHTON, *Romanesque Sculpture in Italy* (1954); HERMANN LEISINGER, *Romanesque Bronzes* (1956); FRITZ SAXL, *English Sculptures of the Twelfth Century* (1954); and M.F. HEARN, *Romanesque Sculpture: The Revival of Monumental Stone Sculpture in the Eleventh and Twelfth Centuries* (1981).

Gothic: Many of the ideas expressed in this section of the article are treated at greater length in ANDREW MARTINDALE, *Gothic Art* (1967). General studies of Gothic art include GEORGE HENDERSON, *Gothic* (1967); JOAN EVANS (ed.), *The Flowering of the Middle Ages* (1966, reissued 1984); and JOHAN HUIZINGA, *The Waning of the Middle Ages* (1924, reissued 1976; 12th Dutch ed., 1973). For the imagery of the period, the reader is referred to ÉMILE MÂLE, *The Gothic Image: Religious Art in France of the Thirteenth Century* (1958, reissued 1972; trans. of 3rd French ed., 1910), and *Religious Art from the Twelfth to the Eighteenth Century* (1949, reissued 1970; originally published in French, 1945). A useful anthology of the literary sources of the period is TERESA G. FRISCH, *Gothic Art 1140–1450* (1971). For a general treatment of English Gothic sculpture, see LAWRENCE STONE, *Sculpture in Britain: The Middle Ages,* 2nd ed. (1972); for France, MARCEL AUBERT, *La Sculpture française au moyen âge* (1947); and for Italy, JOHN POPE-HENNESSY, *Italian Gothic Sculpture,* 2nd ed. (1972).

Renaissance: There are numerous general books on Renaissance art, especially on Renaissance art in Italy, but sculpture is seldom adequately discussed in them. The best introduction to the sculpture is JOHN POPE-HENNESSY, *Italian Renaissance Sculptures,* 2nd ed. (1971). As a succinct guide to the sculpture in Florence, the most consistently important centre in Europe at this time, CHARLES AVERY, *Florentine Renaissance Sculpture* (1970), is recommended. Renaissance sculpture in northern Europe is discussed in ANTHONY BLUNT, *Art and Architecture in France: 1500–1700* (1953); WOLFGANG STECHOW, *Northern Renaissance Art: 1400–1600* (1966); GERT VON DER OSTEN and HORST VEY, *Painting and Sculpture in Germany and the Netherlands: 1500–1600* (1969); and MICHAEL BAXANDALL, *The Limewood Sculptors of Renaissance Germany* (1980). For Spain and Portugal, see GEORGE KUBLER and MARTIN S. SORIA, *Art and Architecture in Spain and Portugal and Their American Dominions: 1500–1800* (1959).

Baroque and Rococo: The best brief general discussion of Western art of this period is MICHAEL KITSON, *The Age of Baroque* (1966, reissued 1976), which includes some consideration of sculpture. For Italian Baroque sculpture, a better guide than POPE-HENNESSY (above) is provided by the sections on sculpture in RUDOLF WITTKOWER, *Art and Architecture in Italy: 1600–1750,* 3rd rev. ed. (1973, reissued 1982). ROBERT ENGGASS, *Early Eighteenth-Century Sculpture in Rome,* 2 vol. (1976); and the first two volumes (1977 and 1981) of FRANÇOIS SOUCHAL, *French Sculptors of the 17th and 18th Centuries,* must also be mentioned. For 18th-century France, the sections by Michael Levey on sculpture in MICHAEL LEVEY and WEND GRAF KALNEIN, *Art and Architecture of the Eighteenth Century in France* (1972), are excellent. For English sculpture, see the admirable account in MARGARET WHINNEY, *Sculpture in Britain: 1530–1830* (1964). For Spain, Portugal, and Latin America, see KUBLER and SORIA (above); HAROLD E. WETHEY, *Colonial Architecture and Sculpture in Peru* (1949, reprinted 1971); and PAL KELEMEN, *Baroque and Rococo in Latin America* (1951).

Neoclassicism and the 19th century: An excellent general account of Neoclassicism, which includes much of value on sculpture, is HUGH HONOUR, *Neoclassicism* (1977). For England, see DAVID G. IRWIN, *English Neoclassical Art* (1966); BENEDICT READ, *Victorian Sculpture* (1982); SUSAN BEATTIE, *The New Sculpture* (1983); and WHINNEY (above). For France and Italy, see GERARD HUBERT, *La Sculpture dans l'Italie Napoléonienne* (1964); JANE VAN NIMMEN and RUTH MIROLLI, *Nineteenth Century French Sculpture* (1971), an admirable introduction; and PETER FUSCO and H.W. JANSON (eds.), *The Romantics to Rodin* (1980), also a good introduction. A superb general introduction—perhaps the only truly comprehensive one—to Western sculpture of the 19th century is H.W. Janson's contribution to ROBERT ROSENBLUM and H.W. JANSON, *Art of the Nineteenth Century* (1984; U.S. title, *19th Century Art*).

Modern: There are numerous general introductions to modern art, but most give little space to sculpture. The best books devoted to modern sculpture are ALBERT E. ELSEN, *Modern European Sculpture: 1918–1945* (1979); HERBERT READ, *A Concise History of Modern Sculpture* (1964); and FRED LICHT, *Sculpture: 19th and 20th Centuries* (1967). Some recent developments are described in ALLEN KAPROW, *Assemblage: Environments and Happenings* (1966); and UDO KULTERMANN, *The New Sculpture* (1968; originally published in German, 1967). For a prominent sculptor's compelling but contentious account of what sculpture consists of, see WILLIAM TUCKER, *The Language of Sculpture* (1977).

(N.B.P.)

Sensory Reception

Sensory reception is the means by which an organism is able to react to changes in its external or internal environment, in particular through the activation of specific parts, which transform the energy change involved into vital processes. In many instances, environmental changes are transmitted from the receptor organ to other parts of the body by specialized cells, but in unicellular organisms such transformations are made by specialized organelles, such as light-sensitive "eye spots" or cilia sensitive to mechanical disturbances. In all higher animals, sensory reception is the special function of sensory neural elements, which "translate" changes into nerve impulses. The biological significance of such a feature is that it allows an organism to make appropriate reactions to changed circumstances; without such provisions it would be in constant danger of injury or death at the slightest environmental alteration. In general, the more highly evolved the organism, the greater the number of environmental changes sensed and the more precisely each is analyzed. Some organisms, however, possess special senses developed for perceiving environmental changes not registered by, nor relevant to, other species.

The purpose of this article is to examine the various sensory capacities of living organisms, to explain the mechanisms of sensory function, and to consider the adaptive advantages of sensing. Human sensory reception is presented in a separate section, followed by a detailed treatment of human vision and hearing. Diseases and impairments of function of these latter two vital senses are also discussed.

This article is divided into the following major sections:

ANIMAL SENSORY RECEPTION

Nature and functions of sensory systems

CLASSIFICATION OF SENSORY SYSTEMS

According to location of receptors. In general, sense cells, or receptors, located superficially in an organism receive signals from outside the organism and are parts of the exteroceptive system. Receptors located inside the body receive signals from changes taking place inside the body and belong to the interoceptive system. On activation, sense cells cause reactions appropriate to their location; they are said to respond with their local sign. For example, a decapitated frog reacts to stimulation of the skin by precisely directed limb movements aimed at wiping away the stimulus. Local sign in man is expressed by a conscious awareness of the spot being stimulated, as when a person locates a thorn in the skin. This is not true for vision, hearing, and smell, however, the sources of which are localized away from the body surface. Although some authorities believe that projection in space is learned, especially in man, for most animals such ability seems to be innate. In many cases, interoceptors stimulate channels that are never brought into consciousness; the presence of a local sign is thus shown only by the appropriateness of the resulting reactions. Internal pain is remarkable in that it is usually "misdirected" (referred) to the body surface in well-established patterns, according to its origin, a considerable help in medical diagnosis.

According to type of stimulus. More than one type of energy applied to a sense cell can, if strong enough, generate a nerve impulse, which will be interpreted by the central nervous system (CNS) as a change in the specific energy to which the cell is sensitive and will cause the same results as if the appropriate stimulus were present. Thus, a specific reflex action can be brought about by natural stimulations, such as touch of the skin, as well as by electrical stimulation of the nerve fibres activated by such touch. Each type of sense cell thus causes a specific output reaction and a specific sensation, which is the modality perceived. In other words, if the optic nerve could be functionally connected to the ear and the acoustic nerve to the eye, lightning would be heard and thunder seen.

Selectivity with regard to specific energy changes comes about in diverse ways, the simplest of which is the localization of the sense cell in such a way that it is protected from unwanted stimuli and by the use of accessory structures that make it extremely sensitive to the wanted one. The sense cells in the eye, for instance, are protected from any but the most severe changes in mechanical pressure; at the same time, the eye's optical properties focus the incoming light on the layer of sense cells constituting the retina. The hair cells of the ear, which are very sensitive to rapid changes in air pressure because of the ear's structure, are also well protected from other mechanical disturbances by shock-absorbing fluid.

Another main factor that differentiates types of sense cells is the presence of specific receptor sites for reacting with the energy to which they are specifically sensitive. Certain cells, for example, can be specifically stimulated by a given substance and no other, at least in the small concentrations required for reaction. Cells with such narrow reaction ranges are rare, however; more often, each cell has a wider spectrum, as is the case of the photoreceptors of the eye with regard to colour. Photoreceptors comprise three types of cell, each with a definite optimum but reactive to well-overlapping band widths, thereby providing for a range of colour vision. In other cases, it is the threshold (the lowest energy level) to any given stimulus that varies in different cells; this variation provides for measurement of the intensity of the stimulus. In many cases, however, intensity is coded by the frequency of the nerve impulses each receptor sends to the central nervous system.

The actual amounts of energy that can be transformed into a nerve impulse are sometimes amazingly small. One or a few photons of light absorbed may suffice not only for reception and transformation into a nerve impulse in several optic fibres but also for visual perception.

Photoreceptors

Photoreceptors are sensitive to light changes. They contain photopigments for absorption of light. The variety of photopigments in different cells determines the number of colours that can be distinguished. It is interesting to note that in insects, among other animals, colour sensitivity is extended into the ultraviolet range, though it is short in the red range. Cells especially sensitive to infrared radiation are found in the remarkable pit organs of vipers, which enable the snake to locate warm-blooded prey from a distance even when it freezes into immobility.

In the skin of warm-blooded animals, nerve endings, with or without accessory structures, are present that react especially to warming or to cooling.

Taste and smell

Well-known organs of chemical reception are those of smell and of taste. Except in cases in which there is great specificity to one substance, as, for example, the sex attractant in insects, the spectrum of chemoreceptive cells is broad. The sense of taste was long thought to be mediated by narrow, separate fibres for acid, bitter, sweet, and sour sensations; this viewpoint is now being replaced by one in which the spectra are considerably wider. Frogs have been shown to have taste cells that react specifically to distilled water. Chemoreceptors are also present as interoceptors, a well-known example being the carotid body in certain vertebrates; this organ monitors oxygen pressure in the carotid artery, which supplies the brain with blood.

Mechanoreceptors are the most widespread type of sense receptor and the most varied with regard to localization, sensitivity, and type of nerve-impulse firing. There are numerous subdivisions of the mechanoreceptive sense, such as touch, pain, sound, gravity, and muscle tone. Examples in humans include the naked nerve endings in the cornea of the eye; the Pacinian corpuscles in the skin, with their multilayered sheathlike covers; and the hair cells in the inner ear. Impulse formation may continue for as long as stimulus lasts, thus giving a continuous (tonic) type of discharge, or be limited and proportional to the rate of change of the stimulus, thus producing an abrupt (phasic) discharge. A remarkable type of mechanoreceptor occurs in the elastic organs of crustacean legs; movement-sensitive cells fire for the time a joint moves in one direction, and others fire for the opposite movement.

Electroreception

Electroreception is known only in certain fishes. Electrosensitive cells are accompanied by an organ that sends out small or large voltage changes. The sense cells occur along the long axis of the fish, enabling it both to discover food objects in the surrounding water and to locate other fish. This system is a great aid for navigation in murky water (see ELECTRICITY AND MAGNETISM: *Bioelectric effects*).

Certain animals appear to be able to orient to environmental changes for which no specific sense cells are known. Among these, magnetism is the most outstanding example. In fish, magnetic fields may well be received by electroreceptors. In insects and birds, which seem to perceive magnetic fields, no special sense cells have been implicated. The wide variety of phenomena considered as extrasensory perception in man may be based on direct influence on central-nervous elements, thus bypassing sensory input channels.

EVOLUTION OF SENSORY SYSTEMS

Specific sensory abilities do not show a clear evolutionary progression, most likely because the development of any type of sense depends on many other factors in the total ecology of a given organism. Vision, for instance, is sometimes poor or absent in a species of a class in which other members have a highly developed visual system: examples include cave-dwelling species, relatives of sighted emergent species.

Mechanical stimuli are effective in all forms of life. Specialized organs, however, appear very early in animal evolution; such organs include gravity and light receptors in jellyfish. In more advanced members of the phyla Mollusca and Arthropoda, greatly developed sense organs occur, some of which show an amazingly close resemblance to vertebrate organs; *e.g., Octopus* eyes and semicircular canals (for equilibrium). There is always a close relationship between the presence of highly developed sense organs and a region of the central nervous system; the latter is needed to "process" the incoming information in order to abstract the cues of importance to a given animal. The fact that such elaborate systems exist does not exclude the possibility of much shorter and simpler pathways, which provide for more localized and quicker reactions; for instance, the blink reflex, caused by the sudden approach of an object to the eye, bypasses the visual cortex of the brain.

INTEGRATION OF SENSORY INFORMATION

Although sensory information must be coded into a flow of nerve impulses for transmittal over distances, interactions between adjacent sense cells and sensory neurons also occur. Nerve cells can influence each other by mutual connections that result in membrane potential changes (electrical differences) in one when the other is stimulated. Similar effects are often caused by nerve impulses. When a number of nerve cells (neurons) with adjacent receptive fields are activated, it is common to find that the ones receiving the strongest stimulation suppress the response of those that are stimulated less. This action leads to a sharper difference at boundaries of stimulated and nonstimulated areas; thus, contrasts can be enhanced by this process, known as lateral inhibition. Certain sense cells have the property of being active, usually at a low rate, when not stimulated. This activity can then be either increased or decreased by appropriate stimuli. It is by such means that neurons indicating visual movements in one direction or sounds changing from one frequency to another obtain their selectivity. Such elaborations can be performed at different levels of the central nervous system. In the higher

Lateral inhibition

mammals, for instance, the fibres forming the optic nerve are mainly of two types, with small visual fields, one in which light in the centre of the field excites while light in the surrounding field inhibits, and vice versa. In animals even as highly developed as the rabbit, more complex integration has taken place in the visual periphery, and optic fibres can indicate such features as the movement of oriented lines in specific directions, which in cats and monkeys does not seem to occur before units in the brain have sampled the incoming information.

From this and other information it is clear that the use the animal makes of its senses is highly correlated with the type of sensory integration taking place in the nervous system. The ways by which a given stimulus is analyzed are varied and as yet only partially understood. It is, however, possible to build models that can be of mutual benefit for engineering and sensory information processing. By feeding back part of the incoming signal to earlier steps in the information processing, stability is greatly enhanced both in organisms and in machines. (C.A.G.W.)

Mechanoreception

Sensitivity to mechanical stimuli is a common endowment among animals. In addition to mediating the sense of touch, mechanoreception is the function of a number of specialized sense organs, some found only in particular groups of animals. Thus, some mechanoreceptors act to inform the animal of changes in bodily posture, others help detect painful stimuli, and still others serve the sense of hearing (see below *Sound reception*).

Slight deformation of any mechanoreceptive nerve cell ending results in electrical changes, called receptor or generator potentials, at the outer surface of the cell; this, in turn, induces the appearance of impulses ("spikes") in the associated nerve fibre. Laboratory devices such as the cathode-ray oscilloscope are used to record and to observe these electrical events in the study of mechanoreceptors. Beyond this electrophysiological approach, mechanoreceptive functions are also investigated more indirectly—*i.e.*, on the basis of behavioral responses to mechanical stimuli. These responses include bodily movements (*e.g.*, locomotion), changes in respiration or heartbeat, glandular activity, skin-colour changes, and (in the case of man) verbal reports of mechanoreceptive sensations. The behavioral method sometimes is combined with partial or total surgical elimination of the sense organs involved. Not all the electrophysiologically effective mechanical stimuli evoke a behavioral response; the central nervous system (brain and spinal cord) acts to screen or to select nerve impulses from receptor neurons.

Man experiences sharp, localized pain as a result of stimulation of "pain spots" (probably free nerve endings) in the skin, and dull pain, usually difficult to localize, associated with inner organs. The sensory structures of pain spots in the skin differ from other receptors in that they respond to a wide range of harmful (noxious or nociceptive) stimuli. Excessive stimulation of any kind (*e.g.*, mechanical, thermal, or chemical) may produce the human experience of pain. Apart from eliciting this subjective feeling of pain, stimulation of pain receptors in the human skin is objectively characterized by such signs of emotional expression as weeping and by efforts to withdraw from the stimulus. The reflex withdrawal of his hand from a burning stimulus may begin even before the person becomes conscious of the pain sensation.

Experience of pain in animals

Judging from objective criteria, responses to painful stimuli also occur in nonhuman animals, but, of course, any subjective experience of pain sensation cannot be directly reported. Still, the question of painful experience among animals is of considerable interest because investigators (*e.g.*, medical researchers) are often obliged to subject laboratory animals to treatments that would elicit complaints of pain from a man. If a cat's tail is accidentally stepped on, the pitiful screeching and efforts to withdraw are so strikingly similar to human reactions that the observer is led to attribute the experience of pain to the animal. If one treads accidentally on an earthworm and observes the animal's apparently desperate struggles to get free, he might again be inclined to suppose that the worm feels pain. This sort of "mind reading," however, is inherently uncertain and may be grossly misleading.

The following observations illustrate some of the difficulties in making judgments of the inner experiences of creatures other than man. After the spinal cord of a fish has been cut, the front part of the animal may respond to gentle touch with lively movements, whereas the trunk, the part behind the incision, remains motionless. A light touch to the back part elicits slight movements of the body or fins behind the cut, but the head does not respond. A more intense ("painful") stimulus, however (for instance, pinching of the tail fin), makes the trunk perform "agonized" contortions, whereas the front part again remains calm. To attribute pain sensation to the "painfully" writhing (but neurally isolated) rear end of a fish would fly in the face of evidence that persons with similarly severed spinal cords report absolutely no feeling (pain, pressure, or whatever) below the point at which their cords were cut.

Aversive responses to noxious stimuli nevertheless have a major adaptive role in avoiding bodily injury. Without them, the animal may even become a predator against itself; bats and rats, for instance, chew on their own feet when their limbs are made insensitive by nerve cutting. Some insects normally show no signs of painful experience at all. A dragonfly, for example, may eat much of its own abdomen if its tail end is brought into the mouthparts. Removal of part of the abdomen of a honeybee does not stop the animal's feeding. If the head of a blow-fly (*Phormia*) is cut off, it nevertheless stretches its tubular feeding organ (proboscis) and begins to suck if its chemoreceptors (labellae) are brought in touch with a sugar solution; the ingested solution simply flows out at the severed neck.

At any rate, responsiveness to mechanical deformation is a basic property of living matter; even a one-celled organism such as an amoeba shows withdrawal responses to touch. The evolutionary course of mechanoreception in the development of such complex functions as gravity detection and sound-wave reception leaves much room for speculation and scholarly disagreement.

RECEPTION OF EXTERNAL MECHANICAL STIMULI

The sense of touch. Sensitivity to direct tactual stimulation—*i.e.*, to contact with relatively solid objects (tangoreception)—is found quite generally, from one-celled organisms up to and including man. Usually the whole body surface is tangoreceptive, except for parts covered by thick, rigid shells (as in mollusks). Mechanical contact locally deforms the body surface; receptors typically are touch spots or free nerve endings within the skin, often associated with such specialized structures as tactile hairs. The skin area served by one nerve fibre (or sensory unit) is called a receptive field, although such fields overlap considerably. Particularly sensitive, exposed body parts are sometimes called organs of touch—*e.g.*, the tentacles of the octopus, the beak of the sandpiper, the snout of the pig, or the human hand.

Stimulation of the human skin with a bristle reveals that touch (pressure) sensation is evoked only from certain spots. These pressure spots, especially those on hairless parts (*e.g.*, palm of the hand, or sole of the foot), are associated with specialized microscopic structures (corpuscles) in the skin. Pressure spots are most densely concentrated on the tip of the human tongue (about 200 of them per square centimetre, or 1,300 per square inch), roughly twice their concentration at the fingertip. A characteristic feature of many tactile sense organs is their rapid and complete adaptation (*i.e.*, temporary loss of sensitivity) when stimulated. Still, in man a distinction can be made between transient and more prolonged pressure sensations.

Relatively little research has been done with regard to the physiology of individual tangoreceptors in vertebrates. The Pacinian corpuscle of higher vertebrates, however, has been studied in isolation (see *Human sensory reception* below, for illustrations). These corpuscles, found under the skin, are scattered within the body, particularly around muscles and joints. Local pressure exerted at the surface or within the body causes deformation of parts of the corpuscle, a shift of chemical ions (*e.g.*, sodium, potassium), and the appearance of a receptor potential at the nerve ending. This receptor potential, on reaching sufficient (threshold) strength, acts to generate a nerve impulse within the corpuscle. Among insects, movements of tactile hairs have been shown (sometimes specifically) to affect the receptor potential and the impulse frequency in the connected nerve fibre.

Many vertebrates and invertebrates can localize with some precision points of tactual stimulation at the body surface. People typically can still distinguish two sharpened pencil points, or similar pointed stimuli, when the points are separated by as little as about one millimetre (0.04 inch) at the tip of the tongue. (When moved closer together, the two points are perceived as one.) The human two-point threshold is about two millimetres at the finger tip, reaching six or seven centimetres (2.4–2.8 inches) at the skin of the back. Such tactual ability serves blind people when they read raised type (Braille) with their fingers. Closely related functions include the ability to distinguish between tactile stimuli that differ qualitatively; for example, between a rough and a smooth surface. This ability is even observable in the ciliate *Stylonychia* (a one-celled relative of *Paramecium*).

Reading Braille

Sensory contact with the ground below often informs animals about their spatial position. Nocturnal animals (for example, some eels) find shelter during the day by keeping as much of their skin as possible in contact with solid objects in the surroundings (thigmotaxis). Animals that live in running water usually maintain their position as they turn and swim head-on against the current (rheotaxis). Study of rheotaxic behaviour reveals that the sensory basis almost exclusively depends on visual or tactile stimuli (or both) arising from the animal's movements relative to the solid bottom or surroundings. The long antennae of many arthropods (*e.g.,* crayfish) and the lengthened tactile hairs (vibrissae) on the snouts of nocturnally active mammals (*e.g.,* cat, rat) serve in tactually sensing objects in the vicinity of the animal's body, extending and enriching the adaptive function of the sense of touch.

Lateral-line organs. *Mechanoreceptor function.* All of the primarily aquatic vertebrates—cyclostomes (*e.g.,* lampreys), fish, and amphibians—have in their outer skin (epidermis) special mechanoreceptors called lateralline organs. These organs are sensitive to minute, local water displacements, particularly those produced by other animals moving in the water. In this way, approaching organisms are detected and localized nearby before actual bodily contact takes place. Thus the lateral lines are said to function as receptors for touch at a distance, serving to perceive and locate prey, approaching enemies, or members of the animal's own species (*e.g.,* in sexual-display behaviour).

From *Experientia* (1952)

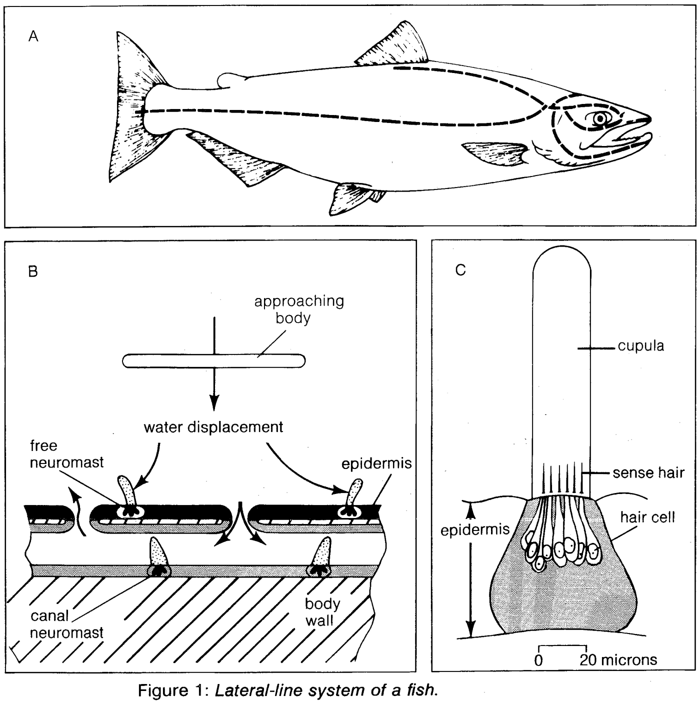


Figure 1: *Lateral-line system of a fish.*
(A) Bodily location of lateral lines; (B) longitudinal section of a canal; (C) superficial neuromast.

Each epidermal organ, called a sense-hillock or neuromast (Figure 1C), consists of a cluster of pear-shaped sensory cells surrounded by long, slender supporting cells. The sense hairs on top of the sensory cells project into a jellylike substance (the cupula) that bends in response to water displacement. The cupula stands freely in the surrounding water, grows continuously (*e.g.,* as a human fingernail), and wears away at the top. Sense organs of this type are distributed along definite lateral lines on the head and body of the animals (Figure 1A), developing in the outer layer of cells (ectoderm) of the embryo from a thickening called the lateral placode. From the central part of the same placode the sensory cells of inner-ear structures (the labyrinth) arise. The common embryologic origin and structural similarities of mature neuromasts and labyrinthine cell groups have led to the designation of all of these organs as the acoustico-lateralis system. The nerves to all the sense organs of the system arise from a common neural centre (called the acoustic tubercle in the wall of the brain's medulla oblongata). Among such amphibians as frogs, lateral-line organs and their neural connections disappear during the metamorphosis of tadpoles; as adults they no longer need to feed under water. The higher land-inhabiting vertebrates—reptiles, birds, and mammals—do not possess the lateral-line organs; only the deeply situated, labyrinthine sense organs persist.

The sensory cell of a neuromast bears one relatively long hair (kinocilium) and about 50 shorter ones (stereocilia). The kinocilium is inserted eccentrically on top of the sense cell; the stereocilia are arranged in parallel rows. In about half of the hair cells of a neuromast, the kinocilium is found on one (and the same) side of the cell; in the remaining hair cells it is found on the opposite side. In most cases these are cranial and caudal side, respectively. In the clawed frog (*Xenopus*), each group of hair cells in a neuromast connects to its own nerve fibre; hence there are two fibres per sense organ. The hair cells send a continuous series of neural impulses toward the acoustic tubercle in the absence of adequate external stimulation. A longitudinal water current along the toad's body surface, however, selectively increases or decreases the frequency of impulses from the cranial and caudal cells, depending on whether the flow is from head to tail or vice versa; current directed at right angles to such neuromasts has no effect. The impact of the moving water moves the cupula to deform the sensory hairs. Even minute cupula displacements of less than one thousandth of a millimetre are clearly effective in altering the impulses.

Detection of water currents

In *Xenopus,* as well as in other animals that have lateral-line organs, there are also some neuromasts with their hair cells asymmetrical at right angles to the head-tail axis. These add directional sensitivity so that other animals moving nearby in the water are well distinguished and localized. The postulated function of the lateral-line organs in the reception of low-frequency propagated pressure waves ("subsonic sound") has not been verified behaviorally. At very short distances, however, a vigorous low-frequency sound source stimulates the lateral-line system on the basis of acoustical near-field effects (water particle displacements), just as does any moving or approaching object.

Cyclostomes, many bony fishes, and all the aquatic amphibians studied have only superficial ("free") neuromasts of the kind described above. In the development of most fish, however, a number of structures called lateral-line canals (Figure 1B) are formed as a secondary specialization. They begin as grooves that develop in the epidermis along the main lateral lines; thus, a number of formerly free neuromasts are taken down to the bottom of each groove. The walls of the grooves then grow together above the neuromasts. Eventually the grown-together walls form canals under the epidermis, containing in their walls a series of canal neuromasts and a chain of openings to the outside (canal pores) along the lateral lines. The cupulae are changed in form, fitting the canal somewhat like swinging doors. The canal is filled with a watery fluid. Stimulation occurs essentially in the same way as with free neuromasts: local, external water displacement is transmitted via one or more canal pores to produce a local shift of the canal fluid to move cupulae. The sense cells in the canal neuromasts are polarized in the direction of the canal.

Canal specialization is particularly well developed in lively species of fish that swim more or less continuously and in bottom dwellers that live in running or tidal waters. Canalization has been interpreted as a case of adaptive evolution, serving to avoid the almost continuous, intense stimulation of free neuromasts by water flowing along the fish body during swimming or, in the case of relatively inactive bottom dwellers, by the external currents. These coarse water displacements probably mask subtly changing stimuli from detection by the lateral-line organs on the surface of the animal's body. Canal neuromasts are shielded in large degree from these masking currents.

Locating prey, predators, and sexual partners

The lateral-line organs function mainly in locating nearby moving prey, predators, and sexual partners. Usually these objects must be much closer than one length of the animal's body to be detected in this way; even intense stimuli are hardly ever detected beyond five body lengths away. Lateral-line function in rheotactic orientation against currents is restricted mainly to inhabitants of small currents such as mountain brooks, where marked differences of water-flow velocity affecting the fish body locally are likely to occur. Compared to their use of other sensory functions (*e.g.,* vision) the animals depend little on ability to sense extremely close, resting objects (obstacles) through the lateral lines. Obstacle detection of this kind does not arise from reflection of water waves; rather, the pattern of water displacement around the moving fish abruptly undergoes deformation at the near approach of an obstacle as the result of compression; the fish encounters a sudden rise in water resistance in the immediate vicinity of the obstruction. Nor do the lateral-line organs function to regulate or coordinate the animal's movements on the basis of the water flow or pressure variations along its body produced by swimming; neither do they serve for the reception of water-transmitted propagated sound waves (hearing).

Ampullary lateral-line organs (electroreceptors). Perhaps the most interesting specialization of the lateralline system is the formation in several groups of fish of deeply buried, single electrically sensitive organs. Such structures, for example, are found on the head of all the elasmobranchs (*e.g.,* sharks and rays), and are called ampullae of Lorenzini. Similar organs include those on the head of *Plotosus,* a marine bony fish (teleost); structures called mormyromasts in freshwater African fish (mormyrids) and in electric eels (gymnotids); what are named small pit organs of catfishes (silurids); and possible related organs in several other fish groups. These are known as ampullary lateral-line organs, and they have features in common. The sensory cells are withdrawn from the body surface, lack kinocilia, and have no mechanical contact with the surrounding water through a cupula. The latter attribute, indeed, is typical for all the acousticolateral end organs, except ampullary sense organs, in which the sense cells lie within the wall of a vesicle (or ampulla) that opens to the surface through a tubelike duct. Ampulla and duct are filled with a gelatinous substance that has excellent electrical conductivity.

Fish with ampullary sense organs are found to be remarkably sensitive to electrical stimuli—*i.e.,* minute, local potential differences in the surrounding water at their body surface. In behavioral experiments with sharks and rays, sensitivity to changes of 0.01 microvolt per centimetre (one microvolt = 1/1,000,000 of a volt) along the body surface has been found for the ampullae of Lorenzini. Similar, though somewhat higher, values have been recorded from the ampullary nerve fibres. A decrease in voltage at the opening of the ampulla causes an increase of the spontaneous nerve-impulse frequency; an increase in voltage at the opening produces the opposite response. Through their electrical sensitivity, such fish can detect and locate other organisms in darkness, in turbid water, or even when these organisms are hidden in the sand or in the mud of the bottom.

Electrical "radar"

Sharks, rays, and most catfishes are able to detect electrical changes (biopotentials) emanating from other organisms. The freshwater mormyrids and eels, on the other hand, have special signal-emitting electric organs. They produce a series of weak electric shocks (up to a few volts), sometimes quite regularly and frequently; for example, about 300 shocks per second in the mormyrid fish *Gymnarchus.* In this way, a self-generated electric field is created in the immediate surroundings. Any appropriate object (for example, a prey animal with good conductivity in relation to fresh water) will cause a deformation of the electric field and can thus be detected in a radar-like manner through the sensitive ampullar electroreceptors.

Some theorists suggest that initially mechanoreceptive lateral-line organs evolved into electroreceptors. At any rate, evidence of a certain double sensitivity—to mechanical and to electrical stimuli—has been observed in electrophysiological experiments with Lorenzinian ampullae. This double sensitivity has not been found, however, in behavioral experiments; alterations in behaviour indicate that ampullary lateral-line organs merely serve the animal as electroreceptors in adapting to the environment.

Other varieties of mechanoreception. *Surface waves.* Several species of animals living at or near the water surface use surface waves or ripples emanating from potential or struggling victims to locate their prey quickly: examples are the toad *Xenopus,* several fish species, and such insects as the back swimmer (*Notonecta*) and the water strider (*Gerris*). The whirligig beetle (*Gyrinus*) also uses surface ripples to avoid collisions with obstacles and companions. The sensory structures involved range from specialized tactile hair receptors (trichobothria) to internally located cells (proprioceptors) in movable body appendages and lateral-line organs.

Water and air currents. Special water-displacement receptors found in lobsters (*Homarus*) are most reminiscent of the lateral-line organs in vertebrates. Water-current receptors also enable several kinds of bottom-dwelling invertebrates to orient themselves (rheotaxis) in rivers and tidal currents. Many predators among these animals also respond chemically (see *Chemoreception* below), moving against the current (positive rheotaxis) until the prey is reached. In this way, for example, certain marine snails easily find their particular prey (sea anemones). Similarly among insects, the chemical "smell" of prey or of potential sex partners elicits a tendency to move against the wind (anemotaxis) until the source of the chemical stimulus is found. Several types of air-current receptors (true mechanoreceptors) on the heads of insects enhance such chemoreceptive behaviour. In flying locusts, an air current directed appropriately toward the head elicits compensatory reflex flight movements. The receptors involved (groups of hair sensilla on the head) mediate small corrections in the maintenance of straight flight; major guidance, however, derives from the insect's visual contact with the ground below.

Vibration reception. Adaptation and recovery occur most rapidly among touch receptors, and they tend to respond well to repeated stimulation, even of relatively high frequency. Thus, a person can feel whether an object is vibrating; above a threshold frequency of about 15 cycles per second (cps), discretely perceived tactual stimuli seem to fuse into a quite new and distinct vibratory sensation. The upper frequency limit of this vibration sense is found at several thousand cps among normal individuals, with sensitivity being maximal in the range of 200 cps (above a threshold amplitude of about 100 millimicrons). Just as pitch is discriminated in hearing, differences of about 12 to 15 percent in vibration frequencies can be distinguished by most people.

Specialized vibration receptors

Vibration sensitivity is not limited to man; fish, for instance, also may respond to low-frequency water vibrations with tactile receptors. In addition, several kinds of animals have special vibration receptors. In some insects, a group of specialized structures (chordotonal sensilla) in the upper part of each tibial segment of the leg signal vibrations from the ground below. In the cockroach, the threshold amplitude for vibrational stimuli of this kind has been found to be less than 0.1 millimicron. Birds have special receptors (corpuscles of Herbst in the tibiotarsal bone of the leg) with which they can detect slight vibrations of the twig or branch on which they sit. Perhaps birds are alerted at night in this way to approaching predators; maximal sensitivity is at about 800 cps, and the threshold amplitude is close to 20 millimicrons. Spiders also use their vibration sense to locate prey in the web.

Generalized hydrostatic pressure. Several types of aquatic animals are sensitive to small changes of hydrostatic, or water, pressure. Among fish, this applies particularly to the order Ostariophysi (Cypriniformes), which includes about 70 percent of all the freshwater species of bony fishes. The swimbladder in these animals is connected with the labyrinth (sacculus) of the inner ear through a chain of movable tiny bones, or ossicles (weberian apparatus). Alterations in hydrostatic pressure change the volume of the swimbladder and thus stimulate the sacculus. These fish can easily be trained to respond selectively to minute increases or decreases in pressure (for example, to a few millimetres of water pressure), indicating that they have a most refined sense of water depth. All of these fish are so-called physostomes, which means that they have a swimbladder duct through which rapid gas exchange with the atmosphere can occur; many live in relatively shallow water. The hydrostatic-pressure sense can function to inform the animals about their distance from the surface or about the direction and velocity of their vertical displacement. It also appears that improvement and refinement of the sense of hearing arises through the swimbladder's connections via the weberian apparatus with the labyrinth.

Sensing of hydrostatic pressure

The sensitivity of several kinds of crustaceans to relatively small hydrostatic-pressure changes (as low as five to 10 centimetres [two to four inches] of water pressure) is most remarkable because these animals have no gas-filled cavity whatsoever. The mechanism by which the stimuli are detected remains a puzzling question, although information about changing water depth during tidal ebb and flow clearly would seem to have adaptive value.

RECEPTION OF INTERNAL MECHANICAL STIMULI

Some proprioceptors (internal receptors) for mechanical stimuli provide information about posture and movements of parts of the body relative to each other; others contribute to an undisturbed course of coordinated muscular actions (*e.g.,* in locomotion). Best known from studies of vertebrates and arthropods, some are tonic proprioceptors (serving to maintain muscle tone in posture); others are of the phasic type (serving movement); still others have a mixed phasic-tonic character. In principle, proprioceptors can be stimulated adequately by pressure or stretching during active movements of the animal (reafferent stimulation) as well as through passive external pushing and pulling (exafferent stimulation). One passive factor, particularly in land-inhabiting animals, is gravity as it acts on bodily tissues or organs. Proprioceptors thus not only serve reflex adjustments in posture and relatively automatic movements of parts of the body with respect to each other (as in driving an automobile) but they also provide gravitational information about the position of limbs or of the whole body in space. To the extent that they are gravity detectors, these sensory structures are properly called external receptors (exteroceptors instead of proprioceptors). For receptors that are diffusely located within the body, a clean distinction between proprioceptive and possible exteroceptive function (gravity reception) is experimentally practicable only under conditions of weightlessness, as in space travel.

Vertebrates. *Muscle spindles.* Well-known proprioceptors of all the four-limbed vertebrates studied are the muscle spindles occurring in the skeletal (striate) muscles; fish muscles show structurally simpler but functionally comparable receptors. Each muscle spindle in mammals consists of a few slender, specialized (intrafusal) muscle fibres that are surrounded by a sheath of connective tissue filled with lymph fluid (Figure 2). The muscle spindle itself is surrounded by and arranged parallel to the ordinary (extrafusal) muscle fibres. Each intrafusal fibre consists of contractile (motor) parts at both ends and a noncontractile sensory midsection that serves as a receptor for stretch (changes of length and tension). There is double (primary and secondary) sensory innervation in mammals, but the secondary endings are lacking in lower vertebrates. Even when the animal is at rest, both types of endings are active (under the tension of normal muscle tonus). Additional stretch (lengthening) of the intrafusal midsection increases the nerve impulse frequency, and relaxation (shortening) causes a decrease. The primary (phasic-tonic) ending responds quickly; responses of the secondary (tonic) endings are slower.

The length of the muscle spindle as a whole varies with the contraction phase and the length of the muscle to which it belongs. The length of the sensory midsection, however, may change more or less independently because its motor nerve endings function apart from the innervation of the extrafusal muscle fibres. Thus the ratio of extrafusal–intrafusal contraction determines whether or not a change of length in the midsection will occur during muscle activity. There are reasons to suppose that midsection stretch remains more or less unchanged during self-initiated ("voluntary") movements; reafferent stimulation of muscle spindles would be avoided in this way. But as soon as an unexpected (exafferent) stretch of a muscle occurs—for example, when a leg pushes against an obstacle during locomotion—the midsections stretch to produce an increase of impulse frequency. This neural activity elicits a compensatory reflex contraction of the stretched muscle, as in the knee jerk during medical examinations: a blow beneath the kneecap causes stretch of a thigh muscle, stimulation of its muscle spindles, and a compensatory jerking contraction of the same muscle.

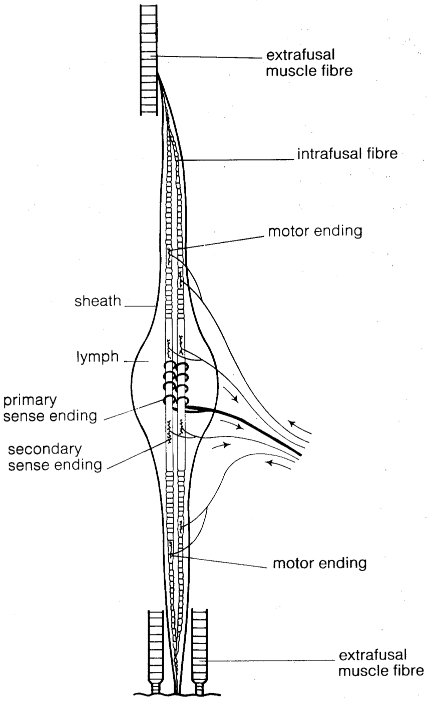


Figure 2: Mammalian muscle spindle.

Knee jerk

Tendon organs. Branched nerve endings on vertebrate tendons (not far from their point of attachment to muscle) also respond to stretch; however, they are decidedly less sensitive than are muscle spindles. These tendon organs produce no impulses under the stretch of normal, resting muscle tonus. Neither is there a mechanism preventing reafferent stimulation of tendon organs, nor does it make any difference whether the stretch is brought about by active muscle contraction or passively following external influence. In both cases tendon receptors respond according to the intensity of the stretch; their response causes relaxation of the attached muscle and may serve (among other functions) to prevent anatomical damage.

Human awareness of posture and movement of parts of the body with respect to each other (kinesthetic sensations) is attributable neither to muscle spindles nor to tendon organs. The sensations are based on stimulation of sensory nerve endings of various types at the joint capsules and of stretch receptors in the skin. There are also mechanoreceptors in the walls of some blood vessels (*e.g.*, in the aorta and the carotid sinus); these are sensitive to blood-pressure changes and play a regulatory role in the circulatory system.

Invertebrates. Among invertebrates, the arthropods exhibit the most readily distinguished proprioceptors, called muscle-receptor organs and chordotonal proprioceptors. Both types of structure occur in crustaceans as well as in insects. Adequate stimuli are variations in length and tension (stretch).

Muscle receptor organs. Although they structurally and functionally resemble the muscle spindles of vertebrates, arthropod muscle receptor organs are always situated outside of the muscles proper. Numerous branches of multipolar primary nerve cells are connected with the noncontractile midsection of specialized muscle fibres, both ends of which are contractile and have an efferent (motor) innervation. In crustaceans, the muscle receptor organ contains two elements: a slowly contracting, nonadapting tonic fibre and a quickly contracting, rapidly adapting phasic element.

Chordotonal proprioceptors. Widely distributed among arthropods, chordotonal receptor organs are thin, elastic, innervated strands of connective tissue, stretched between adjacent segments of the body or of leg joints. The sensory endings of a few bipolar primary nerve cells, each provided with a spiny sensillum (scolopidium), are attached to the strand. Chordotonal proprioceptor organs generate neural impulses that show them to contain both phasic movement receptors and tonic pressure receptors; sometimes two varieties of each. Thus there are receptors that selectively respond only during flexion, only in the flexed position, only during stretch, or only in the stretched state of the given strand. Several kinds of insects, apart from their clearly proprioceptive-chordotonal functions, have other chordotonal elements that serve as typical exteroceptors. Sense organs of this type (tympanic and subgenual organs in legs, organs of Johnston in the antennae) may function in the reception of sound waves, of vibrations in the ground, or of other external mechanical stimuli. Many insects also have a special type of chordotonal-proprioceptor structure (campaniform sensilla) not found in crustaceans. Sensory endings of primary nerve cells are connected with thin, dome-shaped (campaniform) spots on the exoskeleton. These campaniform sensilla respond to external stimuli such as local tensions and deformations of the body surface. They function in the regulation of such movements as the beating of wings in locusts. Similarly functioning proprioceptors (lyriform organs) are also observed among spiders.

In insects, body posture and movements of individual body parts with respect to each other can be detected through groups of external tactile hairs implanted near the joints between adjacent skeletal elements. Some function as rotation receptors or exteroceptors to detect the direction of gravity.

Among other invertebrates, the cephalopod *Octopus* clearly exhibits proprioceptive abilities, though specific receptors have not yet been identified. These animals, however, seem unable to integrate proprioceptive data in the central nervous system with other sensory information in learning. Thus an octopus readily can be taught to discriminate between two small cylindrical objects (both provided with longitudinal ribs) if the ribs on one of them are somewhat coarser than those on the other. But the animal cannot learn to distinguish between cylinders of the same size if the ribs are equally coarse and if they are longitudinal on one and transverse in the other; nor can it learn to discriminate between small objects of different form or different weight. This indicates that an octopus cannot learn any discrimination that depends on sensory information about the position of the arms and suckers making contact.

Octopus learning

MAINTENANCE OF EQUILIBRIUM

Active maintenance of equilibrium during bodily movement (*e.g.*, in locomotion) requires appropriate sensory functions. Although many animals usually maintain their bodies with the long axis horizontal (backside up), man being a notable exception, there are frequent departures from the usual position. A fish may dive steeply downward and a man may alter his normal orientation by lying down at full length. In no case, however, need there be any loss of equilibrium. Every deviation means an equilibrium disturbance and evokes compensatory reflex movements, not only a deviation from the usual position as in most laboratory experiments.

Maintenance of equilibrium is based upon contact of the animal with the external world; several sensory systems may play a role in this context. When an animal moves over a solid surface, tactile stimuli usually predominate as cues. It has been noted above how proprioceptors in vertebrates and arthropods can also contribute to spatial orientation; bodily tissues under gravity weigh vertically down and stimulate internal mechanoreceptors in a way that depends on, and varies with, the animal's spatial position. When they are out of contact with the ground, many animals orient themselves in space by keeping their back (dorsal) side turned up toward the light. Visual cues also can serve equilibration; for example, through compensatory body movements (optomotor reflexes) brought about by the shifts of the image of the environment over the retina of the eye. For the receptors mentioned thus far, however, equilibration is not the unique function. There are other sensory structures that are genuine organs of equilibrium in that they primarily and exclusively serve orientation of posture and movement in space.

Secondary cues

Gravity receptors. Because of the constancy of its magnitude and direction, gravity is most suitable in providing animals with cues to their position in space. The sense organs involved (statoreceptors) usually have the structure of a statocyst, a fluid-filled vesicle containing one or more sandy or stonelike elements (statoliths). Sensory cells in the wall of the vesicle have hairs that are in contact with the statolith, which always weighs vertically down. Hence, depending on the animal's position, different sense cells will be stimulated in statocysts with loose statoliths (Figure 3A); or the same sense cells will be stimulated in different ways in statocysts with a statolith loosely fixed to the sense hairs (Figure 3B).

From W. von Buddenbrock, *Vergleichende Physiologie*

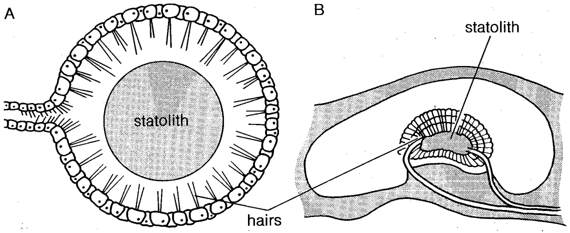


Figure 3: *Statocyst gravity receptors.*
(A) With a free-moving statolith, as in a mollusk (scallop), and
(B) with statolith loosely fixed to hair cells, as in a crustacean (opossum shrimp).

Statocysts are found in representatives of all of the major groups of invertebrates: jellyfish, sandworms, higher crustaceans, some sea cucumbers, free-swimming tunicate larvae, and all the mollusks studied thus far. Analogous receptors that occur generally in vertebrates are the ear's utriculus and probably (to a degree) also two other otolith organs (sacculus and lagena) of the ear (labyrinth). Statocysts (including vertebrate labyrinthine statoreceptors) develop embryologically from local invaginations of the body surface. In primitive evolutionary forms, the interior of the statocyst is in open communication with the surrounding sea and thus is filled with water; statoliths usually are sand particles taken up from outside. In a few animal groups, this developmental stage is only found during the larval phase, the initial opening to the exterior being closed in the adult animal. In more advanced forms, the liquid content (statolymph) and the statoliths are produced by cells in the wall of the organ. This specialized type of closed statocyst is found in many snails, in all the cephalopods such as the squid (except *Nautilus*), and in the vast majority of vertebrates (from bony fishes up to and including mammals).

Geotactic behaviour

Statocyst function may be studied by observing compensatory reflexes under experimental conditions. When the position of a laboratory animal is appropriately changed, movements of such body parts as the eyes, head, and limbs can be observed. Such movements tend to counteract the imposed change and to restore or to maintain the original position. Evidence of statoreceptor function is provided if these reflexes are abolished after surgical elimination of both statocysts. Many animals exhibit locomotion that is gravitationally directed vertically down or up (positive or negative geotaxis, respectively). Geotactic behaviour may be experimentally altered by whirling the animal in a centrifuge to change the direction and to increase the intensity of the force exerted on the sensory hairs by the statoliths. Molting crustaceans shed the contents of their statocysts along with their exoskeleton. If such an animal is placed in clean water containing iron filings, it takes up new iron statoliths instead of the usual sand grains. By moving a magnet to vary the direction of the force exerted by the metal statoliths, the animal can be made to adopt any resting position, even to stay upside down. Statoliths can be washed out of the open statocysts of a shrimp without damaging the sensory hairs. When the hairs are pushed in different directions with a fine water jet, the shrimp exhibits compensatory reflexes. In this way, it has been shown that each statocyst signals a change of position around the animal's long axis; the same reaction is found to occur after removal of the statocyst on one side only. Electrical impulses in the statocyst nerve can be recorded while the animal is in different spatial positions, or during experimental deflection of the sensory hairs. Such experiments reveal that both vertebrates and decapod crustaceans (*e.g.*, shrimp) exhibit spontaneous and statolith-induced neural activity in the lining (epithelium) of the gravity receptor.

Spontaneous activity. The sensory epithelium of a statocyst is spontaneously active, initiating a continuing series of impulses directed toward the central nervous system (even when the statoliths are experimentally removed from the statocyst). This resting frequency of neural activity is fairly constant and completely independent of the animal's position in space. In vertebrates and in crustaceans, spontaneous activity of the left statocyst affects the central nervous system to produce a tendency of the animal to roll to the right about its long axis; spontaneous activity of the right statocyst prompts a tendency to roll to the left. Normally, these rolling tendencies neutralize each other in the central nervous system, not becoming manifest unless the statocyst on one side of the body is functionally eliminated by complete surgical removal, by destruction of its sensory epithelium, or by cutting its nerve. This intervention permits the influence of the spontaneous activity generated in the remaining statocyst to be felt, and the animal tends to roll toward the operated side. Unilateral (one-sided) removal of the statoliths alone, however, does not produce such an effect so long as the sensory cells in the epithelium remain intact. The rolling tendency of a unilaterally operated animal usually diminishes little by little in the course of hours or days, until it finally disappears completely. If the remaining statocyst is then removed, rolling occurs again, but this time to the other (last operated) side. This tendency also diminishes and disappears with time. Apparently the unbalancing effect of the spontaneous influx from a statocyst is gradually counteracted in some unknown way by the central nervous'system.

Statolith influences. Vertebrates and crustaceans have statoliths that are loosely connected to the sensory hairs by a sticky substance. With such a mechanical arrangement, the statolith stimulates the sensory cells by parallel (shearing) motion rather than by pressure or pull at right angles to the epithelium. The effects are demonstrable in experiments with fish, based on the dorsal-light orientation noted above. In a laboratory darkroom, if light shines at a fish from one side, the animal assumes an oblique position. While the fish tends to turn on its side (with its back side to the light), gravity tends to keep it vertical; the oblique position is the result. In a whirling centrifuge, the pressure exerted by the statoliths may be increased. When this is done, the fish rights itself almost precisely to the degree that the shearing force exerted by the statoliths is held constant.

Otolith organs in vertebrates

In vertebrates, statoreception is localized in the head within the labyrinth, particularly within the utriculus, one of the three statolith (or otolith) organs (Figure 4). The statolith is surrounded by a gelatinous substance akin to the cupula of the lateral-line organs. In most higher vertebrates, the head moves rather flexibly because it is not rigidly connected to the trunk. Thus information coming from the utriculi has to be neurally integrated centrally with impulses from proprioceptors that signal the position of the head with respect to the limbs and trunk (for example, neck receptors), if the animal is to orient its head and body appropriately in space.

The roles played by the remaining otolith organs of the labyrinth (sacculus and lagena) in statoreception remain unclear. Their sensory epitheliums (maculae) are roughly at right angles to each other and to that of the utriculus. In view of their arrangement, it was once supposed that the three otolith organs of the labyrinth would serve to detect position in three spatial planes (indeed, the three semicircular canals do serve to detect rotation in different planes). It has been found, however, that the sacculus and the lagena (as far as it is present) can be put out of function bilaterally in representatives of all the classes of vertebrates without causing overt equilibration disturbances. On the other hand, some secondary statoreceptor function has been demonstrated for these otolith organs in all the animals from fish up to and including man.

In the special case of flatfishes (*e.g.*, halibut, sole, flounder), the normal upright position in the juvenile stage changes to one of swimming and lying on one side as an adult. The eye from that side migrates to the upper surface; but the situation of both labyrinths remains unchanged. Hence, the originally horizontal maculae of the utriculi are now oriented vertically. In these fish, the sacculi (usually the major organs of hearing in bony fishes) indeed may be shown to serve as statoreceptors. At any rate, the same otolith organ may function in one fish species as an organ of hearing and in another as a gravity receptor; clearly, both functions depend on basically identical mechanical stimulation.

As receptors belonging to the acousticolateralis system, the otolith organs of vertebrates have hair cells of the same type that is found in lateral-line neuromasts. Under the electron microscope, the sensory hair cells show a pattern of polarization (arrows in Figure 4) throughout the macula, indicating the directions in which the shearing otolith should have an activating or an inhibiting influence. Results of physiological investigations thus far performed agree well with these deductions.

Statocyst research among crustaceans

Among the invertebrates, most statocyst research has been done with such decapod crustaceans as lobsters.

From Cold Spring Harbor Symposia on Quantitative Biology (1965)

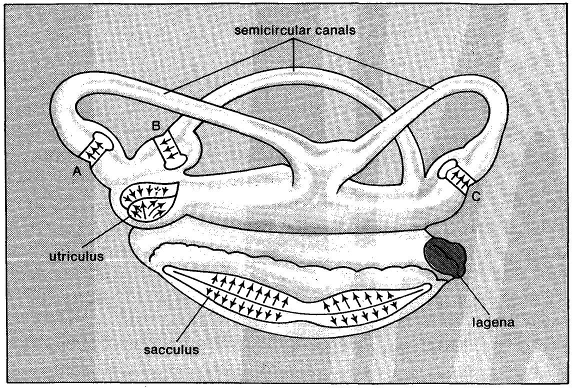


Figure 4: Right labyrinth of a codlike fish called the burbot, seen from above. Ampullae of anterior, horizontal, and posterior semicircular canals are lettered A, B, and C, respectively. Arrows show direction of hair-cell polarization.

The working mechanism of their statocysts conforms with the physiological principles of vertebrate statoreception discussed above. The results of electrophysiological investigations support the conclusions drawn from behavioral observations. In some crustacean statocysts (for example, in the lobster, *Homarus*), special statoreceptors are found that signal the same bodily position differently, depending on the direction of movement through which it was reached (hysteresis effect). The part played by the statocyst in equilibration has been investigated in several other invertebrate groups, among them jellyfish, sandworms, and such mollusks as scallops, common snails, sea hare, and octopus. Each sensory cell from the vertical macula in a statocyst of the octopus bears up to 200 kinocilia, and all the cilia of each cell are polarized in the same direction. On the macula as a whole, there is a radiating polarization pattern, the activating direction pointing everywhere from the centre to the margin. Compensatory eye reflexes resulting from tilting the animal head down or head up around a transverse axis reveal a hysteresis effect. After unilateral-statocyst removal, mollusks do not tend to roll toward the operated side (as do vertebrates and crustaceans) but toward the side of the remaining statocyst.

The almost complete absence of statocysts in insects is remarkable in view of evidence that many of them have a high degree of sensitivity to the direction of gravity. Receptors involved are specialized tufts of tactile hairs at the external body surface; in the honeybee, such groups of hairs are notably found between head and thorax and between thorax and abdomen. The adaptive function of these static (gravity) receptors becomes manifest in the honeybee "dance language" performed on a vertical comb in the hive. The angle between the dancing bee and the perpendicular seems to direct other bees to sources of nectar and pollen.

Rotation receptors. In addition to having tonic statoreceptors (signalling position with respect to gravity), several groups of animals have purely phasic rotation receptors that respond only to angular acceleration or deceleration, as produced on a turntable. Vertebrates, cephalopods (*e.g.*, squid), and decapod crustaceans (*e.g.*, lobsters) have special rotation receptors at the inner surface of the fluid-filled organ of equilibrium (labyrinth or statocyst). This fluid lags inertially with respect to the wall of the organ at the onset and arrest of every rotation. Among crustaceans, such as crabs or lobsters, the rotation receptors incorporate relatively long, delicate hairs that extend more or less at right angles to the wall freely into the statocyst fluid. The hairs respond quickly to fluid motion, swaying around their point of attachment and returning slowly through their elasticity to resting position. Their stimulation causes compensatory reflexes of the eyestalks or of the whole animal.

Eyestalk reflexes can be readily observed when a blinded, legless crab is rotated while flat on a turntable. These reflex movements are called nystagmus. At the onset of rotation to the right, both eyestalks move at about the rotation rate to the left by way of compensation until they reach their maximal deviation. In most cases, one or more jerky movements of the eyestalks in the opposite direction are observed per rotation during the initial period (quick, restoring nystagmus phases). In general, however, the eyestalks remain deviated opposite to the direction of rotation for several revolutions of the turntable. During prolonged constant-velocity rotation, the crab's eyestalks return to their symmetrical position; at this point, inertial lag in the statolymph is reduced to the degree that the fluid finally rotates together with the statocyst wall. Sudden arrest of the turntable under these circumstances causes afternystagmus: the eyestalks move promptly to the right (at about the same velocity as they moved to the left at the onset of rotation) until their maximal deviation toward this side is reached. After a quick jerk in the opposite direction, the eyes continue their slow movement to the right, and in this way as many as three or more after nystagmus jerks may occur with decreasing intensity. Such aftereffects may last many seconds, but finally the eyestalks return slowly to their symmetrical position. All of these nystagmic effects from such horizontal rotation are abolished in a blinded crab, however, after bilateral elimination of the long, delicate statocyst hairs by their denervation or by cauterization of the hair bases.

Rotation of a blinded crab

In vertebrates, rotation reception occurs within the labyrinth. Each labyrinth has three semicircular canals arranged in planes at right angles to each other (Figure 4); the canals communicate with the utriculus. One end of each canal is widened into an ampulla, and the sensory cells (hair cells) are arranged in a row on a ridge (crista) of the ampullar wall. The crista is oriented at right angles to the plane of the canal, and the extended hairs of its sensory cells are imbedded in a jellylike cupula that reaches to the opposite wall of the ampulla. Endolymph displacement through a canal makes the cupula move aside, as if it were a swinging door. In vertebrates, the inertial lag of the endolymph at the onset of rotation is very brief, the fluid catching up with the angular velocity of the labyrinth within a fraction of a second. An ampulla with its crista and cupula is reminiscent of a lateral-line canal neuromast, except that all of the hair cells of a crista are polarized in the same direction. In the cristae belonging to the vertical semicircular canals the kinocilium is implanted at the side facing the canal; in the horizontal cristae all of the sensory cells are polarized toward the opposite side (facing the utriculus). This structural arrangement is in keeping with differences between the vertical and the horizontal canals observed in behavioral and electrophysiological experiments.

A turn of the animal's head around the vertical axis to the left increases the neural-impulse frequency (activation) in the left horizontal crista; a turn of the head to the right causes a frequency decrease (inhibition). Opposite effects occur at the same time in the right horizontal crista. Recordings from the different crista nerves, while the animal is being rotated successively around all three major body axes, show the horizontal crista to respond only to rotation of the animal around its vertical axis; the vertical cristae, however, respond to rotation about all three axes. Behavioral data (compensatory eye reflexes) provide similar results, except that the eyes fail to exhibit an observable response associated with the vertical semicircular canals during rotation around the vertical axis. Stimulation of the vertical cristae under these circumstances gives rise to the simultaneous contraction of antagonistic pairs of eye muscles; hence the absence of a compensatory eye rotation.

In decapod crustaceans, particularly crabs, the statocyst is anything but a simple spherical vesicle; it has a very complicated shape with several curved invaginations and projections. In a small corner in the lowest (most ventral) part of the crab statocyst, a cluster of minuscule sand particles (statoliths) is found in contact with specialized (hooked) hairs. Apart from these hook-hair gravity receptors, there is a single, slightly curved row of relatively straight "thread" hairs atop an oval invagination in the middle of the lower statocyst wall. These hairs are the rotation receptors described above in the blinded crab revolving on a turntable. Bilateral elimination of the thread hairs alters the reflexes of the eyestalks. Instead of reacting immediately, at the very onset of rotation, in the absence of thread hairs on both sides, the eyes initially maintain their symmetrical position. They start their compensatory movement only after rotation has begun or at the end of rotation after the animal has reached a new, steady position. Furthermore, the velocity of this compensatory eye movement seems to be independent of the rate of angular acceleration or deceleration. The delayed nature of the response suggests that loss of rotation sensitivity about horizontal axes results from thread-hair elimination.

That the thread hairs are indeed responsive to rotation about all three major body axes is supported by a number of observations. Bilateral elimination of the impulses from the statolith hairs (position receptors) by selective nerve cutting, for example, does not affect the animal's response to rotation around the vertical axis. Despite their loss of impulses from position receptors, crabs subjected to angular acceleration or deceleration about either horizontal axis exhibit the normal compensatory eyestalk reflexes at the very onset of rotation. When a new (inclined) position of the animal is maintained, the eyestalks again become

Eyestalk reflexes

symmetrical, although complete return to symmetry may require several minutes. On the other hand, when both thread hairs and statolith hairs are eliminated, all such rotation and position reflexes of the eyestalks and related aftereffects are abolished. After unilateral elimination of the thread hairs or removal of one entire statocyst in a blinded crab, both eyes still react to rotation around the vertical axis in both directions. When electrical recordings are made of the activity of the primary sensory neurons innervating the thread hairs, similar results are obtained, the receptors responding only to angular acceleration and deceleration. They are spontaneously active, and the neural response to rotation that is superimposed upon the spontaneous background consists of a coded sequence of impulse-frequency increases and decreases. The same reception unit responds to acceleration about all three major axes.

The statocysts of cephalopods (nautilus, squid, octopus) rival the complexity of crab statocysts. In addition to the perpendicular macula with its statolith (for gravity reception), the octopus has three cristae (containing many hair cells with two-directionally polarized kinocilia) arranged approximately at right angles to each other. Rotation (turntable) experiments and surgical removal of statocyst receptors have shown that the octopus cristae function as rotation receptors. Nystagmus and afternystagmus persist almost unchanged after unilateral statocyst removal, but they are completely abolished after the additional removal of the second statocyst in a blinded octopus. In the cuttlefish (*Sepia*), the statocyst is structurally even more complicated; besides three cristae, it has three maculae (statolith organs) also arranged in different planes.

Rotation receptors of a different type are found in some groups of insects. Dragonflies (for example, *Aeshna*) have external hair receptors between the head and thorax. If a gust of wind turns the animal around its long axis during flight, the relatively heavy head lags with respect to the thorax. The resulting stimulation of the hair receptors in the neck region elicits compensatory flight reflexes and restores the insect to a normal position. These receptors do not respond to static head displacements. In the Diptera (true flies), the posterior knobbed "wings" (halteres) serve as flight stabilizing rotation receptors. During flight, the halteres beat in a vertical plane, synchronously with the forewings. Rotational instability is gyroscopically counteracted by the beating action. Receptors are campaniform sensilla at the base of the haltere. (S.Di.)

Thermoreception

Thermoreception is a process in which different levels of heat energy (temperatures) are detected by living things. Temperature has a profound influence upon living organisms. Active life among animals is feasible only within a narrow range of body temperatures, the extremes being about 0° C and 45° C. On the Fahrenheit scale the same range is 32° F and 113° F. Limitations depend on the freezing of tissues at the lower temperature and on the chemical alteration of body proteins at the higher end of the range. Within these limits the metabolic rate of the animal tends to increase and decrease in parallel with its body temperature.

Body temperature and metabolism among more highly evolved animals (*e.g.,* birds and mammals) are relatively independent of direct thermal influences from the environment. Such animals can maintain considerable inner physiological stability under changing environmental conditions and are adaptable to substantial geographic and seasonal temperature fluctuations. A polar bear, for example, can function both in a zoo during summer heat and on an ice floe in frigid Arctic waters. This kind of flexibility is supported by the function of specific sensory structures called thermoreceptors (or thermosensors), which enable the animal to detect thermal changes and to adjust accordingly.

Temperature of the body directly reflects that of the environment among cold-blooded (poikilothermic) animals, such as insects, snakes, and lizards. These creatures maintain safe body temperatures mainly by moving into locations of favourable temperature (*e.g.,* in the shade of a desert rock). Warm-blooded (homoiothermic) organisms, such as the polar bear, normally keep practically constant body temperature, independent of environment. Homoiothermic animals, including man, are able to control their body temperature not only by moving into favourable environments but also through the internal regulatory (autonomic) effects of the nervous system on heat production and loss. Such autonomic adjustments depend on lower brain centres; the behavioral (movement) responses require the function of the brain's outer layers (the cerebral cortex).

Cold-blooded and warm-blooded animals

A variety of behavioral responses is elicited through stimulation of thermoreceptors, including changes in body posture that help regulate heat loss and the huddling together of a group of animals in cold weather. In some species, thermoreceptors are also involved in food location and sexual activities. Bloodsucking insects, such as mosquitoes, are attracted by thermal (infrared) radiations of warm-blooded hosts; such snakes as pit vipers can locate warm prey at considerable distance by means of extremely sensitive infrared receptors. Man has achieved the widest range of adaptability to extremes in temperature, since his technology allows him to protect himself under a considerable variety of thermal conditions on earth and even in outer space.

Perceptual aspects of thermoreception are found in evidence that man and other animals have conscious temperature sensations and emotional experiences of thermal comfort and discomfort (see below *Human sensory reception*). The effects of temperature on productive efficiency and behaviour (*e.g.,* on one's ability to think) have led to the installation of heat-regulating equipment in homes, public buildings, factories, and similar shelters for people, livestock, and other animals.

Thermoreception can be studied in different ways: (1) on the basis of reports of temperature sensations and thermal comfort by human subjects; (2) through observations of behavioral responses to variations in temperature by all kinds of animals; (3) by the measurement of compensatory autonomic responses (*e.g.,* sweating or panting) to thermal disturbances in the environment; and (4) by recording electrical impulses generated in the nerve fibres of thermoreceptors in laboratory animals and human subjects.

GENERAL PROPERTIES OF THERMORECEPTORS

The concept of thermoreceptors derives from studies of human sensory physiology, in particular from the discovery reported in 1882 that thermal sensations are associated with stimulation of localized sensory spots in the skin. Detailed investigations reveal a distinction between hot spots and cold spots; that is, specific places in the human skin that are selectively sensitive to warm stimuli or to cold. To this extent the different thermoreceptors exhibit sensory specificity. Modern neurophysiological methods show thermoreceptors also to be biophysically specific, in that they include nerve endings that are excited only by or primarily by thermal stimuli.

Early investigations of thermoreception

Extending far beyond the context of conscious temperature sensation as reported by humans, the biophysical definition holds for any thermoreceptive structure. Clearly, electrical responses from thermoreceptors are observable whether conscious sensations are reported by the animal (as in the case of a person) or whether they are not (as in the case of a laboratory rat). Although they are closely related, the concepts of sensory and biophysical specificity are not identical, the criterion being the quality of inner experience (sensation) in the first case and the quality of the neurally effective stimulus in the second. To make the distinction clear, a receptor that is neurally excited by cooling as well as by the application of a chemical (*e.g.,* menthol) might be classified only as a specific (cold) thermoreceptor in terms of human sensation; biophysically, however, it manifestly is a chemoreceptor as well (see below *Chemoreception*).

Most of the modern understanding of thermoreceptors is based on biophysical (electrophysiological) investigations. This approach, introduced in 1936 for recording the electrical signals from single thermosensitive nerve fibres in the tongue of the cat, had been applied by 1960 to similar recordings from single thermoreceptors in the skin of human subjects. Such investigations are made by dissecting single nerve fibres under the microscope and placing them on electrodes or by inserting very fine wires (*e.g.,* tungsten microelectrodes) directly into the intact nerve or receptor. As in the case of other sensory nerve fibres, the electrical signals generated by the activity of thermoreceptors are brief impulses of about one millisecond duration and roughly constant amplitude. They follow in a more or less regular sequence, modulations (changes) in the frequency of which reflect differences in the intensity of the stimulus. (Frequency modulation is widely applied in such devices as radios for information processing.) Sensory structures are called specific thermoreceptors if they respond biophysically to temperature stimuli yet are practically insensitive to such other kinds of stimulation as mechanical pressure.

Cold receptors and warm receptors

The general properties of thermoreceptors in the external parts of the body are found to be similar for any species of animals investigated. Thermoreceptors can be divided into well-defined classes as cold and as warm receptors. At constant temperatures (within an appropriate range), cold receptors are continuously active electrically, the frequency of the steady discharge (static response) depending on temperature. In most cases the static activity reaches a maximum at temperatures between 20° and 30° C (68° and 86° F). On sudden cooling to a lower temperature level, the cold receptors respond with a transient increase in frequency (dynamic response); if the lower temperature is maintained, the frequency drops to a level of static discharge in adaptation. When the receptor is warmed up again, a transient decrease in electrical activity is seen, after which the frequency rises again and finally adapts to the initial static value. Warm receptors are also continuously active at constant temperatures, with a maximum at 41° to 46° C (about 106° to 115° F). On sudden temperature changes, warm receptors respond in the opposite direction from that of cold receptors, temporarily overshooting adaptation frequency on warming and showing transient inhibition on cooling. Thermoreceptors are thus selectively sensitive to specific ranges of temperature as well as to rate of temperature change.

Some receptor cells in the skin of fishes and amphibians respond both to mechanical and to thermal stimulation. In the skin of cat, monkey, and man, receptors have been found that are excited both by mechanical stimuli and by cooling. It seems, however, that these nerve endings are primarily mechanoreceptors (see above *Mechanoreception*), their sensitivity to cooling being much lower than that of specific cold receptors.

THERMORECEPTORS IN INVERTEBRATES

Insects placed on a surface that provides a temperature gradient (warmer at one end and cooler at the other) often congregate in a narrow band of temperature, providing behavioral evidence of sensitive thermoreception. Honeybees (*Apis mellifera*) normally choose a temperature range of 35° ± 1.5° C (95° ± 2.7° F); when repeatedly replaced at the warm end of the gradient, individual bees follow their average chosen temperature within ±0.25° C (±0.45° F). Bees also accurately regulate temperature in the hive between 35° and 36° C (95° and 97° F) by behavioral patterns (*e.g.,* beating wings to circulate air) in the brood season.

Among invertebrates other than arthropods, the leech (*Hirudo medicinalis*) can make temperature discriminations with an accuracy of 1° C (about 1.8° F). The slug (*Agriolimax reticulatus*) reacts at temperatures below 21° C (70° F) by increased locomotor activity in response to 0.3° C (0.5° F) cooling over a period of five minutes.

The temperature sensitivity of bloodsucking arthropods (*e.g.,* lice) is considerably greater than that of nearly all other arthropods; the warmth of the victim's body is the primary influence in stimulating and guiding such blood feeders. The so-called castorbean tick (*Ixodes ricinus*), which sucks blood from sheep, responds when its front legs, which are the primary site of thermal sensitivity, are warmed up by 0.5° C (0.9° F). The bloodsucking assassin bug (*Rhodnius prolixus*) responds with direct movement toward any warm stimuli; *e.g.,* when a glass tube warmed 15° C (27° F) above air temperature is kept within about four centimetres (1.5 inches) of its antennae. Similarly, mosquitoes (*Aedes aegypti*) fly readily to a warm, odourless, inanimate surface as if it were that of a warm-blooded creature. The mosquito's antennae are probably the site of the thermosensors, and the animals manifest sensitivity to changes in air temperature of about 0.5° C. In most insects the thermoreceptors appear to be located in the antennae, since they show impairment of thermoreceptive behaviour when part or all of the antennae are removed. Behavioral studies represent a rather gross method of localizing thermosensitive structures, however. A more direct approach to thermoreceptor function in insects has been achieved by electrophysiological methods. Microelectrodes with tips of very small diameter are inserted near the presumed thermosensitive cells. Any electrical nerve impulses elicited by temperature stimuli are amplified and recorded. This method permits, for example, the study and identification in various insects of receptors that are sensitive to cooling. Cockroaches (*Periplaneta americana*) have two whiplike antennae consisting of 120 to 180 ring-shaped segments that grow thinner and longer with increasing distance from the animal's head. There are about 20 cold receptors per antenna; these are located on the thicker segments (Figure 5), with rarely more than one per segment. Each cold receptor consists of a delicate hairlike structure (sensillum) emerging from a ring-shaped wall. The cold sensilla are mechanically protected by large bristles covering the segments of the antenna. At constant temperatures the cold receptor is continuously active, the average maximum frequency of its discharge being about 16 impulses per second at a temperature near 28° C (82° F). At higher and lower temperatures, the steady frequency becomes lower. When the receptor is rapidly cooled, its discharge frequency rises steeply up to 300 impulses per second and then declines gradually to a much lower constant level. On rapid warming, the opposite response is seen; *i.e.,* there is a transient inhibition of the receptor discharge, followed by a gradual restoration of the steady activity. The cold receptor is thus sensitive to constant temperatures as well as to the rate of temperature changes.

Studies with micro-electrodes

From *Zeitschrift fur Vergleichende Physiologie* (1968)

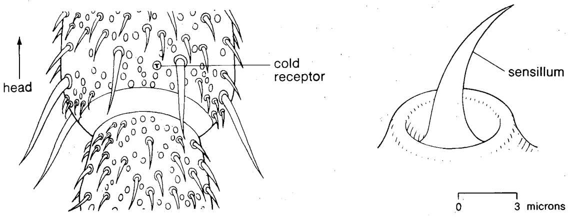


Figure 5: (Left) Portion of cockroach antenna showing location of cold receptor. (Right) Magnification of a single cold receptor of a cockroach.

Caterpillars of various moths (*Lasiocampidae, Saturniidae, Sphingidae*) have cold-receptor cells in their antennae and mouthparts (maxillary palps). Electrophysiological investigations with microelectrodes suggest that just three receptor cells located in the third antennal segment and probably not more than one receptor cell in the maxillary palp are sensitive to cooling. At constant room temperatures, static neural activity from such cells is observed; this activity increases in frequency when the temperature is lowered. During rapid cooling, the frequency rises steeply to a transient maximum of up to 300 impulses per second, while rapid warming produces a temporary inhibition of the discharge. Since only a few cells out of the 20 or 30 that comprise the thermoreceptor structure exhibit the typical electrical response to cooling, specific thermoreceptive function among caterpillars is strongly indicated.

Electrophysiological evidence for the presence of thermosensitive structures also is available for the antennae of honeybees and of migratory locusts (*Locusta migratoria migratorioides*). Temperature-induced changes in the spontaneous electrical activity within the central nervous system of honeybees also have been recorded. Other sensory structures in these animals also can be influenced by temperature, but their primary functions appear to be chemoreceptive or mechanoreceptive.

THERMORECEPTORS IN VERTEBRATES

Fish. Many species of modern bony fish (teleosts) are sensitive to very small changes of temperature of the water in which they live. Various marine teleosts, such as the cod (*Gadus gadus*), have been trained to swim half out of water up a long sloping trough in response to changes of as little as 0.03° to 0.07° C (0.05° to 0.13° F) in the temperature of the water flowing over them.

More detailed conditioning experiments with freshwater fish show that they can distinguish warm from cold, discrimination being made on the basis of thermal change rather than on absolute temperature. Temperature sensitivity persists in these animals when the nerve supplying the lateral line (see *Mechanoreception* above) is cut but is abolished after transection of the spinal cord. When freshwater fish are trained to seek food in response to a change in water temperature, they are found to discriminate differences of less than 0.1° C (0.2° F). Goldfish (*Carassius*) have been trained to discriminate between warm and cold metal rods that have been placed in their tanks. Consistent responses are obtained only when the rod is at least 2° C (3.5° F) colder or warmer than the water. Practically the whole surface of the fish, including the fins, is found to be thermosensitive.

Temperature discrimination in goldfish

This mode of temperature discrimination need not be ascribed to the function of specific thermoreceptors; it could depend on skin receptors that are sensitive to combined mechanical and thermal stimulation. Indeed, electrophysiological recordings from nerve fibres originating in the skin of fish support the latter view. Changes in the electrical activity of these fibres are elicited only when the skin is touched by some solid object; yet the frequency of this mechanically elicited neural discharge is heavily influenced by the temperature of the object used in touching the fish.

From *Journal of Physiology* (1956)

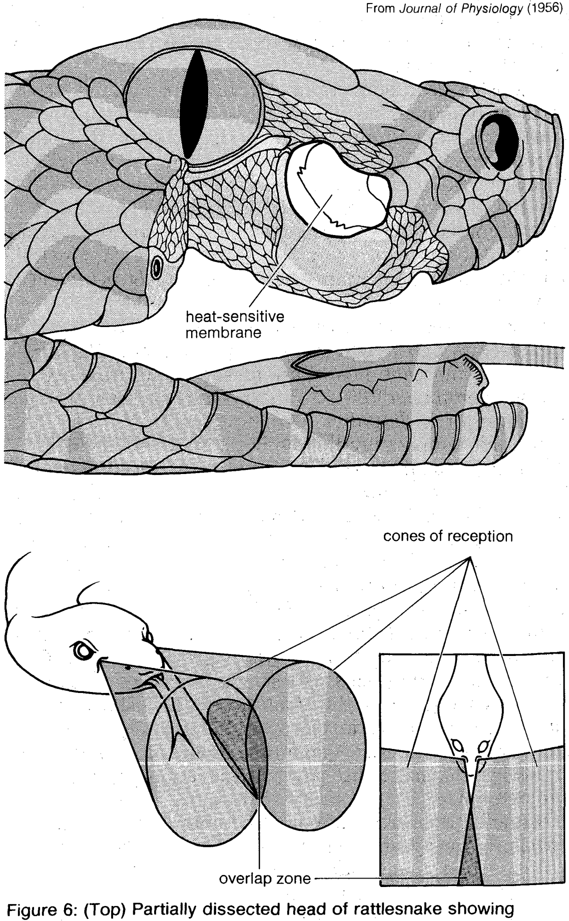


Figure 6: (Top) Partially dissected head of rattlesnake showing heat-sensitive membrane inside pit organ. (Bottom) Cones of reception, directions from which heat energy can be detected (see text).

Elasmobranchs, such as rays and sharks, have distinctive sense organs, called ampullae of Lorenzini, that are highly sensitive to cooling. These organs consist of small capsules within the animal's head that have canals ending at the skin surface. The capsules and their canals are filled with a jellylike substance, sensory-receptor cells being situated within each capsule. Recordings of impulses from single nerve fibres supplying the ampullae of Lorenzini in rays (*Raja*) and dogfish (*Scyliorhinus*) reveal steady activity of the receptors at constant temperatures between 0° and 30° C (32° and 86° F), the average frequency maximum appearing near 19° C (66° F). Rapid cooling causes transient overshooting of the stabilized discharge frequency, while rapid warming produces transient inhibition of the impulses. In some single fibres, cooling by 3° C (5.5° F) leads the frequency to overshoot by about 100 impulses per second. It remains an open question, however, whether the ampullae of Lorenzini are to be called specific thermoreceptors, since they also respond to mechanical stimuli and to weak electrical currents.

Amphibians and reptiles. Rattlesnakes (*Crot alus*) and related species of pit vipers (Viperidae) have a pair of facial pits (Figure 6), sense organs on the head below and in front of the eyes, which are most sensitive thermoreceptors indeed. The pit organs act as directional distance receptors and make it possible for the reptile to strike at warm prey even when the snake's eyes and nose are covered and its tongue has been cut off. Each pit is a cavity about five millimetres deep, equally as wide at the bottom, and narrowing toward the opening at the surface of the head. Inside and separated from the bottom by a narrow air space is a densely innervated membrane of about 10 microns thickness stretching between the walls of the pit. A direct connection between the air space beneath the membrane and the open air maintains equal pressure on both sides of the membrane. Warm-sensitive receptors distributed over the membrane consist of treelike structures of uninsulated (unmyelinated) nerve fibres. Infrared radiation (heat energy) reaches the membrane from an external source through the narrow opening of the pit, permitting the snake not only to detect heat but also to localize coarsely the position of the stimulus. The fields of direction (cones of reception) from which each pit can receive infrared radiation from the environment extend to the front and sides of the head, with a narrow zone of overlap in the middle, as shown in Figure 6.

Under resting conditions, there is an irregular, steady discharge of nerve impulses from the pit organ. Rapid warming by as little as 0.002° C (0.004° F) at the nerve endings elicits a significant increase in impulse frequency; cooling produces an inhibition of the resting discharge. In contrast to the warmth receptors in mammals, the reptile's pit receptors are practically insensitive to steady temperatures, despite their high sensitivity to rate of thermal change. The distinctive consequence in the snake's adaptive behaviour is that gradual variations in air temperature tend to occur without detection, only the more rapid changes in infrared radiation being discriminated. Sensitivity to rapid temperature changes is enhanced by the very limited heat capacity of the receptive membrane (since it is so thin). When an animal that is 10° C (18° F) warmer than the environmental background appears for half a second at a distance of 40 centimetres (about 16 inches) in front of the snake, the heat energy radiated is enough to elevate significantly the frequency of receptor discharge in the pit organ. Indeed, behavioral experiments show that under these conditions the snake is able to discover warm prey through the victim's infrared radiations.

Detection of warmth of their prey by snakes

As poikilotherms, reptiles have practically no internal neural or metabolic mechanisms for maintaining their body temperature within physiologically safe limits. Nevertheless, such reptiles as snakes and lizards are able to keep their body temperature near these safe levels through behavioral regulation (*i.e.*, by moving to cooler or warmer places as necessary). The body temperatures of two samples of lizards (*Sceloporus magister* and *Cnemidophorus tesselatus*), for example, were found to be 34.9° ± 0.6°

C (94.8° ± 1.1° F) and 41.3° ± 0.2° C (106.3° ± 0.4° F), respectively, although the average air temperatures were 33° C (91° F). Highly accurate regulation is recorded for a snake (*Crotalus cerastes*) that moved partially in and out of its burrow into the sun to maintain a body temperature of 31° to 32° C (88° to 90° F) over several hours. The desert iguana (*Dipsosaurus dorsalis*) regulates its body temperature largely by behavioral mechanisms to achieve and hold body temperatures near 38.5° C (101.3° F). These adjustments by iguanas include postural orientation to solar radiation both inside and outside burrows and altered thermal contact of the body surface with the soil. Although supporting direct evidence remains to be more fully developed, it appears that reptiles have thermosensitive nerve structures in the brain as well as in the skin.

There is some electrophysiological evidence of thermal sensitivity among amphibians, but only to relatively large temperature changes. The lateral-line organs in frogs (*Xenopus laevis*), which, as in fish, are sensitive to minute water turbulence, also respond to static temperatures and to temperature changes. Whether these responses have any adaptive, behavioral significance for temperature detection remains to be demonstrated. It has been reported, however, that a frog placed in a pan of cool water will not jump out as the pan is heated, if the temperature changes are gradual enough. Indeed, frogs are recorded to remain in the water this way until they are boiled to death.

Birds. Birds are homoiothermic, normally maintaining their body temperature within a range of less than 1° C (1.8° F). Investigations of temperature regulation in birds suggest the existence of thermosensors both in the lower part of the brain (hypothalamus) and in the skin.

Direct electrophysiological evidence of thermoreceptors has been obtained in the tongues of chickens and in the skin of pigeons by recording from individual fibres of nerves serving the receptors. At a constant temperature of 20° C (68° F), a high level of static activity was observed for cold receptors in the chicken's tongue. When the temperature of the tongue was maintained at 44° C (111° F), individual cold fibres showed a low, steadystate frequency of two to four impulses per second. A temperature drop of 9° C (16° F) was found to elicit an initial response of 30 impulses per second, which gradually declined to a new static frequency of eight impulses per second. Rewarming of the tongue resulted in a cessation of detectable electrical activity for several seconds; no specific warm receptors were found. There is some electrophysiological evidence of cold and warm receptors in the skin of pigeons.

Mound building to regulate nest temperatures

Megapodes, large-footed birds such as the Australian mallee fowl (*Leipoa*), or brush turkey, bury their eggs. They depend on the thermal sensitivity of their face or mouthparts to guide their efforts in controlling the temperature of the eggs during hatching. The eggs are incubated in mounds where heat is generated through the fermentation of rotting vegetation and by irradiation from the sun. For extended periods of time the male bird is busy covering and uncovering the eggs, normally keeping the temperature almost constant at 34° ± 1° C (93° ± 2° F) over the unusually long incubation period (as much as 63 days) that characterizes this family of birds.

Mammals. Detailed information is available from electrophysiological investigations of single thermosensitive nerve fibres in the skin of mammals, particularly cats and monkeys. The nose of a cat contains numerous cold and warm receptors that are highly specific in responding to thermal stimuli; they are not excited by mechanical deformation of the skin. As a rule, each thermoreceptor is connected with a single nerve fibre. By using very finely tipped thermal stimulators, investigators can locate precisely the sites of warm and cold receptors in the skin; the details of the underlying cellular structure at these spots have been studied by electron microscopy. At the site of a cold-sensitive spot in the cat's nose, a thin, myelinated (insulated) nerve fibre penetrates the dermis and divides into several unmyelinated branches about 70 microns beneath the skin surface (Figure 7). The tips of these branches have been shown to be the cold-sensitive nerve endings proper; they come into close contact with the basal cells of the epidermis. In most cases the nerve endings are embedded in small concavities on the lower surface of the basal cells. Warm receptors remain to be identified; they appear to be situated in a deeper layer of the skin. In monkeys and, presumably, in man, warm receptors are innervated by unmyelinated nerve fibres (diameter one micron, impulse conduction velocity of 0.5 to 1.5 metres [1.5 to five feet] per second); cold fibres are served either by unmyelinated fibres or by thin myelinated fibres (diameter two to four microns, conduction velocity three to 20 metres [ten to 66 feet] per second).

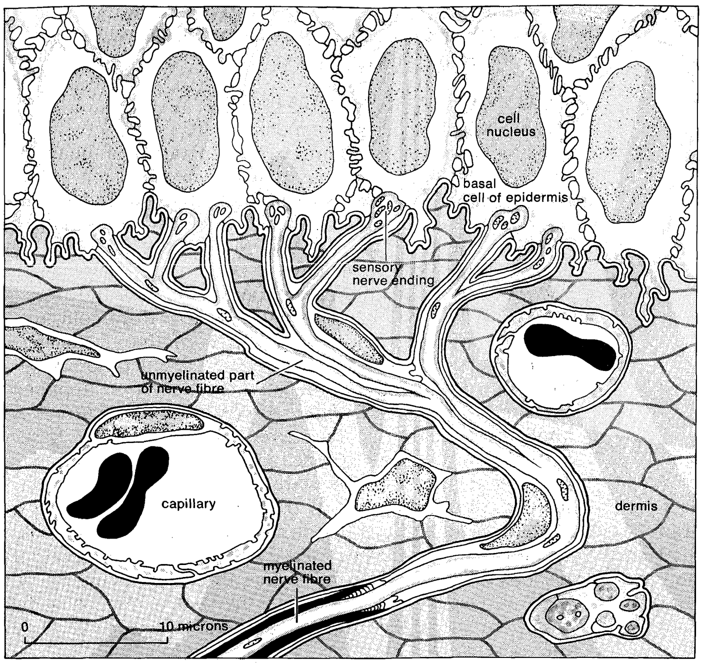


Figure 7: Cold receptor in the skin of a cat as seen with electron microscope.

At constant skin temperatures in the normal range, cold fibres in mammals are found to be continuously active. The average maximum frequency of the static discharge is observed at 27° C (81° F), the extreme limits of the normal range of static activity being 5° and 42° C (41° and 108° F). At skin temperatures above 45° C (113° F), cold receptors again can be activated. This so-called paradoxical discharge corresponds to the similar paradoxical sensation of cold in man when a hot object is touched or the hand is put into hot water. On sudden cooling, temporary overshooting can be 30 times higher than the static frequency of eight impulses per second. Warm receptors in the cat's nose start to show their activity at skin temperatures of 30° C (86° F) and reach an average maximum of 35 impulses per second at 46° C (115° F). Above this temperature the discharge suddenly falls off.

Similar populations of cold and warm receptors have been found in monkey skin, which has an additional group of warm fibres distinguished by an average static maximum at 41° C (106° F). The transient overshooting of warm-receptor activity can be several times higher than the static maximum frequency. When cold stimuli of the same magnitude (*e.g.*, 3° C [37° F]) are rapidly applied to the warmer skin, the degree of overshooting reaches a maximum at a skin temperature of 27° C (81° F). This temperature corresponds to that which elicits maximum static discharge from the cold fibres. Overshooting on rapid warming of warm receptors follows the same general rule: the temporary maximum occurs at temperatures for which the static discharge frequency also appears at a maximum.

Distinctive properties of cold receptors are found in hibernators. European hamsters (*Cricetus cricetus*) tend to maintain body temperatures of about 5° C (41° F) during hibernation. Further cooling, however, elicits arousal reactions from the animal that indicate thermoreceptive function is intact. Electrophysiological investigations have shown that myelinated cold fibres serving the hamster's nose are continuously active at very low temperatures, having a static maximum near 4° C (39° F). In contrast to these findings, myelinated cold fibres in mammals that do not hibernate are blocked at temperatures below 10° C (50° F).

Relatively little is known about the processing of information from skin thermoreceptors in the central nervous system (brain and spinal cord). Responses to cooling the tongue have been recorded from single nerve cells (neurons) of the brain's thalamus in monkeys. In addition to a few brain neurons that are excited both by mechanical stimulation and cooling of the tongue, there are also numerous nerve cells in the thalamus that respond only to cooling. The latter neurons exhibit a static discharge in the temperature range (at the tongue) between 15° and 44° C (59° and 111° F), the maximum frequency being at 21° to 31° C (70° to 88° F). Rapid cooling of the tongue causes considerable transient overshooting in frequency from thalamic nerve cells over the entire range of preadapting temperatures used; by contrast, warming of the tongue results in a transient inhibition but not an increase in the rate of a discharge of any nerve elements in the brain. Thus, the activity of thalamic nerve cells activated by cooling the tongue closely reflects the behaviour of the peripheral (*e.g.,* tongue or skin) cold receptors. Even in the outer layers of the brain (cerebral cortex), nerve cells that respond specifically to cooling of the skin have been found. Other individual cortical neurons in the cat, however, receive signals not only from peripheral thermoreceptors but from mechanoreceptors and taste receptors as well.

Thermal information processing in the monkey brain

Cats can be trained to respond behaviorally to thermal stimulation (*e.g.,* they will press a bar or lever). These experiments reveal cats to be relatively insensitive in discriminating warm and cold stimuli applied to the furred skin of the trunk or the legs, requiring temperature differences amounting to several degrees Celsius. This does not necessarily mean that cats have no finer thermal sensitivity than their grossly observable behaviour suggests. Indeed, autonomic regulatory responses, such as increased blood flow to the ear, can be elicited by mild warming of the paws, even though such warming is inadequate as a signal for training behavioral responses. In contrast to the low thermal sensitivity of their skin elsewhere in the body, cats have been found to manifest behavioral responses when the nose and upper lip are warmed or cooled by only 0.1° to 0.2° C (0.2° to 0.4° F). This corresponds to levels of thermal sensitivity of the face in human subjects and also is in accordance with electrophysiological evidence of high thermal sensitivity of the cat's nasal region.

Mammalian thermoreceptor structures, each containing elements sensitive to warm and cold, are to be found in the skin, in the deep tissues of the body, and in the hypothalamus and spinal cord. While much of the evidence of thermosensors in the central nervous system derives from experiments in neural temperature regulation within the body, in 1963 microelectrodes were used to record directly the activity of thermosensitive neurons in the frontal part of the cat hypothalamus. Many investigations made with this method show that most neural elements in the cat and rabbit hypothalamus are practically insensitive to directly applied temperature stimuli. There are, however, two smaller populations of thermosensitive neurons, one of which responds to local warming and the other to local cooling of the brain tissue.

Warm-sensitive brain neurons in cats and rabbits increase the frequency of their static impulse discharge in direct proportion to the degree that hypothalamic temperature is raised above normal. Similarly, the cold receptors there respond with an increase in impulse frequency when the temperature of the hypothalamus falls below the normal value for the animal's body. Since temperature in the deeper tissues of the body (*e.g.,* the brain) varies quite slowly, the activity of hypothalamic thermosensors seems to be almost entirely a function of the level of temperature alone and not of the rate of temperature change. By contrast, the thermoreceptors in mammalian skin are highly sensitive to rapid changes in temperature. Intervening elements in the nervous system have been identified that integrate temperature signals from the hypothalamus and from the skin. Thermosensors are also found to be localized in the midbrain of rabbits and in various parts of the spinal cord in guinea pigs and dogs. These findings are in good agreement with observations that thermoregulatory responses such as shivering or panting can be influenced by local temperature changes in the spinal region of the animal.

Signals from hypothalamic thermosensors, from deep-body thermosensors, and from those in the skin are integrated in thermoregulatory centres located mainly in the mammal and bird hypothalamus. The integrated signals provide information about the inner (core) temperature of the body and about thermal changes at the periphery (body surface). Such information serves to activate internal mechanisms that maintain body temperature within the normal range of values. When signals from warm receptors (especially from those in the hypothalamus) prevail over signals from cold sensors, such heat-loss mechanisms as sweating, panting, and widening of blood vessels (vasodilatation) in the skin act to reduce body temperature. When signals predominate from cold receptors (particularly from those in the skin), heat-conservation mechanisms are initiated. Heat production rises as muscles expend energy in shivering and through other metabolic reactions (nonshivering thermogenesis); heat loss is reduced by mechanisms that narrow the blood vessels (vasoconstriction) in the skin and that fluff out hairs or feathers to enhance thermal insulation. All these involuntary, or autonomic, regulatory changes continue even without the function of the cerebral cortex; thus, they do not require consciousness, persisting during light anesthesia or during sleep.

Sweating, panting, and shivering

In human beings, the function of thermosensors is closely involved in the highly emotional experiences of thermal comfort and discomfort. Whereas temperature sensations are mainly related to the activity of warm and cold receptors in the human skin, thermal comfort and discomfort reflect the general state of the thermoregulatory system, involving signals not only from thermoreceptors in the skin but from thermoreceptors in deep-body regions and in the hypothalamus as well. Thus, the same temperature at the skin can be experienced as comfortable or uncomfortable, depending on the thermal condition of the person's whole body. When one is overheated, an ice bag applied to the head may be perceived as pleasant; but, if someone is generally chilled just to the point of shivering, the same cold stimulus can be most unpleasant. (H.He.)

Chemoreception

All animals react to chemicals in the environment, initially through a sensory process called chemoreception. The process begins when chemical stimuli come in contact with chemoreceptors, specialized cells in the body that convert (transduce) the immediate effects of such substances directly or indirectly into nerve impulses. A nerve cell (neuron) that makes a direct conversion is called a primary receptor; a cell that is not a neuron but that responds to stimulation by inducing activity in an adjacent nerve cell is called a secondary receptor.

CLASSES OF CHEMORECEPTORS

In man two distinct classes of chemoreceptors are recognized: taste (gustatory) receptors, as found in taste buds on the tongue; and smell (olfactory) receptors, embedded high in the lining (epithelium) of the nasal cavity. These respond to different classes of chemicals: gustatory receptors to water-soluble materials (*e.g.,* salt) in direct contact with them and olfactory receptors to generally water-insoluble, vaporous materials that may arise from a distant source, such as a neighbour's kitchen. The receptors themselves are also different; gustatory receptors are specialized epithelial cells (secondary receptors) with neurons branching among them, while olfactory receptors are nerve cells (primary receptors) with fibres leading to the brain.

In all air-breathing vertebrates (*e.g.,* reptiles, birds, and mammals) the two classes of chemoreceptors are easily identifiable. In fish gustatory organs are on the fins and even the tail, as well as in and near the mouth, all still

Fish that taste with their tails

recognizable as taste buds. The nostrils in fish do not usually open into the mouth, but they are lined with olfactory epithelium. Much lower concentrations of chemicals are needed to elicit responses in fish for smell than for taste. These concentrations are similar to those for air breathers, permitting separate identification of the chemical senses for aquatic and terrestrial vertebrates.

For some invertebrates (*e.g.,* worms), however, distinctions between taste and smell receptors may not emerge. Chemoreceptors of these animals are structurally different from those of vertebrates, and their locations on the body are different. It has been held that invertebrate animals have only one chemical sense, with different sensitivities for various chemicals, as measured by the lowest concentrations (thresholds) of chemicals that can be received. Terrestrial invertebrates, particularly insects, do exhibit separable chemoreceptive capacities, however; additional study seems likely to reveal similar distinctions for other invertebrates. For these animals, the terms distance chemoreceptors and contact chemoreceptors are preferred by many biologists over the terms (*e.g.,* smell and taste) used in human physiology. Separation of these seems feasible because contact chemoreceptors are usually stimulated by nonvolatile, water-soluble chemicals, while distance chemoreceptors typically respond to volatile, oil-soluble chemicals. In addition, thresholds for stimulation of distance chemoreceptors are usually very much lower than those for contact chemoreceptors. Generally the behavioral results of contact chemoreception are feeding, mating, or the deposit of eggs, while those of distance chemoreception are orientation or movement of the animal toward or away from a volatile chemical.

Aquatic animals and terrestrial species with mucus-secreting skins are generally sensitive to chemicals all over the body, reacting with avoidance. This sensitivity has been called the common chemical sense. Man and other terrestrial vertebrates have a remnant of this receptor system that responds to irritants in the mucous membranes of the mouth, eyes, and genital organs. Common chemical receptors are thought to be free nerve endings (branching structures, or dendrites, of nerve cells) in the skin or in moist membranes. Even on the basis of relatively few studies, the common chemical sense is known to be separable from the sense of pain, and thus it is considered as a separate sensory capacity.

Receptors for humidity, particularly well studied in insects, may or may not be chemoreceptors. There is no question that some animals can orient toward or away from regions of high or low atmospheric humidity. The question is whether this is true hygroreception (*i.e.,* stimulation of the receptor by moisture-saturation deficit) or is stimulation by water acting as an odorous chemical. While the matter is far from settled, it seems that some insects and possibly mammals actually may be able to smell water, while others have true hygroreceptors.

Flavour

In common speech the word taste refers to what is more correctly designated as flavour. For man, flavour sensations represent integration by the central nervous system (*e.g.,* the brain) of a complex of stimuli: gustatory, olfactory, common chemical, tactile, thermal, even painful. When carefully studied in other species (*e.g.,* a few other mammals and a few insects), reactions to foods seem to be similar to those of man, with multidimensional stimulation involved in food preferences.

ADAPTIVE FUNCTIONS OF CHEMORECEPTION

For most animals, chemical stimuli are leading sources of information about the environment; even man relies heavily on chemoreception for food selection. Species identification, mate finding, courtship, and mating are also chemically directed among most animals.

Food procurement. Foods are generally located by reception of odours they emit, sampled for palatability by both contact and distance chemoreception, fed upon only if they supply appropriate chemical stimuli during feeding, and laid aside either when the animal is full or when the animal's threshold of response for the stimulating chemical rises above the intensity of stimulation provided by the foods.

At least four classes of chemicals are recognized that affect feeding behaviour: (1) attractants: odours eliciting movement *toward* the source; (2) repellents: odours that prompt the animal to move *away* from the source; (3) feeding stimulants (phagostimulants): tastes and odours that induce the animal to feed; and (4) feeding deterrents (antifeedants): tastes and odours that inhibit feeding behaviour. Chemicals in foods that attract animals or that induce feeding are not necessarily nutritionally valuable in themselves; in food plants, the stimulants often are so-called secondary plant substances (*e.g.,* odorous essential oils) that provide little nourishment. Among animals that are preyed upon as food, the stimulants are often traces of odorous materials present on the body surface. Indeed, animals will feed on nutritionally worthless materials that have been experimentally impregnated with appropriate phagostimulants. Ordinarily, however, specific feeding stimulants are part of an animal's natural food (see also BEHAVIOUR, ANIMAL: *Feeding Behaviour*).

Symbiotic relationships. Most parasites do not just blunder onto their hosts but, rather, orient themselves toward suitable animals or plants. Little is known about the guiding stimuli for most parasites, but for some the odour of the host acts as an attractant, and the taste of the host's body surface functions as a feeding stimulant. Parasitic wasps that lay their eggs on wood-boring insects, for example, locate their targets in logs through olfactory signals. The wasp then drills into the log with a complex egg-laying structure (ovipositor) on the end of which are contact chemoreceptors that allow the insect to sample the prospective host to determine whether or not it is already parasitized. Animals that establish nonparasitic (mutualistic or commensalistic) relationships also find each other by chemical clues; or at least the mobile member of a pair finds the nonmobile member through chemoreception. Sea anemones that attach themselves to shells housing hermit crabs, for example, detect the proper shells with contact chemoreceptors on their tentacles. Annelid worms that are commensal (feeding together) with starfish or sea urchins to which they cling locate the latter by chemicals given off by the hosts.

Chemoreception among parasitic wasps

Communication. Many animals release chemicals that influence other individuals behaviorally or at least physiologically. Usually produced by glands, these chemical communication signals have been named pheromones because they seem to act somewhat like hormones inside an animal's body. Females of some moths, for example, produce scents that attract males from great distances (a behavioral effect). Queen honeybees give off a chemical (so-called queen substance) that suppresses ovarian development in worker bees (a physiological effect). Basically the general classes of information that are coded in chemical signals are concerned with species or individual identification, with social communication, and with sexual or reproductive activity.

Ovarian suppression in bees

In aggregating as groups or in dispersal, animals depend on their ability to identify species or individuals. Thus, honeybees scent-mark their own hive and areas around it with odours that uniquely identify that particular insect community for its members. Many mammals are individually territorial, marking the boundaries of their territories with special glandular secretions (*e.g.,* deer), with body odours (*e.g.,* bears), or with urine (*e.g.,* dogs).

Chemical signals facilitate cooperation among social insects and many mammals. When their colony is endangered, for instance, ants, bees, and wasps alert the group with alarm odours. They also deposit chemicals that serve as guidance signals to indicate the way to sources of food or to living quarters.

Most of the sexual signals that animals produce at all stages of mating are chemical. Females of many mammalian species, for example, produce specific odours that attract only males of the same species. Male bumblebees mark leaves or sticks with a scent that induces females of their species to tarry for mating. In many species mating itself is stimulated in one or both sexes by special chemicals produced by the partners. Male tree crickets, for instance, produce a glandular secretion on which the female feeds during mating.

Orientation. Besides being oriented toward or away from food or mates, many animals are guided to suitable habitats by chemicals emanating from plants or from other environmental features. Fish such as salmon, which return from the ocean to lay their eggs in fresh water, generally come back to the specific stream where they themselves were hatched, guided by the odour of the stream. Other fish recognize their nesting areas by odours produced by plants in the vicinity.

Protection against predators. A most effective form of chemical protection is found in marine slugs and snails that produce strong acid secretions when disturbed. These secretions can injure other animals. Many species of animals produce chemicals that are repellent without necessarily being dangerous; for example, stinkbugs, millipedes, skunks, and some earthworms produce strongly smelling or bitter-tasting secretions when disturbed. An animal that causes a predator to become ill long after contact is not thereby directly protected. If the prey has a special taste or smell, however, the predator that samples it and later sickens learns to avoid the taste or smell, thus sparing other members of the species upon which it might otherwise prey.

Repellents of stinkbugs and skunks

CHEMORECEPTORS IN LOWER INVERTEBRATES

Detailed evidence of chemoreception is available for only insects and mammals. Indeed, chemoreception has been studied in depth for only three or four species of insects and four or five species of mammals. For most animals data for secure generalizations are lacking.

Protozoa. Protozoans, even though they are single-celled, behave as if they had a nervous system. They are sensitive to chemicals in the environment and usually select some foods in preference to others. Carbon dioxide dissolved at low concentrations attracts many protozoans and may be the agent that leads them to foods. Some protozoans (*e.g., Spathidium*), however, can locate specific foods at a distance, presumably by a chemical sense. Ciliates (*e.g., Paramecium*) are most sensitive to chemical stimulation at the anterior (front) end; the receptors are probably special cilia (hairlike structures). *Paramecium* takes nonfoods, such as carmine particles, but soon "learns" to stop this, the change in behaviour persisting for some days. In some ciliates (*e.g., Vorticella*) that reproduce by exchanging genetic material between individuals (conjugation), a motile partner (conjugant) swims to a stationary individual. The swimmer is attracted from up to a millimetre away by a chemical produced by the fixed partner. All of these behaviour patterns performed by only one cell are nevertheless similar to those of multicellular animals.

Cnidaria. Chemoreception is doubtless the most crucial receptive capacity of cnidaria (*e.g., Hydra* and jellyfish), but little is known about the organs involved. Sensitivity to food chemicals is greatest near the mouth and tentacles, but specialized organs remain to be described. Almost all receptors are free nerve endings in the integument (body surface). *Hydra* exhibits feeding behaviour when stimulated by such chemicals as reduced glutathione or tyrosine. This reaction occurs in about half of the tests with weak solutions (1×10^{-6} molar) of these substances. Reduced glutathione acts similarly on the Portuguese man-of-war (*Physalia*) and some other coelenterates called marine hydroids. Amino acids other than tyrosine induce a feeding response in some coelenterates: valine and glutamine in sea anemones and proline in some hydroids and corals.

The feeding sequence of coelenterates is highly coordinated, despite the presence of only a very primitive kind of nervous system called a nerve net. Contact with food causes discharge of stinging or entangling structures (nematocysts), the reaction being released by a combination of chemical and tactile stimuli. The tentacles then draw the prey into the mouth. This response may be evoked by release of glutathione or amino acids from the injured prey.

Other behaviour patterns of coelenterates have been little studied. Anemone fish (*e.g., Amphiprion*) live safely among the tentacles of sea anemones that kill other fishes. Seemingly the mucous coat of the anemone fish develops a chemical that inhibits the discharge of nematocysts, although other interpretations of observations made so far are possible. Many marine coelenterates that live in immobile groups shed sperms or eggs (depending on their sex) synchronously, the activity probably being regulated by chemicals given off by some individuals that trigger discharge in others. A swimming sea anemone, when touched by a starfish that feeds upon it, releases its hold and swims away. Identification of the predator starfish is specifically chemical. Reactions of coelenterates to chemical stimuli are far from stereotyped, a wide range of responses being observable.

Platyhelminthes. Flatworms (Platyhelminthes) have two major life-styles—free-living (turbellarians) and parasitic (tapeworms and flukes)—and their reactions vary accordingly.

For some free-living flatworms (*e.g.,* freshwater planarians) the locations of chemoreceptors in the body are known, but their structure is not. Planarians locate foods at a distance, and their behaviour during this process indicates that earlike protuberances (the auricles) on the head bear the receptors. Water currents elicit orientation movements, the animals crawling upstream when thus stimulated, as if they were making an olfactory response. Removal of a structure called the auricular groove abolishes planarian responses to foods; the receptor organs in the groove are thought to be ciliated glandular patches of nerve cells. Upon reaching food, the worm makes contact with its anterior end and with the tip of its pharynx (proboscis). Ingestion then may or may not occur, the reaction resembling selective taste (gustatory) responses of other animals. The tip of the worm's proboscis has receptors; indeed, an isolated pharynx cut away from the rest of the body will feed on appropriate foods.

Flatworms have been experimentally subjected to stimulation with many pure chemicals, most at concentrations not likely to be encountered in nature. The animals are usually attracted by relatively weak solutions and repelled by high concentrations. They respond to natural food juices and experimentally to pure amino acids and their derivatives. A worm called *Dugesia* reacts positively to such chemicals as lysine and glutamine, negatively to aspartic acid, asparagine, and α-keto-glutaric acid, and gives no observable response to hydroxyproline and glutamic acid. Planarians of different species, when mixed together in the same tank of water, can be separated by species through differences in their chemical-recognition behaviour. These distinctive chemically mediated reactions indicate well-developed sensory function for the planarian nervous system.

Reactions of flatworms to pure chemicals

Little evidence is available about chemical sensitivity among tapeworms and flukes. Tapeworms are said to have only tactile organs, but supporting evidence is almost nil. Adult flukes obviously find their way to specific organs in the bodies of animals they parasitize, but the sensory mechanisms are unknown. The free-swimming stages (miracidia and cercariae) in the life cycle of flukes find their hosts effectively, but there is no general agreement on how this is done. Some workers hold that they swim at random and enter whatever body they encounter; others say that the flukes swim at random but select the host on contact; still others claim that they orient toward the host before contact. Perhaps different species of flukes vary in their behaviour, but the evidence is too sparse to draw general conclusions.

Nematoda. For a phylum with so many commercially and medically important parasites (as well as free-living species), the lack of studies on chemoreception in roundworms (nematoda) is surprising. The integument of these roundworms is supplied with many types of receptors, mostly free nerve endings. These are concentrated anteriorly, particularly on structures around the mouth called papillae. Nematode papillae could be chemoreceptors, but the possibility is supported by no direct evidence. Some roundworms have specialized glandulo-neural structures (amphids at the anterior end of the body and phasmids at the posterior end) that have been claimed to be chemoreceptive, again without critical verifying evidence.

Except for nematodes that parasitize plants, no agreement has been reached on how these animals find their hosts or foods or how they form "social" aggregations, as some free-living species of roundworms do. Parasitic nematodes may attack the roots of plants in response to a chemical attractant in the roots. In some cases the attractant is found to be carbon dioxide that stimulates the worms at a distance, with some other chemical acting on contact. The possibility that control of some agriculturally destructive pests may be achieved by changing the chemical environment in the soil is drawing increased attention to behavioral studies of these nematodes.

Echinodermata. These marine animals (*e.g.*, starfish, sea urchins, sea cucumbers) have also been little studied. They are generally sensitive to chemicals, seemingly most acutely at the tips of their myriad tubular "feet" (podia). Only free nerve endings are present in the integument (skin) of most echinoderm species, but sea cucumbers have sensory pits on their tentacles with more specialized nerve endings. The concentration of primary sensory cells in the integument of many echinoderms is truly striking, upwards of 4,000 per square millimetre (2,600,000 per square inch) being reported for certain starfish. These endings may be multisensitive (to a number of chemicals), or they may be functionally differentiated although structurally they appear to be identical.

The chemoreceptive "feet" of starfish

Reports of studies of chemical reactions among echinoderms are few and spotty. These animals respond positively to natural foods and to some food chemicals (such as glutamic acid) at a distance, and they feed on specific items on contact. They avoid harmful chemicals (*e.g.*, injurious acids and salts). They also form specific aggregations, possibly through chemical responses to their fellows, and are known to spawn synchronously as a result of chemicals released during the process.

Annelida. Annelids (*e.g.*, leeches and earthworms) are sensitive to chemicals all over the body; they are selective in feeding, but no specialized chemoreceptors are yet known for them. Three types of nerve endings in the skin of these animals have been claimed to be chemoreceptive, but without direct evidence: (1) primary sensory cells concentrated at the anterior end, up to 700 per square millimetre (450,000 per square inch) in front of the mouth (on the prostomium) of an earthworm; (2) branching free nerve endings in the skin, possibly mechanoreceptors rather than chemoreceptors; and (3) special concentrations of nerve endings, called integumental sense organs. Some "hairy" marine annelids (polychaets) have a so-called nuchal organ near the head, ranging in complexity from a simple ciliated pit to an elaborate set of folds covering many of the ringlike segments (somites) that form the body. The nuchal organ has been reported as chemoreceptive, but no direct evidence has been produced.

Chemoreception among annelids has been studied mainly by dipping them into or flooding them with various solutions and noting withdrawal or by feeding them natural and man-made foods. The animals respond appropriately, so that thresholds for eliciting responses have been determined. What these mean in the lives of the worms is generally obscure; as usual, low concentrations of many substances are accepted or produce positive responses, whereas high concentrations are rejected or repel. Studies of nerve impulses picked up from receptors in the skin of the body wall have been made with earthworms. The receptors, still unidentified, produce impulses when stimulated with appropriate concentrations of table salt, quinine, and acids, but they fail to respond to ordinary sugar (sucrose). The prostomium, however, does have receptors that are sensitive to sucrose solutions.

Feeding, selection of places on which to settle by some marine annelids, and selection of soil by earthworms have been shown to be chemically mediated. Commensal polychaetes (*e.g.*, *Podarke*) distinguish the organisms with which they live through chemicals coming from their hosts. Synchronous spawning occurs in many anchored (sessile) marine worms, being mediated through the release of signal chemicals. Release of sperms by breeding males of *Platynereis*, a swimming marine polychaete, requires chemical stimuli from the female. Earthworms incorporate an alarm chemical in the mucus given off when they are roughly handled; the effect is to repel other earthworms for as long as several months thereafter.

Mollusca. More information about chemoreception among mollusks (*e.g.*, snails, clams, squids) is available than there is for the groups discussed so far; but these animals comprise a large phylum, and very few species have been studied.

Chemical sensitivity is generally distributed over the mollusk's body, being greatest at the mouth, tentacles, front of the foot, and along the edge of its thin, capelike mantle. The receptors, although not identified with certainty, are thought to be variously branched free nerve endings. Body regions known to be most sensitive to chemicals have high concentrations of these cells. These regions are: (1) tentacles—a variety of projections on various parts of the body; (2) osphradia—ridges or projections near the front of the mantle cavity, best studied in marine gastropods (*e.g.*, snails and slugs); (3) abdominal receptors at the base of the siphons in bivalves (*e.g.*, oysters and mussels); and (4) olfactory pockets behind the eyes in cephalopods (*e.g.*, octopuses and nautiluses). Other organs have been designated as chemoreceptors, but with no critical evidence: (1) so-called subradular organs in the mouths of lower mollusks; (2) a structure called Hancock's organ in some gastropods; and (3) rhinophores (once identified as "olfactory" tentacles) of some gastropods called opisthobranchs. The last, however, are almost certainly established as receptors for water currents rather than as chemoreceptors.

Blinded octopuses

Most of the physiological studies with mollusks have been on reactions to food or to foreign chemicals. Octopuses have been blinded and then trained or conditioned to respond to pure chemicals with specific behaviour patterns. Studies of orientation to or acceptance of feeding stimulants have shown that tentacles and osphradia bear receptors for odorous materials and that receptors near the mouth initiate feeding. Thus separation of contact from distance chemoreception among these animals seems probable; but, until specific receptors are identified through their nerve impulses, the distinction remains conjectural. Although nerve-impulse studies have been made with at least two gastropods (*Aplysia* and *Buccinum*), specific receptors have not been identified thus far. The osphradium has finally been shown to bear chemoreceptors (a matter long debated), and reactions to food extracts and chemicals in natural foods have been studied.

Location of food or prey by many species of mollusks involves what suggests distance chemoreception, generally through the tentacles. Some carnivorous land snails detect and follow (by "tasting") the slime trail left by the prey. Specific "social" aggregations are common among marine bivalves; some of these are brought about by the settling of bivalve larvae near chemically detected members of their group (conspecifics). Chemically regulated synchronous spawning is common among marine mollusks. Land snails and slugs find mating partners by following their slime trails by "tasting" them. Limpets and other snails that live close to the shore emerge to feed when seawater splashes on them at low tide; the sense organs involved differentiate seawater from rain.

Many bivalves and gastropods react strikingly to chemicals from their predators. Herbivorous marine snails, for example, move rapidly away from predators as soon as they touch them. A freshwater snail (*Physa*), when touched by a leech, swings its shell back and forth and then drops to the bottom. These reactions are induced by specific chemicals; the skin of echinoderms, for instance, has yielded such a material, the extract being found to resemble a group of chemicals called saponins.

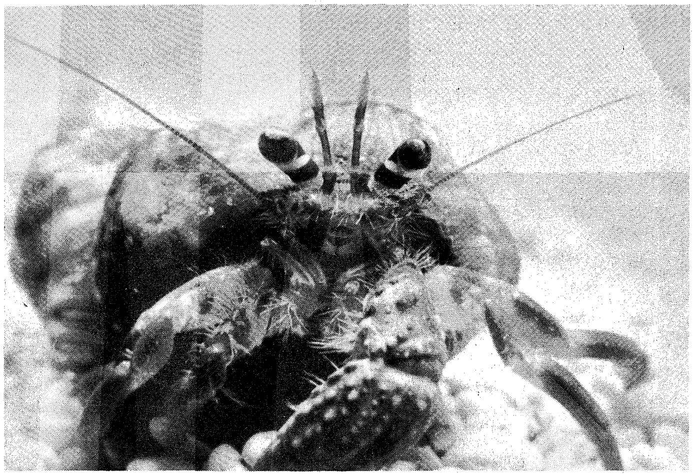
ARTHROPOD CHEMORECEPTORS

In the Arthropoda, which includes more than two-thirds the total number of all individual animals alive, detailed chemoreceptive studies have been reported for less than 10 species of insects and five species of crustaceans; reliable information about other arthropods (*e.g.*, sow bugs and centipedes) is rudimentary. Many of these latter animals have hairs on their outer surface (exoskeleton) that may be chemosensory, since they are similar to those known to be chemoreceptive in insects and crustaceans.

Responses to food and mates, supposedly chemically mediated, have been described for millipedes, centipedes, and a number of arachnids (*e.g.,* spiders). Electrophysiological studies of chemoreceptors have been made with the horseshoe crab (*Limulus*) found on many beaches. The receptors are in spines on the legs and chilaria (flaps behind the mouth) of the animal. Each sense organ has from six to 15 nerve cells that respond or fire when bathed in clam juice or in solutions of amino acids. A tick (*Ornithodoros*), when fed through an artificial membrane, accepts glucose solutions with such substances as reduced glutathione, adenosine triphosphate, and nicotinamide-adenine-dinucleotide; glutamic acid inhibits feeding behaviour in this arachnid. Among some wandering spiders, the male locates the female by the scent of her silken dragline, which serves to identify species and sex. Contact chemoreceptors at the tips of the spider's legs are the sensitive structures. These observations represent a good sample of the scattered work to date with arthropods other than insects and crustaceans.

Chemoreceptors of horseshoe crabs, ticks, and spiders

Crustacea. Crustaceans include such arthropods as crabs, lobsters, shrimps, barnacles, and many other forms. For a number of crustacean species, reactions to food chemicals or other substances have been used to locate the body regions that bear chemoreceptors. The list is impressive. Distance chemoreceptors are borne on the antennae and the smaller antennules, specialized structures (esthetascs) on the tips of the antennules being particularly sensitive (Figure 8). Contact chemoreceptors are borne chiefly on the tips of the walking legs, the mouthparts, antennules, tail flap (telson), walls of the gill chambers, and, in some species, on the general body surface.



Hubert Frings

Figure 8: Hermit crab in shell, showing antennae (long and thin) and antennules (held vertically between eyes) with esthetascs (specific chemoreceptors) along edges near tips.

Locations and structure of chemoreceptors. The sense organs in these regions are various, but only the esthetascs have been shown electrophysiologically to be chemoreceptive. Scattered over the body are so-called funnel canals (or pore organs), which are assumed to mediate avoidance reactions to high concentrations of chemicals. Also widely distributed over the body is a variety of hairlike structures that are similar in appearance to known chemoreceptors of insects. Short blunt projections, resembling certain specialized receptors (basiconic sensilla) of insects, on the body wall of terrestrial isopods (*e.g.,* wood louse or pill bug) are also assumed to be chemoreceptive. The esthetascs at the tips of the antennules are groups of hairlike or spinelike structures. Receptors in these produce nerve impulses when stimulated with a variety of chemicals. Each esthetasc hair receives 100–500 nerve endings from cells aggregated in a ganglion-like structure at its base. The nerve endings have a cilia-like pattern of fibrils, characteristic of the primary chemoreceptors of insects and vertebrates. The outer layer (cuticle) of the esthetascs is very thin, but it has no openings through it, as does the cuticle of the sensory hairs of insects.

Most studies on chemoreception among crustaceans have been made on a few species of crabs and crayfish, with food selection or reactions to chemicals as indicators of reception. Tests before and after removal of parts of the body have led to the discovery of the chemoreceptor locations. There have been a few recent electrophysiological studies with only a very limited number of species.

Responses. In general, crustaceans respond to a wide range of chemicals, negatively at high concentrations and positively at low. In many species, although the body regions that bear chemoreceptors have only one structural type of sensory hair, reactions to different chemicals vary. The antennae of crayfish, for example, have only one distinguishable type of hair, yet the antennae have distance chemoreceptors functionally resembling those of insects and vertebrates, as well as contact chemoreceptors. This has led some to suggest that there is no differentiation between "taste" and "smell" in these animals, merely differences in thresholds. Nevertheless, the behaviour patterns of crayfish stimulated by different classes of chemicals are different. Receptors in the antennules of a shrimp (*Crangon*) respond electrophysiologically to coumarin (usually considered an odour substance) at concentrations of 0.0001–0.00005 percent, to salt (NaCl) at 1.3–7.2 percent, to acetic acid at 0.01 percent, and to quinine chloride at 0.001–0.0005 percent. The observed differences are sufficient to put coumarin in a separate ("smell" or distance) class from the other (contact or "taste") chemicals, as it is for insects and mammals. Thresholds for the other three substances are on the same order as they are for insects and mammals. Thus, although two structurally different receptors have not been distinguished for crustaceans, these animals still show evidence of two types of chemoreception (distance and contact), as in insects and vertebrates. Perhaps the structural similarity of crustacean antennal hairs masks functional differences in their nerve cells.

Questions of the distinction between "taste" and "smell" in crustacea

Behavioral significance of crustacean chemoreception. Chemically modifiable behaviour patterns are wide-spread among crustaceans and have received considerable study. Feeding responses usually occur in two steps: (1) response to chemicals from food at a distance, mediated through receptors on the antennae, antennules, and sometimes the tips of the legs; and (2) acceptance or rejection upon contact with receptors on the antennae, legs, and mouthparts. Barnacles have receptors that mediate feeding responses when stimulated with glutamic acid, proline, or potassium ions. It is believed that these materials initiate ingestion when they are released from prey that is punctured by spines on the entrapping legs of the barnacles. Electrophysiological studies on specialized appendages (dactyls) of the crab (*Cancer*) show that these respond to a variety of amino acids. Among crabs that feed on fish, the receptors respond to trimethylamine oxide and betaine, both chemicals found in fish flesh.

Parasitic and commensal crustaceans respond to chemicals from their hosts. Receptors on the antennules of commensal shrimps initiate nerve impulses when stimulated with fluid discharges (effluents) from their mollusk or echinoderm hosts. Communication by chemicals within any crustacean species is presumably common in the group but has been little studied. Swimming barnacle larvae aggregate specifically, attracted by a chemical given off by settled (fixed) individuals of the same species. This eventually makes reproduction possible among these fixed animals, since their eggs are fertilized internally. Sperms from one barnacle are transferred by a long penis to a neighbouring individual, this being feasible only because the animals aggregate. Sex pheromones have been reported for certain crabs. When ready to moult to sexual maturity, a female crab (*Portunus*) releases a chemical in her urine that attracts the male. In many species of crabs, the male is attracted from a distance by pheromones but uses his contact chemical sense for final identification of the female before mating.

Reactions to environmental chemicals are almost universal in crustaceans. Intertidal barnacles, like intertidal mollusks, respond when splashed with seawater by opening and becoming active, and they react to fresh water by closing tighter. The receptors that mediate this behaviour are along the edges of the mantle. Terrestrial isopods (sow bugs) select places that have specific humidities, the preferences varying with species and other environmental conditions. The receptors have been called osmoreceptors (since they conceivably respond to osmotic pressure), but there is no proof that they are distinct from ordinary chemoreceptors.

Insecta. Among the insects, only the blowfly (*Phormia*), the honeybee (*Apis*), and a few species of caterpillars and moths have been given detailed chemoreceptive study. Otherwise studies are scattered, in detail on only one aspect for some species, in others wide-ranging but without detail. Chemoreception in whole orders of insects has been almost entirely neglected; *e.g.,* among Neuroptera (*e.g.,* ant lions), Trichoptera (caddisflies), Odonata (dragonflies), Mecoptera (scorpionflies), and Plecoptera (stoneflies). For *Phormia* and *Apis,* however, investigative evidence rivals that available for man and rat; and understanding of the mechanisms of taste for *Phormia* is better than that for mammals.

Locations and structure of insect chemoreceptors. There is general agreement as to the parts of the insect body that bear chemoreceptors. Distance chemoreceptors are usual on the antennae and on the palpi of the mouthparts. For most insects, the antennae are probably the major locations of these receptors. In the honeybee, each antenna has about 500,000 receptor cells, most of them probably chemoreceptive, the remainder being mechanoreceptive (for tactile stimuli) and thermoreceptive (for temperature). Contact chemoreceptors are on the following structures: external mouthparts, pharyngeal wall (inner mouth), and ovipositor (egg-laying organ) in both chewing and sucking insects; tarsi (feet) and antennae in sucking species. A form of common chemical sense has been reported for insects but has been poorly studied. The receptors seem to be generally distributed over the animal's body, but they are still unidentified.

Chemoreceptors on insect feet

Regions of the insect body known to bear chemoreceptors have many types of so-called hair sensilla, named on the basis of their shape (Figure 9). The following types of sensilla are known from critical behavioral or electrophysiological studies to be chemoreceptive: (1) trichodea (hairs), distance and contact reception; (2) basiconica (pegs), distance and contact; (3) coeloconica (pegs in pits), distance; and (4) placodea (pore-plates), distance.

The following types of structures are suspected of being chemoreceptive: (1) sensilla ampullacea (flasklike pits), distance; (2) sensory patches in the pharynx, contact; and (3) free nerve endings in hairs and integument, common chemical sense.

The shapes of the sensilla are not fully reliable indicators of function. Trichoid sensilla, particularly, are active not only in both distance and contact chemoreception, but also in thermoreception and mechanoreception. Electrophysiological recording of impulses from specific sensilla should help settle the matter. The designations by shape also are not entirely precise, for many types of insect "hairs" are intermediate between typical long thin types and short blunt pegs, and some have extensive modifications of the walls.

In the central cavity of the hair or peg, chemoreceptive sensilla have terminal strands from neuron cell bodies at the base of the sensillum. The nerve cells are usually few in number, and their terminal strands (dendrites) branch variously to lead eventually to micropores (detectable only by electron microscopy) in the walls of the hair or peg. The taste hairs (labellar hairs) on the end of the extensible proboscis of the blowfly (*Phormia*) have been studied most thoroughly. Each of these has three to five neurons that send their dendrites to the micropores, plus a mechanoreceptive neuron with its dendrite attached to the base of the hair. The discovery of these micropores (formerly the exoskeleton of insects was thought to be imperforate) has necessitated considerable reinterpretation of experimental results.

Insect chemoreceptive processes. In the physiology of chemoreception among insects, many types of studies have been made—unfortunately, however, usually scattered among different species. Behavioral studies of feeding responses and other reactions to chemical substances at a distance and in contact, coupled with experimental removal of body parts and similar manipulations, have produced a large published literature. A few insects have been trained to give special reactions to chemical stimuli, the honeybee having been most extensively conditioned chemoreceptively. Some beetles, wasps, ants, flies, and cockroaches have also been studied in this way. Nerve impulses induced by chemical stimulation of the labellar hairs of *Phormia* have been detected electrophysiologically, representing the first time (1955) that a chemoreceptor of any animal was so studied. Since then, electrophysiological studies have been numerous, but mostly with relatively few species of Diptera (true flies) and Lepidoptera (moths and butterflies).

Among selected examples from the history of research on the functions of insect chemoreceptors, studies before 1950 had shown that the principal loci of distance ("olfactory") chemoreceptors are the antennae and that the end organs (terminal structures) are basiconic sensilla and pore-plates. Determinations of response thresholds, differing with the testing conditions, showed that the classes of chemicals to which insects respond at a distance are about the same as those that elicit responses from vertebrates. (The thresholds for series of chemicals are in the same general order for both groups of animals, although absolute values often differ widely.) Some species of insects are found to have distance chemoreceptors on structures other than the antennae, mainly the palpi of the mouthparts. The exact receptors and their properties were little understood in the 1950s.

Since about 1960, electrophysiological studies have yielded major data about the distance chemoreceptors

Reprinted from Karl von Frisch, *Bees: Their Vision, Chemical Senses, and Language,* (top) 2nd ed., © 1971, (bottom) © 1950 by Cornell University; used by permission of Cornell University Press

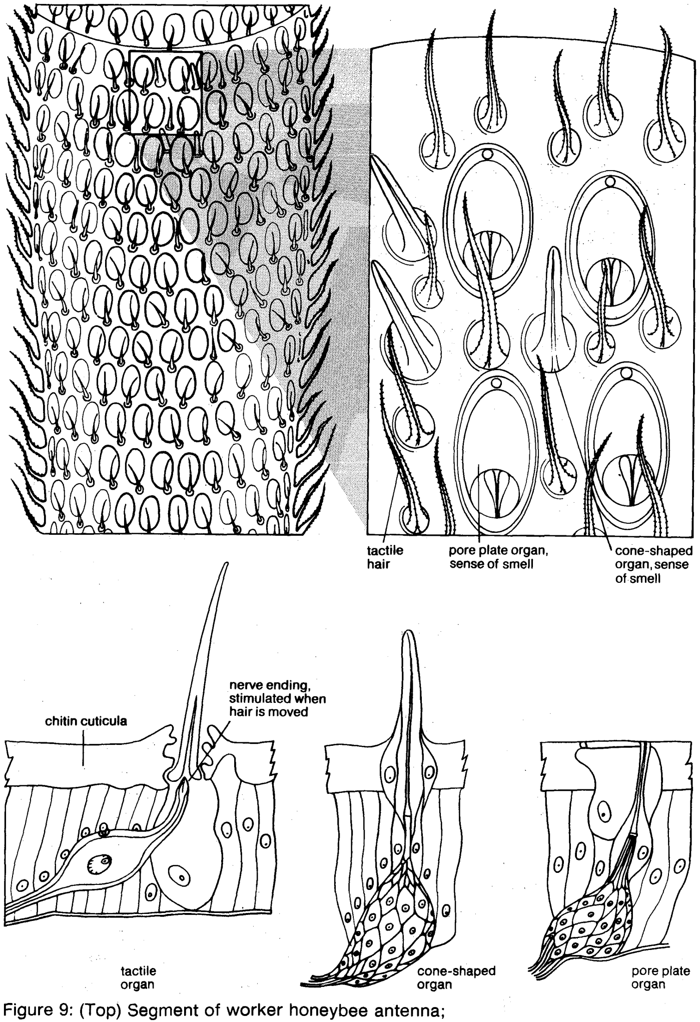

tactile hair
pore plate organ, sense of smell
cone-shaped organ, sense of smell
nerve ending, stimulated when hair is moved
chitin cuticula
tactile organ
cone-shaped organ
pore plate organ

Figure 9: (Top) Segment of worker honeybee antenna; (bottom) cross sections of the antenna's sensory organs.

Electrical activity of silkworm antennae

of insects. Nerve impulses are recorded from the antennal nerves to produce so-called electroantennograms. The major species studied are silkworm moths, both the commercial silkworm (*Bombyx mori*) and the giant silkworms (Saturniidae). Males of these species find their prospective mates by means of a special scent given off by the females; receptors on the antennae of males are remarkably sensitive to these special compounds. *Bombyx* males have about 40,000 sex-odour receptor cells on each antenna, with endings in various hairs and pegs. These structures are generally tuned to specific odours and so are called "odour specialists." They can be stimulated with odorant concentrations as low as 100 molecules of the given chemical per cubic centimetre of air. Females of the *Bombyx* species have distance chemoreceptors that are not so tuned; instead, the cells respond to a wide variety of chemicals, being called "odour generalists." The generalist type of receptor cell can respond both by increased neural firing (excitation) or by decreasing firing (inhibition).

Among caterpillars that feed on plants, odours are detected by similar sensilla on the short antennae. These structures are generalists, each responding to a variety of compounds. Their responses, however, differ in a number of ways: (1) latency, the time needed for response after a stimulus is presented; (2) rate of increase in frequency of firing; (3) rate of adaptation, such as loss of responsive capacity as stimulation continues; and (4) alternation of increase and decrease in the frequency of neural firing. Although there are only a few receptors present in the antennae of such a caterpillar, distinctive patterns of these four modes of response to different compounds represent a kind of code that the central nervous system of the animal seems to interpret as, at least, acceptable or unacceptable chemical stimulation.

After many years of behavioral studies on contact chemoreceptive processes among insects, electrophysiological methods have dominated the field since 1955. In many cases these continue to be supplemented by corresponding behavioral observations. The blowfly (*Phormia*) has become the "standard" subject, just as the fruit fly (*Drosophila*) has served in genetics. The labellar hairs of *Phormia* are known to be contact chemoreceptors; when its tip is inserted into a capillary tube containing a sapid solution, the hair responds with electrical changes that may be picked up through the solution. Thus the animal's responses to specific chemical substances can be readily monitored. An extensive mass of data has been gathered with this fruitful system.

Besides having a mechanoreceptive cell at its base, the blowfly's labellar hair has dendrites from four or five sensory cells. Each of these makes electrical responses that distinguish the cell as one of at least four types: (1) salt receptor (or cation receptor), once called L fibre because it produces large spikelike patterns of electrical activity on the recording screen; this cell is stimulated by positively charged ions (cations such as Na^+) and by acids and mediates behavioral rejection in water-satiated flies; (2) anion receptor, stimulated by negatively charged ions, and mediating rejection under all circumstances; (3) water receptor, once called W fibre; this structure fires when stimulated by water and mediates its acceptance by the animal; and (4) sugar receptor, once called S fibre because of its small electrical spike; stimulated by sugars, it mediates their acceptance by the fly.

Extension of the proboscis

Thus, rejection or acceptance of sapid solutions largely depends on the blowfly's receptors. A sugar solution causes one set of receptors to fire to bring about extension of the animal's proboscis and to stimulate feeding activity. A solution containing salt or acid stimulates another set of receptors to fire to inhibit extension of the proboscis and of feeding behaviour.

The stimulating thresholds for a great number of chemicals have been determined with the blowfly, and some general rules have been propounded. The stimulative effectiveness of cations and anions is proportional to the effective intensity of the electrical field generated by the given ion. At least for cations, stimulative effectiveness also seems correlated with the speed at which they move in solution (*i.e.,* their ionic mobilities). The data suggest that the receptor is stimulated by penetration or adsorption of the chemical on the surface; so far neither ionic mobility nor electrical field has been shown to be the only factor that affects thresholds. Rejection of alcohols and of other organic compounds by blowflies seems to be mediated by inhibition of the animal's sugar receptors. Stimulative effectiveness increases with carbon-chain length in a given series of chemicals up to about 11 carbon atoms. The effectiveness seems best correlated with the comparative solubility of the substance in water and oil, suggesting that penetration of the receptor surface is involved in stimulation. The effectiveness of sugars shows no obvious relationship to any of their chemical or physical properties but loosely seems to depend on their nutritional utility to the insect. Lactose, one sugar that is not adequate for nourishing flies, for example, does not stimulate the sugar receptor. Most stimulating are fructose, sucrose, and glucose, in that order; this is the order of their sweetness as tasted by man. In spite of the large amount of data available, however, neither the exact mechanisms of stimulation nor the details of their interrelationships has been worked out for these insects.

Among the so-called pseudotracheae ("false air ducts") on the labellar pads at the tip of the proboscis of these flies are short peglike sensilla (the interpseudotracheal papillae). Studied electrophysiologically, the papillae show evidence of bearing four kinds of receptors: (1) a mechanoreceptor; (2) a sugar receptor; (3) a salt receptor having other sensitivity as well; and (4) one with chemosensory function unknown as yet, although some data suggest that it may respond to amino acids. Specifically, the labellar hairs do not respond to amino acids, yet amino acids are ingested by blowflies.

The electrophysiological activity of taste receptors in *Phormia* has been correlated with the feeding behaviour of the animal. Attraction to foods from a distance is olfactory, mediated by receptors on the fly's antennae and palpi. Extension of the proboscis (at rest it is folded into the head capsule) is brought about by stimulation of sugar receptors, usually in tarsal hairs, sometimes in labellar hairs. Extension can be inhibited by appropriate stimulation of other sensory fibres by salts, acids, or repellent organic compounds. Stimulation of the labellar sugar receptors brings about sucking as long as stimulation of the other fibres is not too intense or provided that inhibition by other organic substances is not too great. Feeding behaviour is maintained and its level of activity is determined by stimulation of labellar and interpseudotracheal sugar receptors. The higher the concentration of sugar in solution, the more avid the fly becomes and the longer it feeds. As feeding proceeds, the sugar receptors adapt to stimulation, finally no longer firing above their resting levels, and feeding ceases. After this, chemoreceptors in the blowfly's foregut take over and shut off feeding behaviour until the meal is moved out. How widely this *Phormia* scheme will be found to operate remains to be seen, but, as studied so far, it seems generally to hold for other insects.

Behavioral significance of insect chemoreception. Studies on feeding behaviour among insects are extensive. Some insects are strictly monophagous (eating only one food); at the other extreme there are highly polyphagous insects (that eat almost any organic matter). Most insects, however, fall between these rare extremes, showing restricted food preferences that depend on the presence of specific marker chemicals (feeding stimulants) in acceptable items of diet.

Prey location by mosquitoes

Insects engage in a tremendous variety of mutual and commensal relationships; to do so they must find symbiotic partners. Many cases of chemical orientation to partners are recorded, usually in connection with the important communication signals of insects. Host finding by insects that parasitize other animals is likewise influenced or determined by chemical signals. Mosquitoes, for instance, find suitable hosts (*e.g.,* human picnickers) by sensing lactic acid, carbon dioxide, and moisture on the victim's skin, as well as by detecting his body heat and movement.

Chemical communication is probably universal among insects; it is certainly of major importance for the largest and best known groups. The possible practical use by man of sexual communication chemicals (pheromones) produced by insects (or made synthetically) in the control of these animals has led to extensive studies of materials that induce their sexual behaviour.

Social insects (*e.g.*, termites, bees, wasps, and ants) have been known for some time to use chemicals to scent the nest and to recognize individual members of the community. Advances in chemical analysis have facilitated the isolation and identification of many of these compounds. Some of these undoubtedly affect more than the insect's transient behaviour; the so-called queen substance of honeybees (trans-9-oxy-2-decenoic acid), for instance, suppresses development of ovaries in worker bees, often producing (when the swarm is not too large) a community with only one functional female. Similar chemicals are also used for trail marking and as guidance marks to food sources. In ants and stingless bees, deposits of secretions from the mandibular ("jaw") glands (containing such chemicals as geraniol, citral, various terpenes, and methyl ketones) function as guidance spots in the environment to direct fellows to food sources. The most thoroughly studied pheromones of insects are those used for sexual attraction and activation. Specific sexual attractants have been identified in about 250 species of insects. All but about 60 of these are Lepidoptera (moths and butterflies); most of the others are Coleoptera (beetles and weevils). In about 200 of these species, females attract males, and, in about 50 species, males attract females. Generally the attractant substances are what chemists call substituted hydrocarbons, with chain lengths of between eight and 17 carbon atoms in the molecule. It has been theorized that molecules that will allow sufficient structural variety while still being stimulating to insects should have chains of 10 to 17 carbon atoms and molecular weights of 180 to 300. Most of the active substances studied thus far fall within these limits.

Synthetic chemicals that act like the natural pheromones have been prepared for many insect species; these are mainly acetates with chains of 12–16 carbon atoms. Reactions of insects' olfactory receptors to these materials are remarkably specific. In field tests, male moths distinguished the specific chemicals of their own females when these substances were mixed with 26 other pheromones from different species of moths. In the laboratory, where concentrations may be made much higher than in the field, males may confuse some of the compounds, but not under natural conditions. Small differences in molecular structure or configuration can be highly significant. One molecular mirror image (trans isomer) of the Propylure molecule, a substance that attracts pink bollworm males, is active; the other mirror image (the cis isomer) does not attract, yet it masks the trans form when mixed with it.

Remarkably small concentrations of these pheromones can elicit behavioral responses. What was once thought to be the gypsy moth pheromone (isolated in tiny quantities from an extract of hundreds of thousands of female moths) and its synthetic version (Gyplure) have now been found to be inactive in themselves. The active principle seems to be some still unknown impurity present in even more minuscule amounts in the original extracts.

Use of pheromones in insect control

Insect pheromones are thought to be excellent prospects for pest control because of their attractant properties. Unfortunately, most attract males, and even a few fertilized females can maintain a population. At present, the major use of these materials is in population sampling; for instance, male cotton boll weevils (which emit substances called terpenoids) are used in traps to attract females in making a census of their population.

The use of pheromones in insect control is complicated by the finding that high concentrations repel and low concentrations attract. Thus, if high concentrations are used in insect traps to get wide coverage, the animals may be repelled when they get near. Furthermore, a pheromone used in baiting a trap must compete with the attractant from living members of the species. Many pheromones have multiple effects, depending not only upon their concentrations but on environmental factors as well. The so-called Nassonoff gland pheromone of honeybees, for example, consisting mainly of terpenes, serves the insects for attracting workers and queens, for marking food sources, in marking the hive, in scenting prospective hive locations by scouts, and in gathering swarms in flight. Thus, different behavioral reactions to the same pheromone can occur under different circumstances.

As a possible way out of many difficulties, it has been suggested that pheromones could be used to flood given locations with odour. This could fatigue the chemoreceptors of the insects and prevent them from finding mates; their sexual communication channel would be jammed. So far, the few tests of this idea that have been made in the field have not yielded very promising results. Except for short-term, geographically restricted effects, as among insects that live in warehouses where farm products are stored, pheromones for insect control have yet to fulfill earlier optimistic expectations.

Besides responding to food and communication odours, insects are oriented by a variety of other environmental chemical factors. Humidity responses have been extensively studied, but whether the receptors react to water vapour or are hygroreceptors (responding to lack of water) is much debated, with no general agreement. Places for laying eggs are selected by many insects (*e.g.*, mosquitoes and parasitic wasps) by chemical sampling of the prospective sites. Some plant chemicals and a number of synthetic materials repel various insects. There seems to be no generally occurring repellent for all insects, nor has any special relationship between chemical composition and olfactory repellency been discovered.

Protection of man and other mammals from attack by mosquitoes, fleas, ticks (which are arachnids, not insects), and other bloodsucking arthropods has been sought in chemical repellents. Tens of thousands of organic compounds have been tested as insect repellents, mainly for use against mosquitoes. Besides repelling at adequate levels when put on a part of the body that attracts the pests, the compound should not irritate the skin nor be otherwise harmful and should have a reasonable rate of evaporation. In the face of such criteria, few practical repellents have been found. Among those in common use are such substances as dimethyl phthalate, Indalone, Rutgers-612, benzyl benzoate, and Deet; the last is widely used, since it repels many arthropods—mosquitoes, fleas, and ticks. Repellent substances also have been sought among the many warning and alarm chemicals produced by insects, but most of these prove to be irritating to the skin or nose of mammals.

Alarm signals among ants

Alarm pheromones have been studied most intensively in ants, which produce them with special glands to alert their colonies to invaders or to other dangers. The active materials are generally related to hydrocarbons, often ketones; citral and its relatives are important components. Some of these chemicals are also constituents of social and sexual pheromones. Honeybees produce an alarm scent that contains citral and isoamylacetate, among other materials. Formic acid, produced by specialized glands of ants, is found to excite both ants and bees. All of these materials function to alert members of an insect colony when the community is threatened. Other insects (*e.g.*, some beetles) produce strongly repellent chemicals that serve to ward off predators. These chemicals range from apparently harmless but strongly odorous substances to such toxic materials as hydrocyanic acid gas. Among bombardier beetles the ejected spray is even heated by chemical action to about the boiling point of water.

CHEMORECEPTION IN THE VERTEBRATES

Besides the familiar vertebrates (animals with backbones), the phylum Chordata includes some smaller creatures sometimes called protochordates. Little indeed is known about chemoreception in such protochordates (*e.g.*, the lancelets and tunicates) beyond that they seem to show some selection of food and location and that they respond negatively to a variety of foreign chemicals. A group of what are commonly called lower vertebrates is the cyclostomes, such round-mouthed aquatic forms as lampreys and hagfish. Cyclostomes have a well-developed nasal

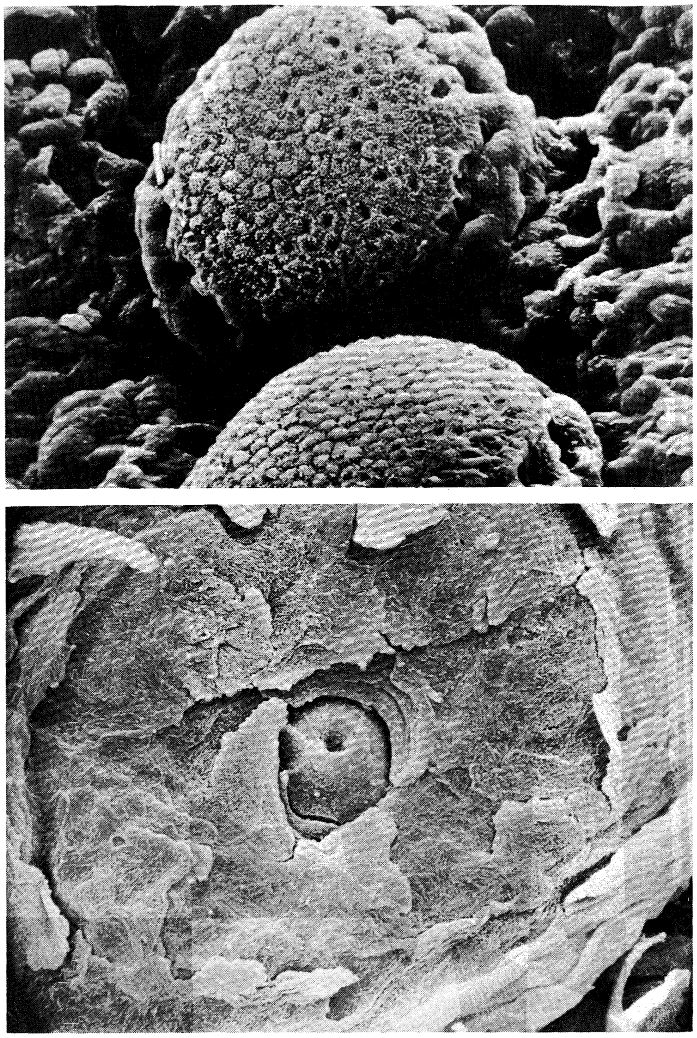


Figure 10: Scanning electron micrographs of (top) two frog fungiform papillae (magnified about 515 ×), and (bottom) taste bud with pore projecting through surface of rat fungiform papilla (magnified about 850 ×).

(Top) P. Graziadei, (bottom) L.M. Beidler, Florida State University

tract, with a single median (central) nostril; they can locate their prey by smell, but otherwise almost nothing is established about their chemical senses. For this reason, the bulk of attention given here to chordate chemoreception will be confined to the five main divisions of vertebrates: fish, amphibians, reptiles, birds, and mammals.

General vertebrate chemoreception. *Gustatory receptors.* The taste buds of vertebrates are secondary sense organs (*i.e.*, sensilla) derived from epithelial cells (Figure 10). Their structure has been well studied by electron microscopy, but in relatively few species (mostly mammals). Each vertebrate taste bud seems to consist of a number of cells of three or four types, but there is some debate as to their exact classification. One widely held view is that the taste bud has four types of cells: so-called supporting cells, sensory cells (the true receptors), basal cells (supplying replacements for old sensory cells), and another type of unknown function. Attempts have been made to designate developmental stages of these types and to view some of them as stages in the development of others, thus giving rise to at least five classes. The sensory cells are continually replaced, each cell having an average life span (at least for rat, mouse, and rabbit) of about 10 days (Figure 11). Each taste bud is innervated by up to 50 nerve fibres entering from below and branching into 200 or more branches to form a basket-like set of dendrites. Presumably chemical stimuli produce electrical changes in the sensory cells of the taste bud, these activating the afferent neurons nearby to generate nerve impulses.

Taste buds of reptiles, birds, and mammals are confined mainly to the upper surface of the tongue, with a few on the pharyngeal walls. In amphibians (*e.g.,* frogs) they are more numerous on the pharyngeal walls and present also on the cheeks and lips. In fish, taste buds are present also on the fins and in some species on the tail. In all cases, vertebrate taste buds are innervated from cranial nerves, mostly the facial and the glossopharyngeal.

Olfactory receptors. Among vertebrates these are the cells of the olfactory epithelium in the nasal cavities. They are primary receptors, true nerve cells the fibres of which form the olfactory nerve leading to the lobe of the brain that mediates the sense of smell. The structure of the cells of this epithelium, as seen with an ordinary (light-wave) microscope, appears remarkably similar for all vertebrates. Electron microscope studies reveal much more structural detail but have not changed the general interpretations. There are three fundamental cell types in the olfactory membrane: receptor cells, supporting cells, and basal cells; in addition, numerous gland cells furnish a mucous covering for the epithelium. Ramifying (branching) among the cells are very delicate terminal fibres of neurons leading to the brain through the trigeminalnerve. These are thought to be receptors of the common chemical sense, responding chiefly to irritants. The olfactory receptor cells have terminal cilia, which are fused into olfactory rods projecting outward.

Olfactory cell types

Man has about 40,000 sensory cells per square millimetre (26,000,000 per square inch) of olfactory epithelium, while the rabbit has about 120,000 per square millimetre, with an estimated total of 100,000,000 such cells. (Fish average between 45,000 and 95,000 per square millimetre, the eel having a total of about 800,000.) A significant discovery made with the electron microscope is that the olfactory sensory cells seem to be synaptically related. Such an arrangement would permit the cells to interact through mutual excitation and inhibition, thus allowing versatility of response at the receptor level itself.

The olfactory epithelium forms at least one wall of the nasal cavity of vertebrates. In fish, the nasal cavities are mostly paired pits or tubes just in front of the eyes, each with two nostrils, one anterior, the other posterior. In terrestrial vertebrates, the paired nasal cavities have external openings, the nostrils (external nares), and paired or unpaired internal openings (internal nares) into the mouth or pharynx. In all cases, water or air is moved through the nasal cavity and over the olfactory epithelium.

Another olfactory receptor of many vertebrates is the so-

Adapted from A.J.D. De Lorenzo, "Ultra-Structure and Histophysiology of Membranes" in Y. Zotterman (ed.), *Olfaction and Taste* (1963); Pergamon Press

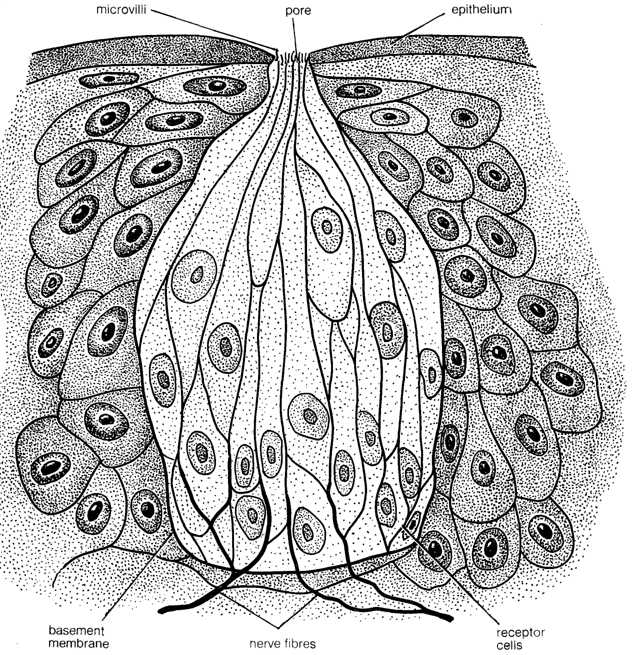


Figure 11: Microscopic section of taste buds of circumvallate papilla.

called Jacobson's organ (vomeronasal organ). This structure is variously developed; absent in fish, birds, and some mammals, it is highly developed in lizards and snakes. Nerve fibres from this organ lead to the accessory olfactory lobe of the brain and so are closely related to the primary olfactory system.

Common chemical receptors. Mucous membranes in vertebrates have receptors that respond to the presence of chemicals rather indiscriminately and, when stimulated, tend to evoke avoidance reactions from the animal. In mammals these common chemical receptors are restricted to the mucous membranes of the nose, mouth, pharynx, eyes, and genital organs. Free nerve endings in the olfactory epithelium of mammals are believed to respond to irritant chemicals.

Free nerve endings

In fish and larval amphibians, free nerve endings all over the animal's body seem to be sensitive to chemicals, their excitation eliciting avoidance reactions. These free nerve endings send their fibres to the central nervous system through spinal nerves. The free nerve endings of the head region enter the brain via the trigeminal nerve. These widely responsive receptors are vitally important in enabling the animals to escape from harmful chemicals in the environment, but relatively few studies have been made on them.

Process of gustation (taste). Among vertebrates other than man, the usual types of behavioral studies (*e.g.,* involving feeding responses) have been made, and training or conditioning procedures also have been used. Gustatory thresholds for detection, acceptance, and rejection have been determined. In more recent years electrophysiological techniques have been most numerous. Human reactions to tasted chemicals can be studied by experiments involving recognition of materials and verbal specification of preference or aversion. Aside from man, the animals most studied are frog, monkey, rabbit, rat, and cat; the investigations have focussed on taste qualities and on the action of sapid substances.

During the 19th century it was widely held that there are four primary taste qualities (salt, sweet, sour, and bitter) and that all other gustatory experiences represent combinations of these. Some investigators have added to these an alkaline and a metallic taste, but others claim that they are not primary qualities. On the assumption that there are four primary taste qualities, chemicals supposedly exemplifying each of the classes (NaCl for salt, sugars for sweet, acids for sour, and alkaloids for bitter) have been applied to the tongues of man and laboratory animals in attempts to find regions of selective sensitivity or (by electrophysiological tests) to locate different types of taste receptors.

Unfortunately taste buds are compound structures, and their neural connections are complex. At any rate, impulses recorded from nerves, or even from single taste buds, fail to give direct evidence about what the individual receptor cells can do. While recordings can be made by inserting fine wires into individual taste buds, the exact cell sampled is not known. It is clear, however, that vertebrate taste receptor cells are not classifiable as sugar, cation, anion, and water receptors as they are among insects. Some vertebrate cells respond to a fairly narrow range of chemicals, but most do not; those cells that respond to salts may also react to acids and sugars, or even water. Certain regions of the tongue tend to be selectively sensitive (*e.g.,* the tip of the human tongue seems highly responsive to sweet chemicals, but not uniquely so). It is no longer expected that, by studying impulses in single gustatory nerves, specific salt, sweet, sour, and bitter receptor cells will be discovered. It seems that patterns of response (rather than specific receptor activation) set up among the sensory cells on the tongue mediate the different taste sensations in man.

As in the case of insects, there is no general agreement on how sapid substances stimulate vertebrate taste receptors. For related series of organic chemicals, stimulative effectiveness is proportional to carbon-chain length up to some maximum and is also related to the comparative solubility of the substance in water and oil. Among inorganic materials, cations generally seem to have stimulative effects that are proportional to their mobilities, but there is great variability in response to the same ions from one vertebrate species to another. Sweet substances are not chemically definable; at least there is no obvious relation of taste with molecular structure. Although many sugars apparently stimulate the same receptors, man and other mammals often can easily distinguish one sugar from the other. Activation or inhibition of receptor cells occurs upon stimulation with different materials. The idea of four primary taste qualities or senses (modalities) has semantic utility, but to date it has not proved useful to investigators as a central dogma in understanding fundamental mechanisms of taste.

Process of olfaction. Studies of smell reception among vertebrates have been similar to those with taste, with electrophysiological methods dominating modern research. The literature on the subject is large, particularly with respect to man.

While attempts have been made to categorize odours in classes that could be considered primary, they have not produced a generally accepted system. The smallest number of primary odour qualities suggested is four, but more than 30 have been offered by some theorists. Attempts to relate odours to chemical structure or to other generalizable physical characteristics of odorous materials have not succeeded. Studies on mechanisms of stimulation of olfactory cells have similarly given rise only to theories, none generally acceptable.

The most active research on human olfaction is concerned with attempts to link odours, such as those of foods or perfumes, with specific chemical structures. Newer analytical techniques, as with insect pheromones, have facilitated the determination of the chemical composition of odorous materials present in the tiny amounts typical of natural products. By these means, extracts from foods can be separated into components with characteristic odours and chemically identified. From the standpoint of olfactory physiology, these studies emphasize the immense capacity of individual olfactory cells to detect a tremendous variety of chemical materials.

From white bread alone, for example, approximately 70 odorants have been identified, including alcohols, organic acids, esters, aldehydes, and ketones. From coffee, 103 separable volatile compounds have been isolated and many chemically identified; it is estimated that at least

From J.A. Nicol, *The Biology of Marine Animals;* Sir Isaac Pitman and Sons Limited

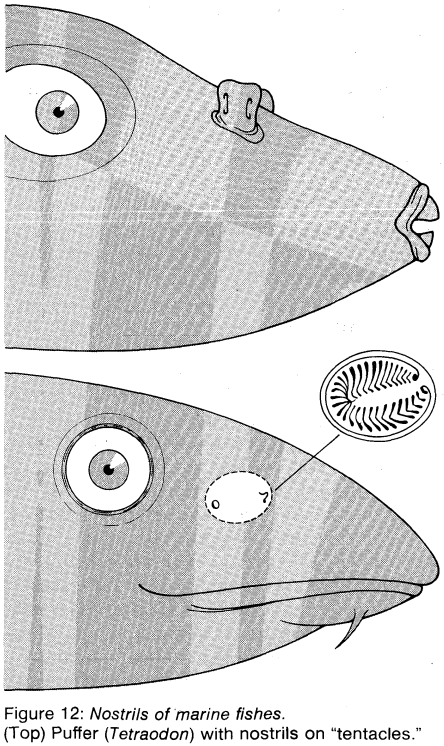


Figure 12: *Nostrils of marine fishes.*
(Top) Puffer (*Tetraodon*) with nostrils on "tentacles."
(Bottom) Cod (*Gadus*) with typical form of nostrils (inset shows detail of folds in nasal cavity).

150 substances contributing to the flavour of coffee will be discovered. Since many of these are present in extremely minute quantities, the capabilities of the human olfactory epithelium, usually regarded as having low sensitivity as compared with that of other mammals, seem remarkable. For substances called mercaptans (*e.g.,* in the skunk odorant), only about 40 receptor cells in the human nose need be stimulated by no more than nine molecules each to give a detectable odour sensation.

Skunk odour

Chemoreception in the main vertebrate divisions. *Fish.* Structure, location, and innervation of fish chemoreceptors are like those of terrestrial animals; thus separation into distance and contact chemoreceptive channels is possible. Taste buds are more widely distributed over the body in fish than in terrestrial vertebrates. In teleosts (*e.g.,* herring, trout, perch) they occur not only in the mouth and pharynx but also on the lips and regions nearby, on whisker-like barbels where present, on fins, and (in some fishes) on the tail. These taste buds are all innervated by branches of the facial nerve. The olfactory epithelium in the fish is in nasal cavities through which water passes; the nasal cavities do not, except in lungfishes, open to the mouth. There is no true Jacobson's organ, although some authors believe that structures near the nostrils may represent a rudiment of this organ (Figure 12).

Feeding behaviour among fish, as with all animals, is determined primarily by the chemical senses, smell being used to find food and taste to determine final palatability. Odours from foods excite movement in hungry fish, but true orientation toward food requires a current to indicate direction.

Social and sexual chemical signals are widespread among fish, though they are probably not as important as visual and possibly acoustic signals. In darkness, or for blindfish in light, species odours are important in schooling. Species differentiation may be excellent; a minnow (*Phoxinus*) can be trained to distinguish 14 different species of fish by their odours, even when the odours are offered in up to 15 different combinations. Mouthbreeders, fishes that hold eggs and young in the mouth, are able to distinguish their own offspring from those of others by odour. Some fish chemically mark their nests with mucus. Redfin shiners, fishes that lay their eggs in the nests of green sunfish, find these nests by the sunfish odour.

Some fish have remarkable powers of olfactory orientation to specific geographic locations. Minnows can distinguish the galaxy of odours of aquatic plants in their home streams and return to them when displaced. The most noteworthy of homing fish are salmon and eels, which return to the fresh water (where they began life) after some years in the sea. Each fish returns to the precise stream in which it was hatched. Many experiments have shown that this is possible only because they sense the odour of the natal stream. Apparently a form of learning called imprinting occurs in these baby fish. The hatchlings learn (or are imprinted) to associate the particular odour of a specific stream with home base. Orientation to the mouths of the streams from the sea requires some other talents as well; but, once the fish enters its home river, it unerringly finds its way to the headwaters where it started life.

Among the earliest reports of animal warning odours was that of the so-called *Schreckstoff* (German for "fright substance") given off by agitated fish. Injured fish produce chemicals that alarm other members of their own species, generally causing them to flee. The material is detectable to fish at extremely low concentrations. Some predators have turned this to their advantage; for example, sharks can detect the odour of an injured fish and swim toward it. In the hope that some chemicals besides those naturally occurring could repel sharks from swimmers, considerable effort has been expended to try to find a suitable shark repellent. So far the results have not been promising, but compounds that are remarkably effective in stimulating other fishes have been found; for example, phenacyl bromide repels teleosts at 0.01 part per million, but unfortunately it does not do this to sharks.

Shark repellents

Amphibians. In spite of the widespread use of frogs in physiology laboratories, understanding of chemoreception in amphibians is extremely meagre, particularly as related to their normal life. The gustatory organs are typical taste buds not only on the tongue and walls of the mouth and pharynx but also variously distributed on the lips. The nasal cavities of urodeles (*e.g.,* salamanders, newts) are relatively simple, but those of anurans (toads and frogs) are complex, with three chambers. The olfactory epithelium is of the usual type; in *Triton,* a salamander that lives both in water and air, cells of the olfactory epithelium have long cilia when the animal is an air breather and short cilia when it is a water breather. The Apoda, wormlike amphibians that are blind, have well-developed nasal cavities and olfactory epithelium. The organ of Jacobson in urodeles is a mere grooved channel in the nasal cavity, but in anurans it forms one chamber of the three nasal cavities.

As usual, feeding is chemically mediated, at least in part, although anurans mostly use their eyes for food capture. There have been few studies on chemicals that determine feeding among amphibians. Chemical communication is probably dominant in salamanders, odours of females attracting males in many species. Females of aquatic salamanders are induced to mate by chemicals produced by males. The chemicals are wafted toward the females by tail-wagging on the part of the males. Frogs and toads seem not to use chemical signals for communication, relying instead on auditory and possibly visual signals. As among fish, salamanders displaced from their home stream are able to find their way back by chemical sensing. Tadpoles, like fish, produce a *Schreckstoff* when injured, its discharge causing other tadpoles to scatter.

Reptiles. Chemoreception among reptiles has been very poorly studied. The major physiological work has been with turtles, mostly with objectives that are totally unrelated to the normal lives of the animals. The taste buds of the turtle are restricted to the tongue and the walls of the pharynx; the olfactory epithelium is in the nasal cavity. All reptiles but turtles have well-developed nasal cavities, crocodilians having exceedingly complex cavities and accessory sinuses. Jacobson's organ reaches its acme of development in lizards and snakes, where it opens into the anterior part of the mouth. Nearby, the lacrimal ducts (tear ducts) open, thus irrigating Jacobson's organ and possibly aiding in its function. This organ is absent in crocodilians and indistinct in turtles. While there has been considerable argument about the function of Jacobson's organ, it is now generally believed that it acts as a second olfactory organ in snakes and lizards at least. The forked tongue of some snakes can be inserted into the openings (inside the mouth) of Jacobson's organ, thus bringing chemical particles picked up on the tongue into contact with the olfactory epithelium. All snakes cannot do this, however, and apparently in some species the materials are dissolved off the tongue in the secretion from the lacrimal glands and thus brought to the organ.

Jacobson's organ in reptiles

Feeding behaviour among reptiles is probably determined to a large extent by chemical stimuli, but there have been few verifying studies. Some snakes find their prey by using the sense of smell, as is shown when newborn young of some snake species attack objects scented with extracts from the skin of species upon which they prey. The receptors involved in this case are in Jacobson's organ. Some predatory snakes cannot trail their prey when this organ is destroyed.

Communication in lizards, turtles, and crocodilians seems to be mostly by visual signals, although some tortoises have glands that secrete chemicals, which they distribute in their territories. Male snakes track females by detecting an odour on their skin; the male will not court his prospective mate if his nostrils are plugged. Rattlesnakes react defensively to the odour of king snakes; conversely, the predatory king snakes track rattlesnakes by their odour. Snakes seem to be more reactive to olfactory stimuli than are other reptiles; nevertheless, reasonable generalizations about the role of chemoreception in the lives of reptiles can only be expected to come from much more study than these animals have had to date.

Birds. General opinion among ornithologists is that birds are predominantly auditory and visual creatures; certainly among birds these senses are usually well developed.

This opinion, however, has led to a possibly unwarranted lack of interest in chemical reception among birds. There is growing evidence that chemoreceptors are well developed in at least some birds. The receptors are well-known: taste buds on the tongue and olfactory epithelium in a rather uncomplicated nasal cavity. The olfactory lobe of the brain in many birds (*e.g.,* kiwis, albatrosses) is large, suggesting a high degree of olfactory sensitivity. Birds have no organ of Jacobson.

Early observations on birds in the field led to the belief that their chemical senses were poorly developed, or even totally absent. Later studies, though few and scattered, suggest otherwise, however. Birds taste water before drinking or bathing, for instance, and their thresholds for rejection are similar to those for mammals. Indeed, birds drink water containing some chemicals at concentrations higher than those they tolerate in bathing water. The bathing thresholds may be well below thresholds for gustatory stimulation of mammals. The large olfactory brain areas of many birds certainly indicate that older ideas about their being olfactorily impoverished need reexamination. The few recent studies that have been made show that birds do, indeed, have an olfactory sense. Quail, for instance, can be trained easily to respond to odours and apparently can mark feeding locations by scent from their bodies, much as mammals do. The chemical senses and the place of chemoreception in the lives of birds—as well as reptiles and amphibians—deserve much more study than they have received up to now.

Mammals. Chemoreceptively, mammals are the best studied of vertebrates by far, and man is probably the most studied of all, although experimental techniques that can be used with blowflies (*Phormia*) and rats are inappropriate for humans. The taste buds of mammals are mostly on the upper surface of the tongue, on so-called vallate, foliate, and fungiform papillae. Some taste buds are also on the palate and in the walls of the pharynx. The olfactory epithelium lies dorsally in the nasal cavity, which in most mammals is extensive and complicated. Bony structures (conchae) subdivide the nasal passages, and sinuses extend into the bones of the mammalian skull. Aquatic mammals alone have relatively small olfactory areas. Jacobson's organ is absent in aquatic mammals, bats, and primates (*e.g.,* monkeys and humans). In almost all other mammals the organ is in the nasal septum (central dividing wall), being small but functional. In a few groups of mammals, Jacobson's organ opens into the mouth cavity through special (nasopalatine) ducts.

Mammalian feeding behaviour is dominated by the chemical senses; indeed, mammals generally are activated and oriented primarily by chemical stimuli. Food finding usually involves olfaction, and food testing involves gustation and olfaction together. Flavours of foods seem to determine acceptance or rejection in all mammals. (Flavour refers to the combined experience of taste, smell, texture, and temperature.)

Flavour testing

Flavour testing of foods for human use is an important factor in the economics of commercial food processing. Therefore an extensive literature exists on the techniques and results of flavour testing and on the production of synthetic flavouring materials. The gustatory organs supply rather restricted information to the brain, but the olfactory receptors supply a vast set of information. As an example of the wide array of volatile chemicals in foods, strawberries contain at least 35 chemical constituents contributing to their odour. These vary from time to time and with conditions in the same berry; for example, crushing converts some materials present in the intact fruit to other substances. The human olfactory organ easily detects these subtle changes, and responses in the brain are thereby affected.

It seems clear that the most important communication signals of mammals are chemical. Social aggregation and territoriality are guided by marking scents secreted by a variety of special glands in different places on the animals' bodies (*e.g.,* on the flanks, back, belly, and near the anus). The secretions are wiped onto objects or sprayed over terrain or are deposited by discharge of urine and feces at particular locations. Almost all mammals chemically mark their nesting or resting areas and quickly detect intruders. Members of a flock or herd (*e.g.,* of sheep) identify one another mainly by scent, apparently producing not only the species scent but also an odour distinctive of that flock or herd alone. Man's use of incense and perfumes in social and religious activities is probably rooted in the basic mammalian pattern of odour sharing within a group.

Similarly, chemical sexual signals are general among mammals. When their nostrils are plugged, male rhesus monkeys and males of some herbivores (*e.g.,* cattle) show no interest in females in heat. Among mice, the odour of strange males (from other communities of mice) interferes with the normal development of fertilized eggs in females; yet, signs of sexual arousal (estrus) can be induced in female mice and other rodents by the odour of a strange male. In probably all terrestrial mammals, arousal of the estral state in females is in response to odours produced by the male genital glands. While fastidious people often may say that sexual odours do not exist for man, the widespread use of perfumes (which supply masked sexual odours) attests to the importance to man of chemical channels in sexual communication.

Orientation to chemical cues is also general among mammals. Many mammals find water or home territories, even when far from them, by the sense of smell. As with fish, it is probable that the total odour complex from soil and plants of a region is detected by mammals.

Mammalian alarm odours

Alarm odours are part of the general communication system of most mammals. Many herbivores have special glands that release odours that alarm the herd when the animals are frightened. Similarly the odour of blood is repellent to many mammals. Many animals (*e.g.,* skunks) have warning odours that repel prowling predators. The tendency of mammals to discharge feces or urine when frightened is also adaptive, for these may act as olfactory repellents to enemies.

Man seems to be an unusual mammal in his limited use of the sense of smell. Other land mammals use olfactory function as their primary sensory basis for interacting with the environment. The sensitivity demonstrated for the human nose with respect to flavour discrimination suggests that even man relies much more than he realizes on the array of olfactory stimuli reaching him from the environment as sources of information. (The human senses of smell and taste are discussed further below; see *Human sensory reception.*)

THEORIES OF CHEMORECEPTOR ACTION

Attempts to create theoretical concepts to explain the actions of chemicals on chemoreceptors have generally been directed toward answering one, or both, of two questions:

1. What characteristics of chemical molecules are critical in producing responses by receptor cells?

2. What molecular characteristics elicit the experiences of particular tastes and smells?

There is still no generally accepted answer for either of these. The theoretical constructs developed have been somewhat different for taste and smell.

Taste (contact chemoreception). Many kinds of actions of sapid substances at gustatory receptor cells have been postulated, and some evidence has supported each. Unfortunately, much evidence militates against each. The most widely accepted possible mechanisms for stimulation of gustatory receptors are the following: (1) chemical reactions at the cell surface; (2) adsorption of molecules on the cell surface; (3) penetration of substances into the cell; (4) enzymatic reactions at the cell surface; and (5) protein bonding in the cell membrane.

Not all of the many adherents of theory (1) select the same type of chemical reaction at the receptor-cell surface. A few physicochemical models of the theory have been proposed, but none fits all the data. The adsorptive theory (2) is probably most widely believed now; while a wide spectrum of data fits this well, not all of the evidence is explained. The penetration theory (3) is supported by correlations between the oil–water solubility and stimulative effectiveness of sapid substances, but the mechanism seems to be too slow and too long lasting. The enzymatic theory (4) can be made to explain almost any data, if one just imagines the existence of the right enzymes (yet to be discovered). Nevertheless, the temperature independence of taste stimulation militates strongly against it, for enzymatic reactions are strikingly influenced by temperature. The protein-bonding theory (5) is weakened, as is the penetration theory (3), because these processes would be slow to reverse; otherwise good fits with data can be obtained by postulating the existence of appropriate proteins.

Some correlation between human taste responses and chemical composition has been found for sweet, salty, and sour substances, but the results are much less clear-cut for bitter materials. Most substances do not have one of the four simple tastes, and there are other suggested primary taste qualities, the validity of which has not been settled. Almost all investigators who have studied contact chemoreception in detail have come to doubt the validity of any theory of four primary tastes, at least for mammals. More recent electrophysiological data, although gathered mainly by workers who originally adopted the concept of four primary qualities as their guide, do not support the theory. Individual receptors (except labellar hairs of blowflies) are generally not excited by only one of the four presumed primary categories of sapid compounds. The intergrading of tastes for a large series of chemical compounds and the variety of electrical response patterns of receptors obtained in the laboratory suggest more of a continuum of taste-response patterns in a population of receptor cells than the existence of four specialized receptors for primary gustatory qualities.

Smell (distance chemoreception). Theories of olfactory stimulation are even less satisfactory than are those for taste. The events that have been suggested as occurring at the receptor cell to trigger off an olfactory response include the following: (1) chemical reactions at the cell surface; (2) solution of odorant molecules at the surface, thus altering surface tension; (3) radiant energy from an odorant affecting the cell without actual contact of the chemical molecule with the cell; (4) adsorption of the odorant on the cell surface; (5) effect of molecular internal vibrations (molecular resonance) on some aspect of cellular function; (6) enzymatic reactions; (7) penetration of odorant molecules with disruption of receptor-cell membranes; and (8) effect on an olfactory chemical or pigment within the receptor cell, similar to the effect of light on visual pigments such as visual purple (rhodopsin) within the retina of the eye.

Stages in the olfactory response

As would be expected with this array of olfactory theories (and not all proposed ideas are included), there is even less agreement than in the case of taste. The first, fourth, sixth, and seventh of these theories of smell are similar to their counterparts suggested for taste and have the same strong and weak points. The solution theory (2) seems too slow, particularly in accounting for recovery of olfactory sensitivity after adaptation to an odorant has occurred. There is no good positive evidence for olfactory theories based on radiant energy (3) or on olfactory pigments (8). Neither the molecular-resonance theory (5) nor the penetration theory (7) has even majority acceptance right now, although the idea that adsorption (4) is the critical step in stimulation seems to attract adherents.

Attempts to find and name odour primaries have proved more difficult than in the case of postulated taste primaries. A major stumbling block is that none of the theorized primary odour qualities can be related to specific classes of chemical compounds. The postulated primary odours have received such names as: foul, fruity, ethereal, fragrant, resinous, and burnt. The number varies from as few as four primaries to as many as 32, the most usual number being six or eight.

Some current theories relating olfactory experience to chemical or physical characteristics of odorous materials rely upon some postulated selection of primary odours; others do not. Although the first class of theories is based upon attempts to relate specific odours to particular chemicals, no reasonable correspondence between chemical structure and odour has yet been found. Attempts to correlate solubilities or other physicochemical characteristics with odours have been equally unsuccessful. Because many workers believe that the first step in olfactory excitation is adsorption of odorants on the surface of receptor cells, extensive studies have been made on correlations between odour and adsorptive behaviour of chemical compounds at interfaces between water and lipids (*e.g.,* fats or oils). The correlation is surprisingly good in some cases and poor in others. By changing postulated cell-surface characteristics, good correspondence with experimental data can sometimes be obtained, but the theory then potentially seems to fit any data and therefore is suspect.

Two newer, widely discussed theories are based, at least in part, on molecular shape rather than on chemical structure alone. One theory is based on the assumption that odorant molecules puncture the receptor-cell surface, thus releasing ions, and that the ability to puncture the surface depends not only upon the molecule's chemical properties but also on its shape. The olfactory quality experienced is believed to be the result of differential ability of molecules to puncture the receptor cells, determined by the size and shape of the molecules, and by differential rates of healing of the punctures by the cell. Not enough observational data are available on the fundamental events assumed here to make evaluation possible.

Theory of molecular shape

An alternative theory starts with the postulate that there are only seven primary odours, each of which results from the fitting of molecules of seven specific sizes and shapes into special receptor sockets imagined to exist on the cells. Thus molecules of compounds with a similar odour should have similar size and shape, and proponents of this idea believe that this is so. Others, however, find situations that are inexplicable by this "socket" theory. A most critical objection to this theory is that it is impossible to code the tremendous variety of definable smells with a system of only seven units. This has led some investigators to postulate many more than seven primary odours, separable molecular shapes for all of which have yet to be discovered.

Still another theory (5) of odour qualities starts from observations of high correlations between low-frequency molecular vibrations (resonances) and odours. This theory assumes different primary receptor cells, the number still unknown but probably relatively large. The primaries, in this case, are not postulated ahead of time (a priori). Since the theory depends on experimental evidence for its detailed development, only time will tell how or if the correlations will emerge. It is not assumed that the molecular characteristics being measured (*e.g.,* resonances called Raman spectra) are in themselves the stimulative factors; instead it is theorized that they are accompaniments of molecular energy characteristics that are the actual factors in olfactory stimulation. Thus, the unspecified molecular vibrational characteristics are postulated as acting upon energy-transfer mechanisms in the cell membrane or as determining orientation of odorant molecules on the cell surfaces.

None of these theories of smell at present has wide enough acceptance to be said to be the dominant idea. The general attitude is one of wait and see, while proponents of each gather data. Only further research will decide whether any one of these, or none, fits the observed evidence. Theories of gustatory qualities, starting with widely accepted agreement on primary tastes, and those on olfaction, starting without a generally accepted scheme of primary modalities, have now come to about the same conceptual turning point. (H.W.F.)

Photoreception

Photoreception is the activation of a biological process by means of illumination. Most organisms, including man, respond to visible light; some react to wavelengths of light not seen by man; and still others can react to properties of light not detectable by man, such as polarization (vibration of light waves in a definite pattern). This section is concerned with the sensory processes by which animals detect information carried by light (a detailed discussion of the human eye and its function is to be found below; see *Human vision*).

Light energy is necessary for life on Earth. Green plants require light for photosynthesis, the process by which water and carbon dioxide are transformed into carbohydrates; plants also show adaptive responses (*e.g.*, germination and flowering) to annual changes in daily light periods. Animals depend on plants for food and thus are indirectly dependent upon photosynthesis. In some animals, response to variations in day length is of great importance in the regulation of annual reproductive cycles. (For additional information about the above responses to light, see PHOTOSYNTHESIS and BEHAVIOUR, ANIMAL: *Stereotyped Response* and *Photoperiodism.*)

Light, the name given to the mediator of the sensation of sight in higher animals, including man, and the equivalent of this sensation in lower animals, is the part of the electromagnetic spectrum that is visible to animals; it includes the range of wavelengths from about 300 nanometres (1 nm = 10^{-6} millimetre) in the near ultraviolet to about 700 nanometres in the deep red (300 nanometres is beyond violet and does not evoke sensation in the human eye).

The entire cell of a unicellular animal such as *Amoeba* may be sensitive to light so that the cell moves toward or away from it. Some unicellular animals (*e.g.*, *Euglena*) have developed a light-sensitive receptor, or eyespot—a region with a lower threshold for light stimulation than occurs in the rest of the cell. Some multicellular animals have photoreceptive cells, or eyespots, scattered in various parts of or throughout the body; those in the outer covering of the earthworm (*Lumbricus*) serve in directional orientation, which involves comparison of light intensities at different directions. Most animals have localized photoreceptors of varying complexity—*e.g.*, the ocellus of certain mollusks and arthropods; the compound eyes of arthropods; and the camera eyes of cephalopods and vertebrates.

Evidence indicates that the eyes of certain insects can make use of the information carried by near ultraviolet wavelengths of light as well as that carried by visible wavelengths; both carry information related to the sensation of colour. The eyes of many invertebrates, such as certain arthropods and mollusks, have evolved in such a way that they can detect polarized light; *i.e.*, it evokes a sensation and provides information used for navigation. Visual sensation in higher organisms is primarily a complex response to the intensity and the spatial and temporal distribution of light on the photosensitive retina (the innermost layer of nervous tissue within the eye). Different eyes, different specialized parts of the same eye, and even the same parts of the same eye vary in their responses to illumination. Both the properties of light and those of the eye are thus important determinants of visual sensation. The great differences in the light-analyzing capacities of animals are reflected in the great diversity of gross structural organization involved in photoreception. The fundamental mechanism of photoreception—photochemical activation of a light-receptive pigment and the primary excitation-initiating process—seems to be similar among most animals, however.

This section deals with the optical properties of eyes, including the basic arrangements for image formation and light detection and the morphology of photoreceptors; the photochemistry of light detection; and the physiological functioning of the receptor cells that initiate nervous activity. These initial processes of photoreception provide information for the neural centres of the retina and higher nervous centres. The neural events involving the higher centres lead to visual perception.

THE OPTICAL PROPERTIES OF EYES

The first active step in vision is the absorption of light by a photosensitive substance, a visual pigment. Various devices within the eye assist vision by directing incoming light to this pigment; *i.e.*, they act as light guides by refracting (bending), reflecting (turning back), or guiding light. The arrangement of the optical structures influences the resolving capability and other basic properties of visual sensation.

Camera eyes. *In vertebrates.* The mammalian eye (Figure 13), somewhat like a camera, has a cornea (a transparent, anterior portion) and a lens; it functions as a dioptric system—*i.e.*, a system in which light rays are refracted so as to focus on the retina; the image projected on the retina is inverted. In the retina the light first passes through several layers of nerve cells before impinging on the photoreceptor cells, called rods and cones. The dimensions and refractive powers of all the optical parts of the human eye are known, making it one of the best understood vertebrate eyes (see below *Human vision*).

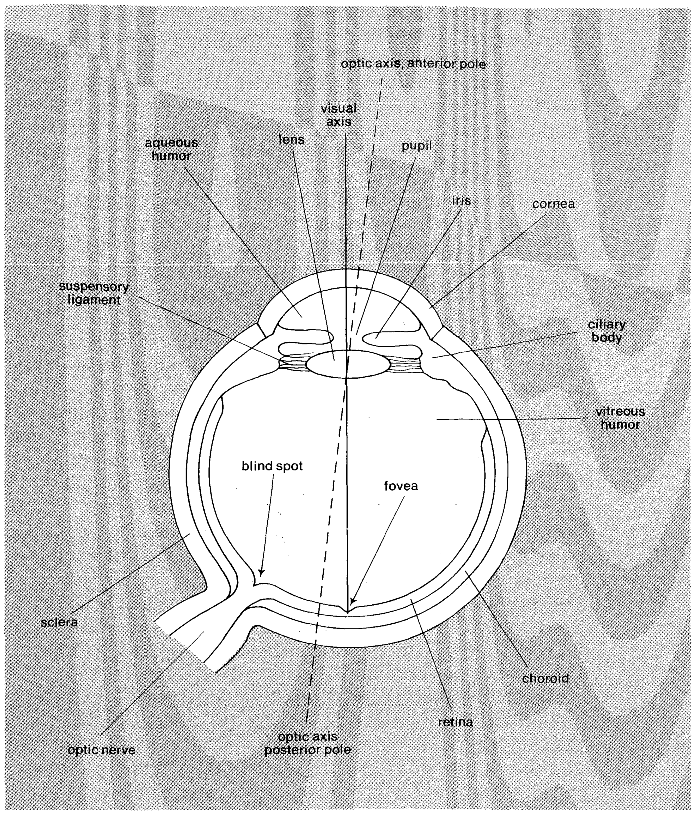


Figure 13: General structure of the mammalian eye.

From P.B. Weisz, *The Science of Zoology*, copyright 1966; used with permission of McGraw-Hill Book Co.

The influence of environment on vertebrate eyes

The large size of the camera-like vertebrate eye makes it potentially the most efficient of all eyes because it can project a large image on a large surface area containing a high density of receptors. Both vertebrate and invertebrate eyes reflect the influence of the animal's environment. The optical arrangement of the eyes of animals active during the night (*i.e.*, nocturnal) suggests that resolution is sacrificed for light-gathering power (see Figure 14). The opossum lens, for example, is so large that it almost touches the retina—*i.e.*, it has a short focal length, the distance from the centre of the dioptric system to the place at which the image of distant objects is focussed (retina). The short focal length combined with a wide aperture results in a low focal ratio, or f-number

$$\left(\frac{\text{focal length}}{\text{aperture diameter}}\right),$$

and ensures high light-gathering ability. In the eyes of animals active during the day (diurnal), the lens is smaller; as a result, the optical centre is closer to the front of the eye,

From G.L. Walls. *The Vertebrate Eye*, Cranbrook Institute of Science

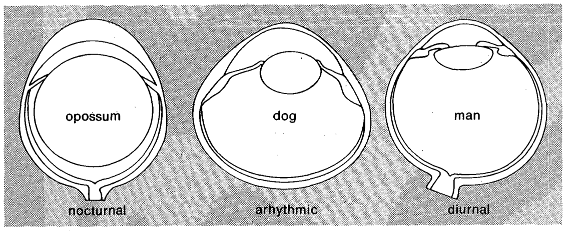


Figure 14: Influence of the environment on the optical arrangements of animal eyes (see text).

and its front surface is flatter. Thus, the focal length of the system is longer, the f-number is higher, and the image on the retina is larger and dimmer than in the nocturnal eye. Assuming that the large image can be detected by the photoreceptors, resolution is improved at the expense of the speed of the lens system.

In order to utilize efficiently a large image, the retinas of diurnal animals have localized areas with many photoreceptors; *i.e.,* a higher receptor density. The receptors in this area, called the area centralis, are usually cones, the receptors of daylight and colour vision. These areas for sharp vision, often circular, are seldom located exactly in the optic axis (an imaginary line drawn through the centre of the cornea and the central point of the eye, see Figure 13). The eyes of most birds have two such areas, the centre of each of which is specialized by a thinning of the retina to include only the receptors. This gives rise to a depression called the fovea, also found in teleost fishes, certain reptiles, and man. When the area centralis contains a yellow pigment, it is called the macula lutea. The macula lutea is found in higher primates (simians) and possibly chameleon lizards. This pigment filters out the shorter wavelengths of light and improves the sharpness of the image by reducing the chromatic aberration (variation of the focal length with different wavelengths of light) that would be caused by the inability of the lens to bring the long and short wavelengths to the same focus.

Accommodation in the eyes of vertebrates

Distant objects are in focus on the retina of the normal human eye. In order for objects closer than about six metres (20 feet) to be in focus, however, an adjustment called accommodation is necessary; otherwise, the image would fall in back of the retina, and the object would appear fuzzy. In mammals, birds, and reptiles other than snakes, the accommodative adjustment consists of sharpening the curvature of the lens so as to shorten its focal length. In snakes, elasmobranchs (*e.g.,* sharks), and amphibians, accommodation is achieved by moving the lens—hence its focal plane—forward. In lampreys and teleost fishes the eye is adjusted for near objects, and accommodation for distant vision is carried out by a backward movement of the lens. Some species have evolved adaptations that make accommodation unnecessary. The retina of the fruit bat (*Pteropus medius*) is in folds, ensuring that some part of it will intercept an image at any location. The ray *Raja batis* and the horse have ramp retinas, in which a continuous and gradual change occurs in the distance between the lens and retina in certain parts of the eye. Specific areas in these animals' eyes are presumably used to view objects at varying distances much as the human eye directs the image for detailed vision onto the fovea.

Because vertebrate species are adapted to almost every aquatic and terrestrial environment, they have evolved equally diverse eyes. In air, for instance, the front surface of the cornea can function effectively for image formation; in underwater eyes, however, the refractive index (the ratio of the speed of light in air to that in a given medium) of water and the cornea are almost identical, and the corneal front surface does not refract light. In these eyes the lens does much of the image formation.

Vertebrates have two types of photosensitive cells, rod cells and cone cells. The rod cells, which are long and fat, contain large amounts of visual pigment; they are the photosensitive cells for vision under conditions of dim illumination (scotopic vision). The cone cells, which are relatively small, mediate daylight vision (photopic vision) and colour sensation in many animals. The photosensitive photoreceptor outer segments of rods and cones are stacks of disks, or lamellae, with the planes of the disks at right angles to the long axis of the rod and cone cells. The retinas of animals active both day and night, as are those of humans, contain both rods (for night vision) and cones. In parts of the human retina the rods and cones are intermingled; elements of the nervous system provide the switching mechanism that permits adjustment for light conditions. The specialized fovea contains only cone cells; in the fovea the switching function is accomplished by eye muscles that change the direction of the field of vision in order to bring the image to the fovea.

The amount of light reaching the photoreceptor cells is controlled to some extent by the pupil, the opening of the eye through which light passes. The iris, the coloured portion of the eye surrounding the pupil, constitutes a diaphragm. Its muscles cause the pupil to change in diameter, decreasing the size of the pupil when light enters and increasing it when little or no light enters. The area of the pupil increases about 15 times in going from one millimetre to four millimetres (0.04–0.16 inch) in diameter, a relatively small increase in comparison with the range of light intensities under which the eye effectively operates. Since the amount of light entering the eye is proportional to the size of the pupil, it can be seen that changes in pupil size modify the amount of light only over a small range. Changes in pupil size are important in the human eye because they allow the lens to be used most effectively for visual acuity. When the whole lens is used, as in dim light when the pupil is large, the image formed by the lens is rather poor, chiefly because of chromatic aberration. The neural image on the retina is already poor, however, because the responses of thousands of rods must be pooled to obtain maximum sensitivity. Use of the whole lens is beneficial because it adds further light without reducing the image quality. When illumination is bright, the pupil is small, and only the aberration-free central part of the lens is used. This high-quality image is used effectively by the cone receptors of the fovea. There, no pooling of receptor responses occurs.

Adaptations of the eyes of nocturnally active animals

Pupils that form a circle when closed cannot greatly change in area; however, a pupil that forms a slit when closed can close almost completely. When nocturnally active animals find themselves in bright sunlight, they need additional protection for their sensitive rod-containing retinas; such animals have evolved pupils that close to form a slit. Many nocturnal vertebrates also show eyeshine (*e.g.,* the glow of a cat's eyes reflecting light at night). Eyeshine, which is caused by a mirror-like reflection from either the retina or choroid (a layer of blood vessels and connective tissue), enhances the sensitivity of the eye. The reflection of the light outward means that it passes the receptors a second time, giving them a chance to absorb light that was not absorbed during the inward passage through the receptors. Some animals thus have smaller rod receptors than they would otherwise need.

In cephalopods. The eyes of the invertebrate cephalopods—octopus, squid, and cuttlefish—are usually cited as examples of convergent evolution because they have independently evolved large camera-like eyes similar to those of vertebrates. The cephalopod eye lies within a cartilaginous cup. It consists of a cornea, lens, iris, and retina with the same basic relations to one another as are found in vertebrate eyes. Iris muscles can enlarge and narrow the pupil. Many details, of course, are different; for example, although the maximum density of photoreceptors in cephalopods is high—about 50,000 per square millimetre (32,000,000 per square inch) in *Loligo* and 100,000 per square millimetre in *Sepia*—the retinal structure otherwise bears little resemblance to that of the vertebrate. The photosensitive cells are of two types and are organized to detect polarized light (see below *Morphological features*). In addition, unlike vertebrate receptors, those of the cephalopod are the first element of the retina to be illuminated (rather than the last, as in vertebrates), and the optic ganglion (a mass of nerve tissue) is separate from the retina (rather than an intimate component, as in vertebrates). Both vertebrate and cephalopod retinas show pigment migration and movement of the receptors in response to adaptation for conditions of light and dark.

The cephalopod cornea does not have any focussing function. Image formation is accomplished entirely by the lens, which is forced forward for viewing nearby objects. The pupil is round in deep-sea cephalopods, such as *Loligo,* and slit-shaped in shallow-water dwellers, such as the octopus. Cephalopod photoreceptors are very long, an adaptation for nocturnal or deep-sea and low-light level conditions. The length allows for the presence of more visual pigment and hence greater absorption of light by each receptor cell.

Eyespots. Eyespots, the most primitive eyes, are found in the protozoan flagellates (unicellular animals with a flagellum, or whiplike structure, used for locomotion), flatworms (Platyhelminthes), and segmented worms (Annelida). An eyespot may be a specialized part of a cell as in protozoans, a single photoreceptor cell, or a small cluster of receptors with few or no accessory optical and neural structures. The entire epithelium (skin) of the annelid earthworm *Lumbricus* contains isolated light-sensitive cells that are eyespots. These photoreceptor cells are rather spherical in shape. A rhabdomere—a structure containing photosensitive pigment—lines a vacuole, or internal cavity, of the photoreceptor cell. The function of the vacuole may be to gather light.

Eyespots in planarians and amphioxus

The flatworm *Planaria* has a more highly developed eyespot. A number of photoreceptor cells are clustered under the epidermis. All of the rhabdomeres, which occur together within a cup-shaped collection of pigment cells, are located on slender filaments some distance from a cell that also gives rise to a nerve fibre.

A third type of eyespot is found in the nerve cord of the cephalochordate amphioxus. Each of a small cluster of photoreceptors, the Hesse cells, has a rhabdomere along the edge that faces a pigmented cell.

Ocelli. The ocellus, which is recognized as a true eye, is similar to a camera in that it usually projects an inverted image onto a light-sensitive layer. The ocellus is distinguished from the compound eye, which has many lenses, and from the more highly developed camera-like eye of the mollusks and vertebrates.

In mollusks. In a simple ocellus, that of *Nautilus,* the photoreceptor cells are bipolar—*i.e.,* the rhabdomere is at one end, the nerve fibre at the other—and arranged in a cup-shaped sheet. There is no lens or cornea, only a pinhole opening.

The more complex ocellus of the slug *Agriolimax reticulatus* is located at the tip of the tentacle (see Figure 15); there is a cornea under the epithelium, a vitreous body (a mass of clear jellylike material), and a lens, as well as a main retina and an accessory retina. The accessory retina is believed to function as an infrared receptor. As the tentacle is withdrawn, the accessory retina is rotated so that it is exposed to incoming radiation. The few photoreceptor cells are surrounded by pigment-containing cells. The ovoid eye is about 0.18 millimetre (0.007 inch) in its longest diameter. Distant objects appear in focus in the photoreceptor cells of *Agriolimax,* and the shape of the eye changes as the state of retraction of the tentacle varies. The change in shape may provide a mechanism for accommodation, although accommodation may not be particularly useful, because indications are that this ocellus does not clearly distinguish form.

From P.F. Newell and G.E. Newell, *Invertebrate Receptors* (1968); The Zoological Society of London, No. 23

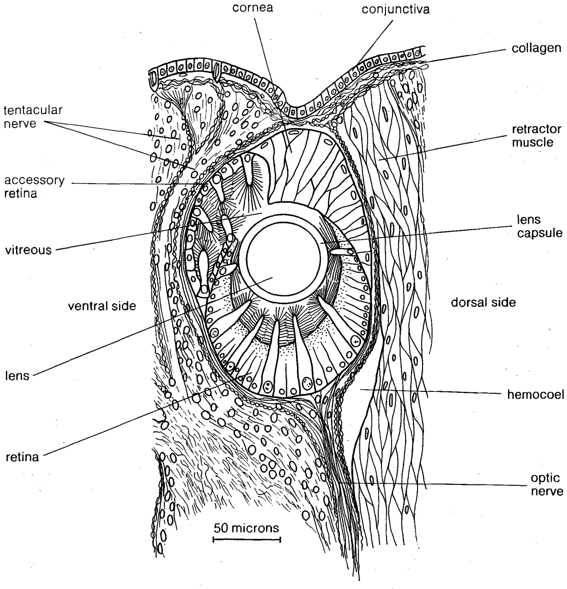


Figure 15: Longitudinal section of the eye of the slug *Agriolimax reticulatus.*

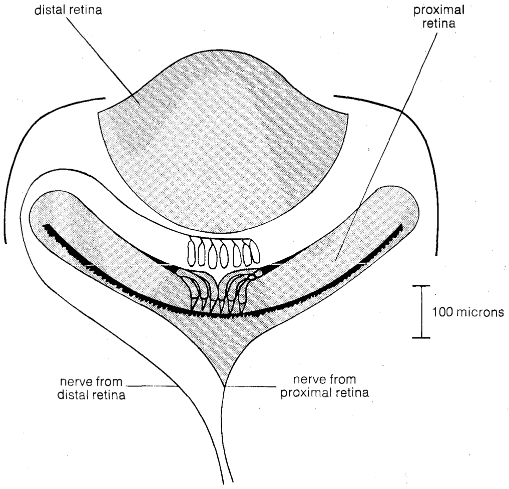


Figure 16: The central region of the eye of the scallop *Pecten.*

From M.F. Land, *Symp. Zool. Soc. London No. 23* (1968); The Zoological Society of London

The molluscan ocelli described above resemble a miniature simple camera, in which an inverted image is projected onto a photosensitive retina. The optics of other ocelli, however, are more sophisticated; for example, the scallop *Pecten* (Figure 16) and related genera have about 100 eyes located along the fringe of the mantle, the lining of the inner surface of the shell. The eyes are one millimetre (0.04 inch) in diameter and have a double retina; one is called a distal retina, the other a proximal retina. The nerve fibres of the distal (*i.e.,* farthest from the body axis) retina respond only to decreases in light intensity; the nerve fibres of the proximal (*i.e.,* closest to the body axis) retina, on the other hand, respond only to increases in light intensity. Because the photoreceptors in the proximal retina are inverted, light passes through all parts of them before reaching the light-sensitive pigment, and the photoreceptors just touch a reflecting structure, the tapetum lucidum. The scallop tapetum contains about 35 layers of thin crystals; the thickness of and the degree of separation between crystals are precise. The crystals and the intervening spaces function as an interference filter. The small reflection from each crystal interface adds to give a large net reflection from the surface of the tapetum. The fact that the photoreceptors just touch the tapetum ensures that they are illuminated on the second pass (after tapetal reflection) by the same light that passed through them the first time. The tapetum in the scallop eye acts as a concave (*i.e.,* depressed toward the centre) mirror, projecting light through the proximal retina and focussing an image of a distant object on the photoreceptors of the distal retina. The combination of tapetum lucidum and inverted photoreceptors in the proximal retina may thus enhance sensitivity without sacrificing resolution. The structural details of the tapetum lucidum have been described not only for the eye of the scallop *Pecten* but also for several arthropod and vertebrate eyes.

The tapetum lucidum

In arthropods. Among arthropods, ocelli are the main organs of sight in arachnids such as spiders and in insect larvae that undergo complete metamorphosis (*i.e.,* a radical physical change during development). Insects that undergo incomplete metamorphosis have three ocelli arranged in a triangle on the dorsal, or top, part of the head; these are subsidiary, however, to the main organs of sight, the compound eyes.

Spiders have two kinds of ocelli, the principal, or antereomedian, eyes and the lateral eyes. The principal eyes of jumping spiders are used when stalking prey and in courtship; the photoreceptors are arranged in four layers in the optic axis (see Figure 17). The two deepest layers cover the entire retina; the two most superficial layers are confined to the retina's central region. The rhabdomeres of the three deepest layers are rod shaped. Those of the most superficial layer are ovoid and oriented in the direction of propagation of light through the eye. Because the receptors are layered, the images of objects at differ-

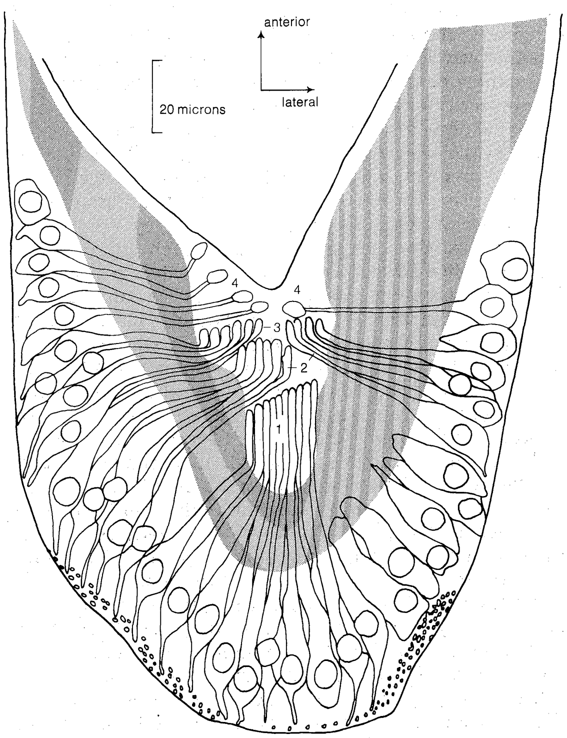


Figure 17: Frontal (horizontal) section close to the centre of the retina of the spider *Metaphiddipus aeneolus,* with the layers of receptor endings numbered 1 to 4. In layers 1, 2, and 3 the presumed receptive part of each receptor is the straight terminal portion; in layer 4 it is the terminal ovoid swelling.

From M.F. Land, *Journal of Experimental Biology,* vol. 51 (1969)

ent distances from the eye appear in different layers; it is possible that this type of retinal system is used for depth perception. It appears more likely, however, that the arrangement is used for colour vision. Images of distant red objects appear in the first retinal layer, blue-green images appear in the second layer, and ultraviolet images in the third layer, because the primitive lens system shows considerable chromatic aberration.

No image for visible light appears in the most superficial ovoid rhabdomeres of the principal eyes of spiders, and it has been postulated that they function to detect polarized light. The principal eyes also have a set of six muscles that produce eye movements when an image falls on the retina. The eye follows the image, which remains centrally fixed.

The lateral ocellus of the caterpillar, which has a group of five or six ocelli on each side of the head, resembles one ommatidium, or unit, of a compound eye. The optical apparatus includes a lens, or cornea; a crystalline cone, which is the principal focussing device; and two layers of photoreceptors (distal and proximal) in the optic axis, each of which is radially arranged like sections of an orange. The rhabdomeres of both cell types are in close apposition over much of their length. Objects at different distances project images at different vertical locations in the rhabdom—*i.e.,* a group of rhabdomeres in physical contact with one another. The proximal and distal rhabdoms form an optically continuous structure, a light guide. The optical information in this structure depends on the light accepted at the distal part of the rhabdom, rather than on the locations in the rhabdom at which images are projected, because the rhabdoms actually touch one another and effectively form one optical structure (see *Optical properties of photoreceptors* below).

Spatial resolution in ocelli and camera eyes

No optical system can form a point image of a distant luminous point, such as a star; because of diffraction (a modification in which a redistribution of energy occurs), the image is instead a small spot, the Airy disk, with a bright centre enclosed in concentric, alternately bright and dim rings. The radius of the first dark ring indicates the size of the diffraction pattern. The smaller the aperture and the longer the wavelength, the bigger the Airy disk and the poorer the resolution.

The degree to which images can be detected is determined both by the size of the Airy disk and the size, separation, and density of packing of photoreceptors. Compared with the eyes of most invertebrates, a relatively large number of photoreceptors occurs in the principal eye of the jumping spider. This suggests that the image is good, that its quality is primarily determined by the size of the lens, and that the mosaic of receptor units makes maximal use of the image quality. By comparison, the theoretical resolution of the human eye is 10 times better than that of the spider. This difference results solely from the difference in the size of the eye. In both the human eye and that of the spider, the focal length is adjusted to transmit the image onto a retina containing about the same density of photoreceptors. Actual resolution, however, is complex and depends on other factors, such as the nature of the target, properties of the illumination, the form of measurement, and neural processing in the eye and brain.

In the lateral ocellus of insect larvae, the size and density of the receptors are probably the limiting factors in resolution. In addition to poor spatial resolution, the light-gathering power of the lateral ocellus is also inferior to that of the spider eye.

Compound eyes. Aggregations of ocelli are found in several lower animals—in polychaete worms, for example. True compound eyes are found only in the arthropods, however. The compound eye of insects is composed of hexagonal or rectangular-shaped, closely packed optical units called ommatidia (small eyes); each ommatidium is virtually a single eye. In different species the size, number, and structure of ommatidia vary. An ommatidium (Figure 18) is composed of a corneal lens, or facet, which consists of a modified extension of the cuticle (the hard outer covering of arthropods) on the surface of the eye; four cells called Semper's cells or cone cells, which form the crystalline cone; and a sensory region called the retinula (small retina). In primitive insects (*e.g.,* the springtail *Lepisma*), in which the transparent cone cells are not specialized, the ommatidia are called acone ommatidia. In the more common eucone ommatidium, which occurs in moths and butterflies, the cone cells have a more complicated structure and contain granules of glycogen, or animal starch; because the granules are packed at various distances from each other, the refractive index varies in different positions in the cell. In certain beetles (*e.g., Lampyris*) and in the horseshoe crab *Limulus,* the crystalline cone is an extension of the cornea.

From *Journal of Cellular and Comparative Physiology* (1965)

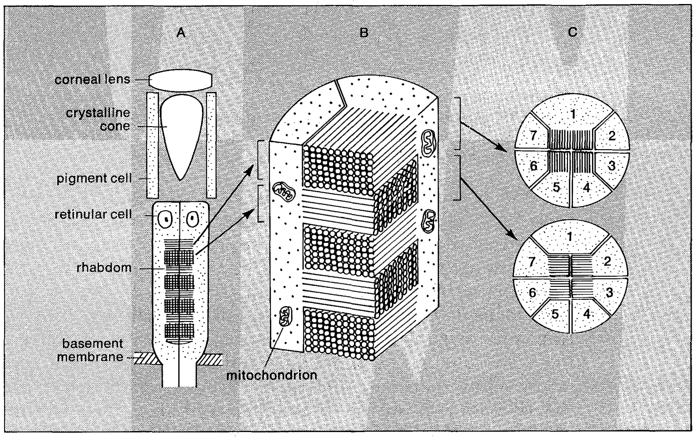


Figure 18: *An ommatidium from the compound eye of the crayfish Procambarus clarkii.*
(A) Longitudinal section through the optic axis with the cone stalk between the crystalline cone and the retinula omitted.
(B) Part of the rhabdom. The direction of the closely parallel microvilli in one layer is perpendicular to that in the next.
(C) Cross sections through the retinula at the levels of two neighbouring layers, rhabdomeres from retinular cells 1, 4, and 5 constituting one layer (upper figure); 2, 3, 6, and 7 constituting the next layer (lower figure); 1, 4, and 5 the third; and so on alternately.

Components of the ommatidium

The sensory part of the ommatidium, the retinula, consists of several radially arranged cells (retinular cells); each has a photoreceptor component, or rhabdomere (see Figure 18). The rhabdomeres of neighbouring retinular cells may be either in contact (forming a rhabdom) or completely separate. The optical isolation of each ommatidium is enhanced by its being surrounded by light-screening, pigment-containing cells. During adaptation to light and dark conditions, migration of pigment in cells around the crystalline cone, the corneal process, the proximal part of the ommatidium, and within the retinular cells has been reported to occur in most types of compound eyes.

The apposition eye. In the compound eyes of diurnal arthropods, each ommatidium is separated from its neighbours by pigmented, or iris, cells under all conditions of illumination. As a result, the rhabdom of each ommatidium receives light only through its own corneal lens; light from the lenses of other ommatidia is blocked by the pigment. This is the basic structure of the apposition eye (Figure 19). There are, however, variations in structure.

From G.A. Mazohkin-Porshnyakov, *Insect Vision* (copyright 1969); Plenum Press

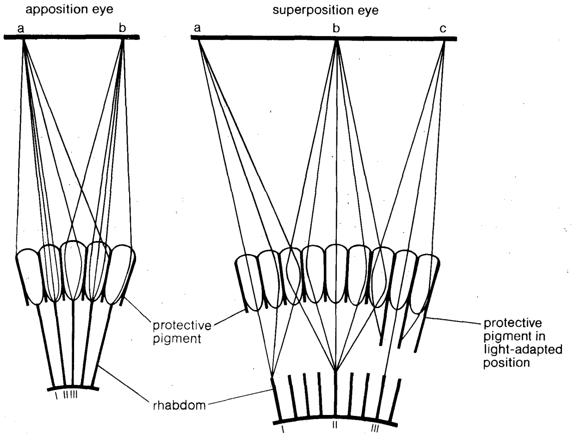


Figure 19: Image formation in apposition and superposition eyes: I, II, III represent different ommatidia; a, b, c represent point objects (see text).

In the honeybee *Apis mellifera,* the cornea is layered, each layer having a different refractive index. It has been shown that the corneal lens projects a small inverted image of a distant object in the crystalline cone some distance short of the rhabdom of an ommatidium. If the image were focussed exactly on the rhabdom, it might effectively transmit image-location information by means of differing responses of the individual retinular cells within an ommatidium. Because the image is focussed in front of the rhabdom, however, individual retinular cells within an ommatidium do not respond to image location. The defocussing also has the effect of broadening the field of view of the ommatidium. As the image moves, it can still be guided into the rhabdom by the crystalline cone, which is considerably wider than the rhabdom.

Specialization of the butterfly eye

It is not possible to generalize from the known optics of the compound eye of the bee to other apposition eyes as, for example, those of the butterfly, which superficially resemble the bee's in having a cylindrical rhabdom. Certain features of the butterfly eye are known, however (see Figure 20). Under the rhabdom of each ommatidium, the tracheole, or respiratory tubule, which supplies air to the ommatidium, has become specialized to form a tapetal filter comparable with that of the *Pecten* eye; it reflects highly saturated colours, usually blue colours in the dorsal ommatidia and more reddish colours in the ventral part of the eye. Each ommatidium thus has its own mirror. (The swallowtails, *Papilio,* are the only butterflies lacking this adaptation.) The light in the rhabdom of an ommatidium is tinted by the coloured light reflected from the mirror, and can be seen only with the aid of an ophthalmoscope—an instrument for viewing the interior of the eye. This filter system probably aids vision by enhancing contrast between objects of certain colours.

Butterflies with a tapetal filter, as well as nocturnal species (with a tapetum of different structure), have a corneal surface structure that acts as an anti-reflector. It consists of minute conical corneal "nipples" that provide a gradual transition between the refractive index of air and that of the cornea, effectively eliminating front-surface reflection for many wavelengths of light. One of the functions of this coating may be to minimize reflection of images from the tapetum lucidum back into the eye.

The rhabdomeres within an ommatidium in the compound eyes of flies (Diptera) have a microvillar structure (*i.e.,* one involving minute hairlike projections; see below *Morphological features*) and are completely separate from one another, in contrast to the apposition eyes above, in which the rhabdomeres are in contact. Each ommatidium contains eight retinular cells and eight rhabdomeres.

From any relatively distant point in space, a small cluster of ommatidia in a fly's eye appears darker than other areas because these ommatidia are the ones that are best aligned to absorb light coming from the direction in which the ommatidia are pointed. This cluster, called the pseudopupil, is often found in insect eyes; it moves, appearing to follow an observer viewing different parts of the compound eye. In the compound eye of the butterfly it is the pseudopupil that lights up with colours when observed and viewed from the same direction as the illumination with the ophthalmoscope. The pseudopupil of flies usually consists of seven ommatidia. In this apposition eye a distant point forms an image on the distal end of the rhabdomere, not within the crystalline cone as in the apposition eye of the bee. From any distant point in space only one rhabdomere in each ommatidium of the pseudopupil is illuminated by a point source; *i.e.,* one rhabdomere in each ommatidium of the pseudopupil is directed toward the same point.

Spatial resolution in the apposition eye

Experimental results suggest that the resolution of the compound eye (*i.e.,* the capability of the eye to distinguish between two separate but adjacent objects) is determined by the divergence angle between the ommatidia. If a compound eye views a pattern of alternating black and white bars in which the angle formed by one bar is the same as that between the ommatidia, the image of one black bar falls on one ommatidium, and the image of the flanking white bars falls on its nearest neighbours. This defines the limit of resolution; if the image of more than one bar falls in an ommatidium, the bars are not resolved.

The image is out of focus at the rhabdom in the bee eye. The angle between fly rhabdomeres is the same as the angle between ommatidia. The fused rhabdom of the bee and the separate rhabdomeres of the fly probably convey

From N. Yagi and N. Koyama, *The Compound Eye of Lepidoptera: Approach from Organic Evolution* (1963); Maruzen Co., Ltd.

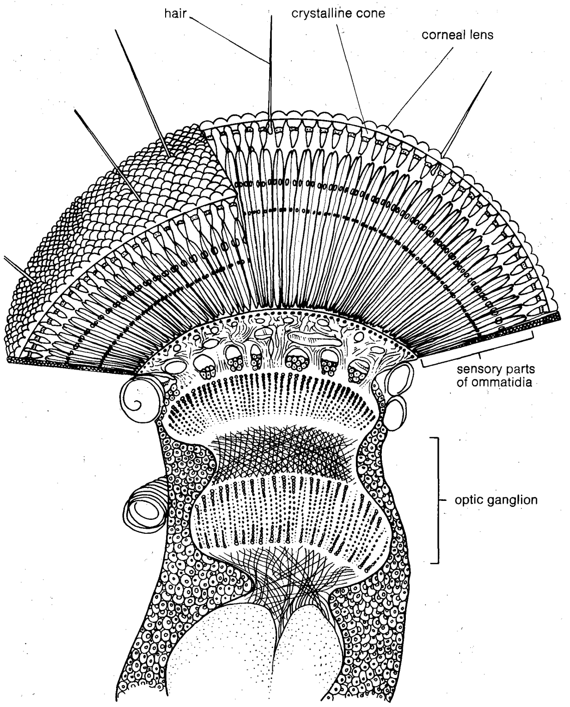


Figure 20: Section through the compound eye and optic lobe of a butterfly.

polarization and colour information rather than spatial information.

The array of corneal lenses on the surface of the eye can be considered as sources of light that interfere deep within the eye to give high-order diffraction images. The structure of the eye suggests that this effect is not detectable by the photoreceptors. Substantial shielding pigment between the ommatidia reduces the intensity of such patterns, which thus should be relatively ineffective in exciting the rhabdom.

The superposition eye. In the compound eyes of nocturnal arthropods, the rhabdoms are deep within the eye, far from the cornea and crystalline cone. In 1891 Sigmund Exner, an Austrian physiologist, reported experiments that showed how these eyes function. He demonstrated that there is pronounced pigment migration within the iris cells. In eyes adapted to darkness, the pigment of these cells migrates upward and into spaces between the crystalline cones of neighbouring ommatidia. In the light-adapted eye, on the other hand, the pigment migrates into the region between the cones and the rhabdom, in effect isolating each ommatidium by surrounding it with a light-absorbing tubular pigmented structure.

The pigment migration has the effect of changing the sensitivity of the eye. When a light-adapted eye is placed in the dark, there is an initial increase in sensitivity of about 10 times with no accompanying pigment movement. In the following 25 minutes in the dark, the sensitivity increases about 100 times, as the pigment gradually withdraws from between the cones.

Important differences between apposition and superposition eyes

Exner found that the corneal process of the firefly *Lampyris,* a structure that resembles the crystalline cone of butterflies and other nocturnal insects, has a higher refractive index at the centre than near the periphery and effectively bends incoming light. When the eye is dark-adapted, light from many facets converges in the rhabdoms of many ommatidia forming a so-called superposition image (Figure 20). This eye differs from the apposition eye in that light from many facets is involved in forming an image in the rhabdom of an ommatidium; in the apposition eye, on the other hand, light from its own corneal lens reaches the rhabdom within a particular ommatidium. The mosaic image of the superposition eye, although less sharply defined than that of the apposition eye, is brighter and thus a valuable adaptation for nocturnal insects. In the light-adapted condition, inward migration of pigment in the iris cells effectively prevents the spread of light to adjacent ommatidia, and the superposition eye acts somewhat like the apposition eye; *i.e.,* only the light from one facet reaches the rhabdom within an ommatidium.

THE PROPERTIES OF PHOTORECEPTORS

The photoreceptor cell absorbs light energy and transforms it into a nervous response. The actual photoreceptor component, or organelle, of the photoreceptor cell contains a coloured substance (visual pigment) that absorbs light and initiates the chain of chemical reactions leading to nervous excitation. The vertebrate photoreceptor cells, the rods and cones, are so called because of the shape of their photoreceptors, which are found in the outer segments of the cells. Invertebrate photoreceptor cells, the photoreceptors of which are the rhabdomeres, show greater diversity of structure than do those of vertebrates.

The function of photoreceptors is dependent on the visual pigment they contain; the identification of photoreceptors ultimately depends, therefore, on correlating the light-absorbing properties of the pigments within them and the physiological response of the receptor cells. Such evidence exists for a number of invertebrates and vertebrates; the identity of some vertebrate and many invertebrate photoreceptors, however, has been inferred from their location, colour, and structure after comparison with known photoreceptors.

When viewed with the light microscope, the outer segments of vertebrate receptors and the rhabdomeres of invertebrates can be recognized by their shape, location, and high refractive index. In the electron microscope, all photoreceptors are seen to consist of a dense collection of membranes. The outer segments of vertebrate photoreceptors are lamellate (layered). Invertebrate rhabdomeres usually consist of microvilli (minute projections of the receptor cell membrane) and resemble a tiny honeycomb in three dimensions. In certain invertebrates, however, the rhabdomeres are lamellate, resembling those of vertebrates. Some features of photoreceptors from various animal groups are discussed below.

Morphological features. *Invertebrate photoreceptors.* The photoreceptor of the jellyfish *Polyorchis penicillatus* (Figure 21) is a modified cilium, or hairlike structure, as are the photoreceptors of many higher animals. Cilia are found in the cells of most animals; their motion is used either to move the animal or to move materials within or outside the body. As seen in the electron microscope, a cilium characteristically consists of nine double, peripherally arranged filaments (very fine threadlike structures) and a central pair of filaments in a so-called 9 + 2 pattern. A cilium arises from a cell structure known as a centriole, near which is often found a second centriole; ciliary rootlets extend from the centrioles to deep within the cell. In sensory cells a cilium often develops into a sensory organelle. Sensory cilia invariably lack the central pair of filaments, having what is called a 9 + 0 pattern. The sensory cilium of *Polyorchis* contains numerous coarse microvilli, which constitute the sensory portion; centrioles and a ciliary rootlet are also present. Near the base of the photoreceptors are frequently found large numbers of cell components called mitochondria. Mitochondria manufacture the compound adenosine triphosphate, which seems necessary for the photoreception process.

From *Proceedings of the U.S. Academy of Sciences* (1962) in D. Mazia and A. Tyler, *General Physiology of Cell Specialization,* Copyright 1963; used by permission of McGraw-Hill Book Company

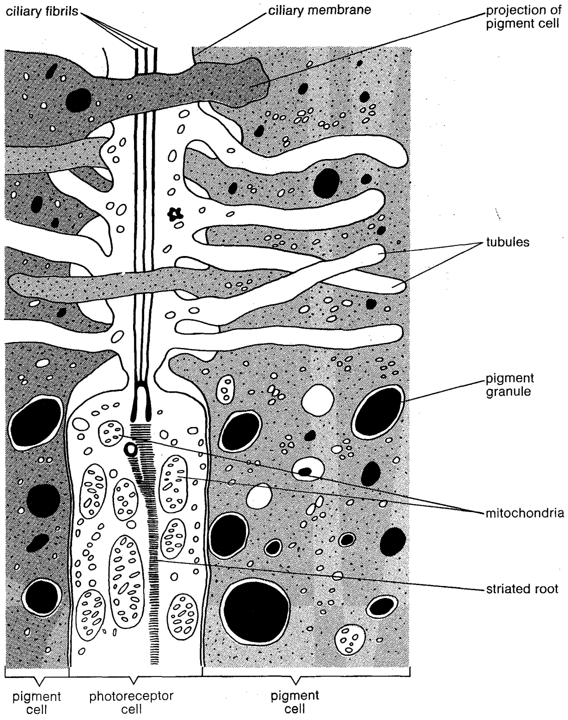


Figure 21: The distal part of a photoreceptor cell and two adjacent pigment cells in the ocellus of the hydromedusan *Polyorchis penicillatus.*

Photoreceptors in lower invertebrates

The microvilli found in the primitive photoreceptor of *Polyorchis* are atypical in two respects: first, it is unusual for the microvilli of rhabdomeres to be derived from cilia; second, the microvilli of the rhabdomere, which are usually packed together very densely, are, in the jellyfish, mingled with the microvilli of the pigment-containing cells that envelop the rhabdomere. Migration of pigment into or out of the pigment-cell microvilli provides a mechanism for regulating the amount of light absorbed by the receptor. The identification of photoreceptor cells in the jellyfish has been based on anatomical and behavioral evidence. The cells are located in ocelli, the removal of which renders the jellyfish incapable of responding to illumination.

Although the ctenophorans (comb jellies) usually have been assumed to lack both a response to light and photoreceptor cells, recent evidence has raised some doubt about such a conclusion. Electron microscopic examination of *Pleurobrachia pileus* has revealed a radial arrangement of four groups of lamellate bodies; the lamellae are composed of membranes of about 12 sensory cilia (*i.e.,* 9 + 0 type). The sensory cilia, instead of developing microvilli as in *Polyorchis,* become individual platelets. These lamellate structures, which are located in infolded regions of the presumed photoreceptor cell, resemble similar lamellate structures in certain molluscan and vertebrate photoreceptors; it is presumed by some investigators, therefore, that the lamellate structures of *Pleurobrachia* may be photoreceptors.

The planarian flatworm *Dendrocoelum lacteum* (phylum Platyhelminthes) has photoreceptor cells with well-developed rhabdomeres (see Figure 22). The cytoplasm of the receptor cell, which has a tubular extension, contains many mitochondria; the surface membrane, which forms the tightly packed microvilli that constitute the photoreceptor organelle, is presumed to be the site of the visual pigment. No evidence has been found that either this photoreceptor or those of other flatworms are derived from cilia.

From *Zeitschrift fur Zellforschung und mikroskopische Anatomie* (1961) in D. Mazia and A. Tyler (ed.), *General Physiology of Cell Specialization* (Copyright 1963); used by permission of McGraw-Hill Book Company

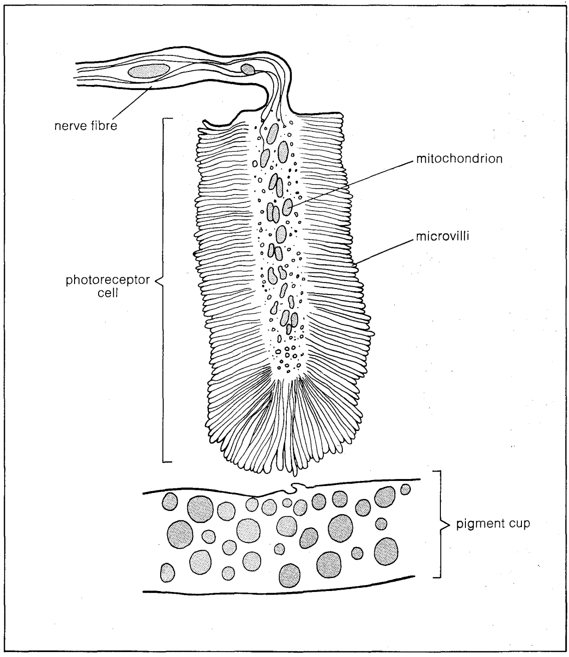


Figure 22: The photoreceptor cell in the flatworm *Dendrocoelum lacteum.*

The nemertine worms (phylum Nemertea) have photoreceptor cells that resemble those of planarians when viewed with the light microscope. The photoreceptor cells on the eye of *Lineus ruber* resemble those of planarians in the electron microscope in that both have a rhabdomere consisting of microvilli. An important difference exists, however: the photoreceptors of *L. ruber* possess a filament that resembles a ciliary rootlet, even though no evidence exists that the rhabdomere is derived from cilia; in fact, cilia have not been observed in this cell. The function of the filament is not yet known.

Photoreceptors in the eye of the rotifer *Asplanchna brightwelli* (phylum Aschelminthes) are lamellate. Thin lamellae, arranged as leaves of a cabbage, apparently are folds of the receptor membrane. There is no evidence that the receptor is derived from cilia.

The photoreceptor cells in the eyes of the arrowworm *Sagitta scrippsae* (phylum Chaetognatha) have rhabdomeres consisting of structures of ciliary derivation that are called microtubules. The microtubules arise from a cone-shaped body filled with granules and cordlike structures. The cone-shaped body is derived from a sensory cilium. The microtubules comprising the rhabdomere are 50 nanometres in diameter and 20 long.

Photoreceptors of annelid worms

The segmented worms (Annelida) are the first invertebrate animal group in which the photoreceptors of a considerable number of species have been investigated. Studies indicate that annelids generally have rhabdomeres consisting of microvilli and are not of ciliary origin; however, at least one exception (*Branchiomma*) to this pattern has been observed.

The photoreceptor cell of the earthworm *Lumbricus terrestris* is an example of a rhabdomere comprised of microvilli; as mentioned previously, the microvilli of the membrane of the rhabdomere border on a vacuole within the receptor cell rather than on its outer surface. A number of sensory cilia are mixed with the microvilli and extend from the cell cytoplasm into the vacuole. The microvilli are not derived from cilia, and their function is as yet unknown. Although rhabdomeres consisting of microvilli have been found in several other annelids, no evidence of cilia has been reported. In the polychaete worm *Nereis vexillosa,* however, the microvillar rhabdomere arises from a tubular extension of the cell containing a centriole and a small fibre (fibril) with the structure of a ciliary rootlet; no direct evidence has been found thus far to link these ciliary vestiges to the rhabdomere.

The photoreceptor of the polychaete *Branchiomma vesiculosum* is a well-documented example of an annelid photoreceptor derived from cilia. The photoreceptor cell has a large invaginated (inpocketed) cavity filled with about 450 flattened lamellar sacs, which are the expanded flattened membranes of cilia having the 9 + 0 configuration. A collection of mitochondria and the nucleus of the receptor cell are displaced from the light path.

The Onychophora represent a transitional group that combines features of both the annelids and the arthropods. The photoreceptor cells of several species of onychophorans have been studied with the electron microscope. The photoreceptor cell of *Peripatus* closely resembles that of some annelids. The rhabdomere, which consists of microvilli, has a sensory cilium near its base. In studies of the *Peripatus* eye during development, no connection has been found between the sensory cilium and the developing microvilli. The function of the sensory cilium in this and in other photoreceptors in which no clear developmental relationship exists between the two is obscure.

Arthropods have few cilia, although they do occur in a few organs and provide the developmental basis for one type of arthropod ear, the chordotonal organ. Cilia have not been found in association with arthropod photoreceptors, however. The photoreceptor cells of all arthropods studied thus far are generally similar. Part of the cell surface forms densely packed microvilli, which is the photoreceptor containing the visual pigment.

The presumed photoreceptor in echinoderms is a collection of loosely packed microvilli extending from each receptor cell into the central cavity of the ocellus. In some species, each receptor cell has at least one cilium with the 9 + 0 pattern; this pattern is presumed to exist in other species. No relationship exists between the cilia and the microvilli; thus, the same condition prevails as in certain mollusks, annelids, and *Peripatus.* The echinoderms studied thus far appear to have a rhabdomere composed of microvilli and in close association with a sensory cilium.

Photoreceptor cells in mollusks and cephalopods

The mollusks (*e.g.,* scallops, clams, squid) have two morphologically distinct types of photoreceptor cells. One contains microvilli; of these, some photoreceptors have rudimentary cilia and fibrils such as those of certain annelids. Most mollusks appear to have this type of photoreceptor. The other type contains cilia; about 100 cilia on one surface of the cell develop into flattened sacs that are either packed tightly, like slices of bread, or curl on one another, like cabbage leaves. The ciliary type of rhabdomere is found in *Onchidium verruculatum* and in *Cardium edule.* The eyes of the bivalves *Pecten, Chlamys, Spondylus,* and *Amussium* contain double retinas, each with one of the types of photoreceptor. Evidence indicates that both types are photoreceptors and that they have different functions. The distal retina, with a ciliary photoreceptor, responds to a decrease in illumination; the proximal retina, with a microvillar photoreceptor, responds to increasing illumination.

The cephalopod retina has long cylindrical photoreceptor cells with rhabdomeres consisting of microvilli. In some cells the long axes of the microvilli are horizontal. The other type has vertical microvilli. The axes of both types of microvilli are parallel to a plane that is tangent to the retinal surface. This orthogonal arrangement of the microvilli in different cells—*i.e.,* in which they are perpendicularly disposed to one another—forms the physical basis for polarized light detection in the cephalopods.

Vertebrate photoreceptors. The photoreceptor of such primitive chordates as the urochordate ascidian tadpole is a modified cilium with the $9 + 0$ pattern; the cilium has developed into numerous lamellae. The cephalochordate *Branchiostoma californiense* may have both microvillar and ciliary photoreceptors. The Hesse cells, shown experimentally to be photoreceptor cells, have typical microvillar rhabdomeres with no cilia. Two other types of cells that may have a photoreceptor function are the dorsal ependymal cells (*i.e.,* special cells lining the cavities of the brain) and cells of the infundibulum, a ventral portion of the brain. Both cells have appendages containing cilia that form lamellae.

Other vertebrates have the two types of photoreceptor cells mentioned before: rods and cones. The rods are the photoreceptors for vision under conditions of dim illumination; the cones mediate daylight vision and colour sensation in many animals. The photoreceptors in both the rods and cones are composed of stacks of disks derived from cilia; the lamellae are stacked at right angles to the long axis of the cell and constitute the outer segments of rod and cone cells.

Most of the lamellae of the outer segment of the rod consist of free-floating disks within the limiting membrane of the outer segment (see Figure 23). The free-floating disks are formed at the base of the outer segment by a process of continuous infolding of the membrane of the outer segment. The process of new disk generation proceeds continuously during the life of the animal, and the entire outer segment of the rod is replaced in about one week. As new disks are being formed at the base of the rod, disks at the other end of the outer segment are being broken off and ingested by pigment cells; the rod length thus remains constant. As infolded regions of the membrane at the base of the rod move outward, they are pinched off to form free-floating disks. The outer segments of the cones, sometimes conical, or cone-shaped, and shorter than the outer segments of the rods, do not have free-floating disks. The process of infolding stops just after formation of the outer segments of the cones; infoldings occur along the entire length of the cones (see Figure 23). Because the membrane of the outer segments of the cones is not formed continuously during life, membrane components are continuously replaced. It has been shown that renewal takes place continuously along the entire surface of the membrane in cones. In rods, on the other hand, the renewal occurs only at the base, as new disks form.

From R.W. Young, "Visual Cells," Copyright © 1970 by Scientific American, Inc., all rights reserved

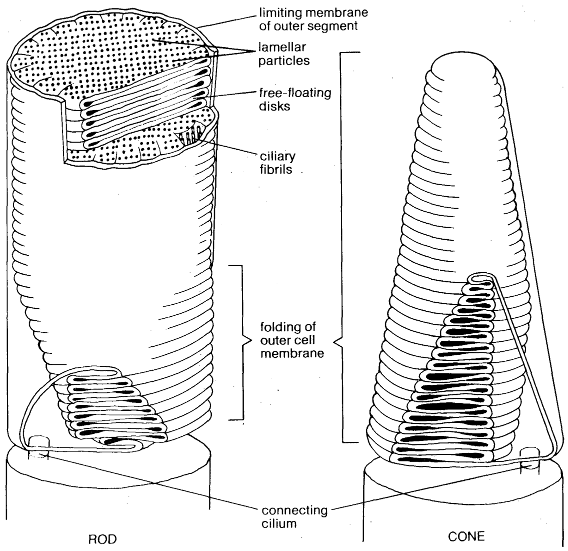


Figure 23: Structural relations between the outer segments of rods and cones (see text).

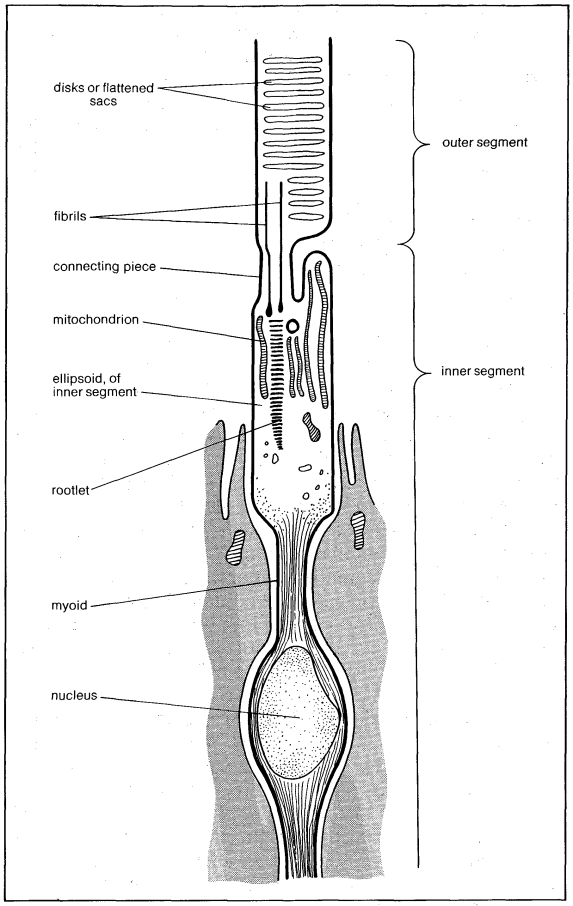


Figure 24: Part of a rod cell in the eye of a vertebrate.

From F.S. Sjostrand in G.K. Smelser (ed.), *The Structure of the Eye;* Academic Press, Inc., in D. Mazia and A. Tyler (eds.), *General Physiology of Cell Specialization,* Copyright 1963, McGraw-Hill Book Company, used by permission of McGraw-Hill Book Company

Structural differences between rods and cones

The terms rod and cone are misleading, because these receptors cannot always be identified by shape; considerable variation exists between species. There are, nevertheless, a number of morphological differences between the rods and cones. The outer segments of the rods of many species are deeply incised—*i.e.,* lengthwise clefts in the disks give the outer segment a scalloped appearance in cross section. Both the rods and cones contain dense aggregations of mitochondria in a section of the inner segment called the ellipsoid (see Figure 24). In the same region the cones of many species have an oil droplet, sometimes coloured, that filters the light before it reaches the receptor. An important difference between the photoreceptors of vertebrates and those of most invertebrates is that the latter give rise to nerve fibres that synapse either with cells in the central nervous system or with cells in a nerve ganglion within the eye; the vertebrate receptors, on the other hand, end in a region called a foot piece, into which nerve cells of the retina send synaptic projections.

Embryology and evolution of photoreceptors. There are two types of vertebrate eyes. The more familiar is the highly developed lateral eye. The other is the primitive median, or pineal, eye (in the top of the head). Only cyclostomes, reptiles, and amphibians have a median eye. It is best developed in lizards, in which the cornea is translucent and the number of receptors is small. The photoreceptor cells of the pineal eye are homologous (*i.e.,* structurally or developmentally related) with those of the lateral eye. They are, in fact, cones, the outer segments showing the characteristic infoldings along their entire length. In the median eye, the ends of the outer segments face the lens and light; in the lateral eye the retina is inverted, so that the ends of the outer segments face away from the light and toward the back of the eye.

The vertebrate retina is literally a part of the brain. The photoreceptor cells are derived from cells that line the neural tube, a hollow, dorsal structure that appears early in the embryo. The median eye develops from an outpocketing of the neural tube, and the cilia that develop into outer segments face the lens.

Invertebrates lack a neural tube; their photoreceptor cells develop from embryonic epithelial cells that send out axons—threadlike extensions—that grow into the central nervous system. Often the epithelial cells contain cilia, and the photoreceptors develop from the cilia. The ciliary photoreceptors of invertebrates do not appear to be homologous with those of vertebrates. As mentioned above, it is of particular interest that photoreceptors of diverse structure and in many different phyla develop either from cilia or in close association with them. Still unknown is the developmental and functional role of the sensory cilia that are frequently found in association with photoreceptors of apparently nonciliary origin.

Optical properties of photoreceptors. *Rhabdomeres and outer segments as light guides.* The densely packed membranous structures of photoreceptors have a higher refractive index than the surrounding substances. The refractive index of the rhabdomere of the blowfly (*Calliphora*), for example, is 1.349, and that for the surrounding substance is 1.336; any light reflected from the inside surface of the rhabdomere is called dense to rare reflection. When the angle at which the light strikes the inside surface (angle of incidence) is such that the change of the angle of the light when entering the photoreceptor (*i.e.,* the angle of refraction) is 90°, all of the light is reflected back into the rhabdomere. The angle of incidence at which total internal reflection occurs is called the critical angle. The rhabdomere and outer segment trap the light entering them at angles equal to or greater than the critical angle and propagate it, essentially without loss of energy. The rhabdomere and outer segment thus act as an optical wave guide, or light pipe (see OPTICS).

Although all light entering the photoreceptor organelle—*i.e.,* all light incident at angles equal to or less than the critical angle—is propagated by successive internal reflection, only certain angles of reflection occur; these are determined by various physical characteristics of the organelle. As a result, the light within the photoreceptor light guides is propagated in modes, or patterns, of energy. This is important for the photoreceptive function of the rhabdomere and outer segment because the amount of information the organelle can carry is related to the number of modes propagated. The thinnest photoreceptor organelles, such as those of certain fly rhabdomeres with a diameter of 0.5 nanometre, can, for example, support only one mode, called the lowest order mode. The amount of energy propagated in the rhabdomere in this mode depends on the wavelength of light, so that the shorter wavelengths are best propagated. Thus reddish light is not made available to the rhabdomere's visual pigment as readily as bluish light; the light guide properties of the organelle, therefore, can influence the physiological response of the photoreceptor. The light guide also profoundly affects the amount of image information transmitted because the greater the number of modes, the more image information. This last consideration is of particular importance for fused rhabdoms, in which image information apparently is not transmitted (although the question is not yet entirely settled). These light guide properties are of importance for still another reason. A small amount of the modal energy propagated by a light guide is actually outside the structure. Thus when, as described in the next section, pigment granules come near the photoreceptor organelle, they control the amount of light in the organelle by absorbing the energy propagating outside, so that it can no longer be transmitted by the photoreceptor. Such pigment granules may exert an even more profound effect by raising the refractive index outside the organelle, thus destroying its guiding properties and causing the light to spread into adjoining structures where it is absorbed, but not by visual pigment; absorption thus does not result in sensation.

Migration of pigment granules

Photomechanical light and dark adaptation. A mechanism that may partly control the amount of light within the photoreceptor under widely different illumination conditions involves the migration of photoprotective pigment granules either within the receptor or within neighbouring cells so as to envelop the photoreceptor; it has been observed in both vertebrates and invertebrates. Pigment migration in response to illumination takes place in the retinas of most vertebrates except mammals. The most pronounced effects are found in fishes, frogs, and toads; they also occur in reptiles and in diurnal and nocturnal birds.

From Dr. Samuel R. Detwiller, *Vertebrate Photoreceptors* (1943); Macmillan & Co.

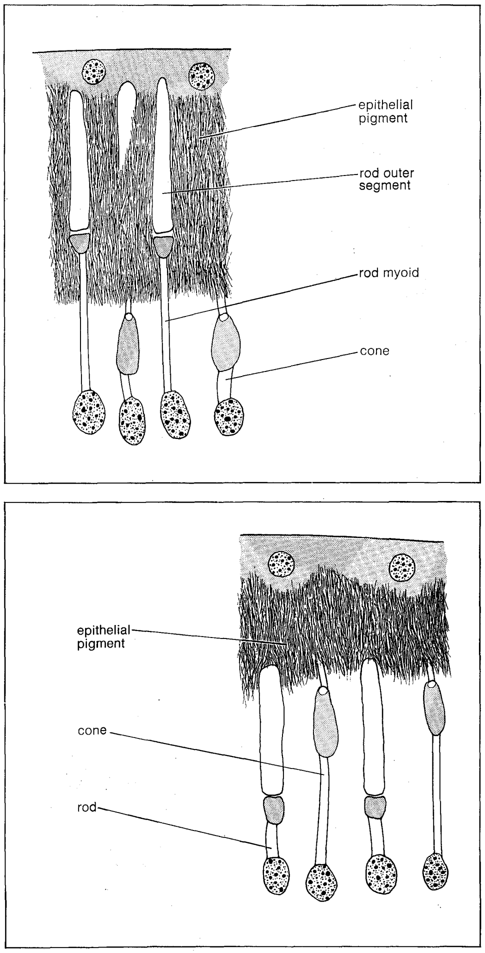


Figure 25: *Adaptation of retina.*
(Top) Theoretical light-adapted retina, showing migrated epithelial pigment, elongated rods, and contracted cones. (Bottom) Theoretical dark-adapted retina, showing contracted pigment, elongated cones, and contracted rods.

Pigment migration is combined with photomechanical movements of the photoreceptor cells of many lower vertebrates (Figure 25). In darkness the pigment granules of the pigment epithelium are withdrawn, as shown in the figure. The rod myoids (contractile elements) are contracted in darkness so that the rod outer segments are withdrawn from the pigment and exposed fully to the available light. On the other hand, the cone myoids are extended, so that their outer segments are within the pigment, which functions effectively to reduce the amount of light in the outer segment, as explained in the previous section. The cone outer segments are enveloped by pigment, presumably to help suppress responses from the daylight receptors under conditions of dim illumination. Under the influence of light the rod myoids extend within several minutes, and the pigment envelops the rod myoids to help suppress the diurnal activity of the rod system (pigment surrounds only the rod myoids, not the outer segments [see Figure 25, top]). At the same time the myoids of the cones contract to withdraw the cone outer segments from the pigment. As mentioned above, neither photomechanical nor pigment movements occur in mammalian eyes; switching between rod and cone systems is accomplished neurally.

The envelopment of rhabdomeres by pigment migration is widespread in invertebrates. An example of this phenomenon occurs in the ommatidium of the amphipod *Gammarus ornatus,* in which the rhabdomeres of five radially arranged receptor cells form a star-shaped rhabdom. Extensions of the receptor cells are directed toward the cornea next to the lenslike crystalline cone and away from the cornea to cell bodies beneath a so-called basement membrane. In the dark-adapted condition the pigment is located around the cone and in the cell body. The rhabdom functions as a light guide within which is the visual pigment, which absorbs light. In the light-adapted condition, the pigment migrates from the receptor cell body to surround the rhabdom completely. Similar cell pigment migrations occur within the receptors of many invertebrates. In the compound eye of the horseshoe crab *Limulus,* the rhabdom is also star-shaped in cross section, but the rays of the star are much longer than are those of *Gammarus.* In the eye of *Limulus* the pigment migrates radially within the receptor cell, completely enveloping the rays in a matter of minutes in bright light and requiring about one hour to withdraw entirely from the rhabdom area in darkness.

The rhabdom of the migratory locust (*Schistocerca*) shows an interesting variation. In the light-adapted retina, mitochondria migrate and become tightly packed around the rhabdom. This may be a device for providing energy for the rapidly metabolizing light-activated photoreceptor. There is physiological evidence, however, that, in the light-adapted condition, the mitochondria, which have a high refractive index, increase the critical angle by coming close to the rhabdom. They thereby cause loss of light from the rhabdom and a decrease in the "field of view" of the rhabdom. When the eye is dark-adapted, the space around the rhabdom is replaced by fluid-filled spaces, with a low refractive index; they return the rhabdom to an efficient light guide mode of operation.

Birefringence and dichroism. The physical properties of certain crystals, glass, liquid, and gas, are the same regardless of the direction of the light propagated through them; for example, the speed and thus the refractive index of light are the same regardless of its direction of propagation. Such a medium is said to be isotropic. In other crystals, the atoms are arranged so the refractive index varies with the direction of propagation of light; such crystals are said to be anisotropic. If a medium has a crystalline arrangement that retards light in a particular direction, it is birefringent. In such an arrangement an entering ray of light is divided into two rays that are polarized in planes at right angles to one another.

When birefringence can be made to disappear by immersing a substance in a liquid with an appropriate refractive index, the birefringence is caused by an orderly arrangement of isotropic particles submicroscopic in size, such as the limiting membranes surrounding cells. Such birefringence is called form birefringence to distinguish it from the intrinsic molecular crystalline birefringence described above. Very small rods, lying parallel to each other, cause positive birefringence; parallel platelets cause negative birefringence. Studies have shown that rods in frogs have a negative form birefringence caused by the platelets and a positive intrinsic birefringence. These phenomena are thought to result from the ordered arrangement of a lipid (fat) layer two molecules thick in the platelet membranes.

Birefringence in the rods of frogs

Variation in the colour of light absorbed dependent on the direction of polarization of the light is termed dichroism. This property is a sensitive indicator of the orientation of molecules in a structure; dichroism in photoreceptors, for example, results from the ordered arrangement of the visual pigment molecules. The visual pigment of outer segments is dichroic, as are the outer segments. Dichroism of the outer segment can be demonstrated only by measuring the absorption of polarized light shone through the side of the outer segment. Light that propagates in the usual direction, that is down the long axis of the outer segment, does not show different absorption properties that depend on the direction of polarization. The fact that there is not dichroism for light propagating along the long axis, but there is for light propagating perpendicular to the long axis, indicates that the visual pigment molecules are oriented at random but with the long axes in the plane of the disks.

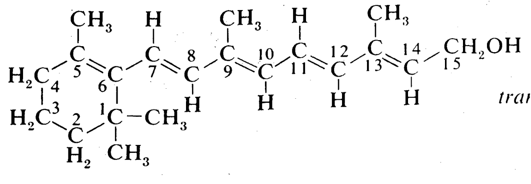
Rhabdoms consisting of microvilli can detect the plane of polarization of polarized light. Invertebrates with such rhabdoms use polarization properties of the blue sky opposite the Sun for navigation. In animals with the ability, the rhabdomeres comprising a rhabdom show orthogonal orientation. It has been shown that the microvilli are dichroic and that the dichroism is caused by the visual pigment. The orientation of the microvilli provides the mechanism for sensation of polarized light. Vertebrates cannot detect polarized light because of the random orientation of visual molecules within the plane of the platelets.

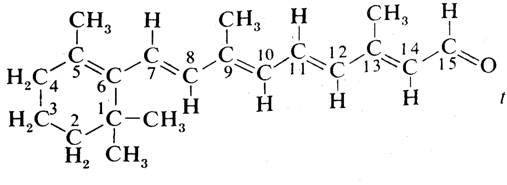
Stiles-Crawford effect. In the Stiles–Crawford effect, first observed in humans, the luminous efficiency of light has been shown to be greater when entering the centre than when entering the edge of the pupil. This effect may be considered an optical property of the cone receptor cells. The energy acceptance, or radiation, pattern of photoreceptors is a consequence of the optical wave guide properties resulting from the small size of the receptors. The Stiles–Crawford effect is not observed with dim illumination (rod vision), probably because the acceptance angle of the rods is wider, and it essentially fills the pupil.

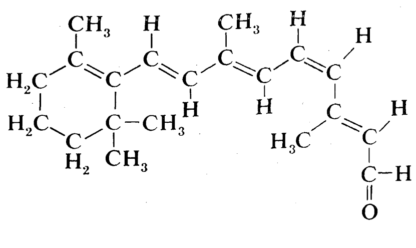
Visual pigments. As light propagates through the photoreceptor, the visual pigment absorbs photons (light particles). The absorption of one photon by a visual pigment molecule can initiate events that lead to nervous excitation. The photoreceptors in the eyes of both vertebrates and invertebrates consist of tightly packed membranes (lamellae), the function of which is to support the visual pigment and other molecules necessary for excitation. The photoreceptor membrane is a lipid layer two molecules thick, with protein either between the layers, coated over the layers, or both. The outer segment of the frog rod is 40 percent lipid and 60 percent protein; more than 80 percent of the protein is visual pigment.

Components. Visual pigment has two components: the light-absorbing chromophore (a chemical group that produces colour) and the protein moiety (opsin) to which it is chemically attached. The structure of the naturally occurring chromophore, retinal, is thought to be the same in all animals. Before it has been exposed to illumination, the chromophore is 11-*cis*-retinal, an alternate structure of vitamin A aldehyde. Upon exposure to light each

Characteristics of retinal


trans-retinol


trans-retinal


11-*cis*-retinal

chromophore molecule absorbs one photon of light. The absorption of a photon changes 11-*cis*-retinal to *trans*-retinal. (See structural formulas.) The resulting change in the shape of the molecule causes the separation of retinal from its protein component and gives rise to a number of coloured intermediates. In the dark, 11-*cis*-retinal combines spontaneously with the protein component to regenerate the visual pigment.

Hundreds of visual pigments have been identified and characterized by the wavelength of light that is best absorbed by them (absorption maximum, or λ max); the wavelength maxima range from 432 nanometres for the green rod of the frog to 625 nanometres for the red-absorbing cone of the goldfish. The reason for such a great range when the chromophore is the same in all visual pigments lies in part in properties of the chromophore; much of the variation, however, resides in the protein component. Variation in protein structure among different animals gives rise to different absorption maxima.

There are two forms of vitamin A: retinol A_1 and retinol A_2. The structure of A_1 differs only slightly from that of A_2. The effect of this slight structural difference is to increase the wavelength for maximum absorption by as much as 20 to 60 nanometres, depending on the pigment involved. The pigments with the shortest λ max are A_1 pigments; those with the longest are A_2. There is a broad area of overlap within the two groups, however. In general, marine fishes, land vertebrates, arthropods, and mollusks have A_1 pigments; freshwater fishes have A_2 pigments.

The protein components, or opsins, of visual pigments have not yet been fully characterized. The visual pigment of rods, a combination of retinal and opsin, is called rhodopsin, or visual purple. Estimates for the molecular weights of the opsins found in rods centre around 40,000. Squid opsin, however, has a molecular weight of about 70,000. Analyses of the visual pigments from the rods of squid and vertebrates suggest that the protein components are quite different. The nature of cone pigment proteins is thus far virtually unknown.

Important to an understanding of the photochemical reaction is the nature of the attachment of the chromophore to the opsin. Although controversy still exists on this point, it is clear that the visual pigment is part of the photoreceptor membrane and is very closely attached to its lipid and protein.

The role of visual pigments in vision. The sensitivity of the dark-adapted human eye can be measured as a function of wavelength of the stimulating light falling on that part of the retina in which the density of rods is greatest. When such a curve is compared with a curve of the absorption spectrum (the entire range of wavelengths of light absorbed) of human rhodopsin in the outer segments of the rods, the two curves correspond exactly. This is strong evidence that rhodopsin initiates the visual process for light absorbed in the outer segments.

The problem of cone pigments is an important one. More than 100 years have passed since it was concluded after a comparative study of retinas that the rods of vertebrate retinas mediate vision in dim light and the cones mediate daylight and colour vision. Recently, a method called microspectrophotometry has been developed for the study of cone pigments and other visual pigments. By projecting a minute beam of light onto a single photoreceptor, the absorption spectra of a variety of vertebrate and invertebrate photoreceptors have been measured. The absorption spectra of single human and other primate cones have thus been characterized. There are three types: a blue absorbing cone, a green absorbing one, and a yellow absorbing one. Originally suggested in 1802, these three cone types, now positively identified, can quantitatively explain human colour sensation.

Absorption spectra of photoreceptors

Such methods have been helpful in identifying receptors and their visual pigments in animals other than man; for example, in the retina of the frog *Rana pipiens,* a green rod and a red rod have been identified. A double cone with principal and accessory pigments and a single cone have also been characterized. Microspectrophotometry has been used to obtain evidence of visual pigments in rhabdomeres of the fly *Calliphora.*

PHYSIOLOGICAL RESPONSE OF PHOTORECEPTORS

The initial molecular events of photoreception can be summarized: after the absorption of a photon of light, 11-*cis*-retinal is converted into *trans*-retinal; the process disrupts the binding of retinal to opsin and results in the sequential formation of a number of coloured photoproducts.

The final physiological effects of photoreception have also been well defined. Excitation of visual pigments results in a change in certain properties (*i.e.,* permeability and potential) of the limiting membrane of the photoreceptor cell. Because the initial and final molecular events of photoreception are now understood, there is considerable hope that the entire chain of molecular events between the absorption of light by the visual pigment and the changes in potential of the cell membrane can eventually be reconstructed. This challenging problem has stimulated research concerned with the nature of the photoreceptor membrane and the chemistry of the photoproducts.

The internal environment of cells is generally different from that outside. The main cation (positively charged atom) inside cells is usually potassium; the main cation outside is sodium. In the resting, or unexcited, state, the cell membrane of most excitable cells (such as receptor, nerve, and muscle cells) is impermeable to sodium ions but permeable to potassium ions. A separation of charges occurs on either side of the membrane because the positively charged potassium ions diffuse outside of the cell, leaving a surplus of nondiffusing negatively charged atoms (anions) inside the cell. The result is a difference in potential across the cell membrane; this difference is proportional to the logarithm of the ratio of the concentration of potassium ions outside to that inside the cell. Photoreceptor cells are negative by some 30 to 60 millivolts (1 mV = 0.001 volt) inside relative to the outside, the exact value depending on the specific cell. This negative potential difference is termed the resting potential.

Definition of receptor potential

Receptor potential. The absorption of a single photon by the visual pigment in a photoreceptor may result in a physiological response of the receptor cell. Such a response of the cell can be measured as the receptor potential—that is, a change in the potential of the cell membrane caused by a change in the permeability of some part of the membrane to certain ions. The permeability change is in turn related to the absorption of light by the chromophore through a chain of molecular events as yet unknown. The receptor potentials of the eyes of animals from a number of phyla have been investigated. In certain photoreceptors the receptor potential is positive in sign, or depolarizing; in others it is negative, or hyperpolarizing.

Depolarizing receptor potentials. The depolarizing receptor potential, a change in potential response to illumination that is positive in sign, is characteristic of photoreceptor cells in many invertebrate phyla. One of the best known photoreceptors is the retinular cell in the compound eye and in the ventral eye of the horseshoe crab *Limulus,* an arthropod. The membrane potential of this cell is easily measured by means of a very small electrode (a device that can be placed inside a cell to measure the potential across the cell membrane). When the electrode penetrates the cell in the absence of light, the potential falls to some negative value, perhaps −40 millivolts, which is the resting potential. Even in complete darkness, however, variation in the resting potential occurs in the form of slow fluctuations. Although these spontaneously and randomly occurring fluctuations presumably seldom result in a physiological visual response, they are miniature receptor potentials. Each is accompanied by a decrease in the resistance of the membrane (*i.e.,* it becomes a better conductor) and an increase in the permeability of the membrane to sodium ions. When the receptor is illuminated with dim light, the slow potential fluctuations increase in frequency, and, with brighter light, they become briefer and higher in frequency, finally fusing to form a more or less steady depolarizing receptor potential. Thus, a short time after the cell is exposed to light, the membrane becomes depolarized; when the light is removed, the receptor potential decays, and the resting potential is restored. The extent of the receptor potential increases with increasingly stronger illumination; it has a linear relationship to the logarithm

Relationship between receptor potential and illumination of the intensity of illumination. Whatever properties are responsible for this transformation are responsible for the photoreceptor's enormous dynamic range: despite the fact that the receptor potential can change at most on the order of 100 millivolts, the receptor can respond over a range of more than 10 orders of magnitude, a range of intensities of over 100,000,000,000.

In all species with this depolarizing receptor potential, the fact that the resistance of the photoreceptor cell membrane decreases during the receptor potential suggests that it becomes more permeable to some ion or ions during this time. Evidence as to the nature of the permeability change suggests that it is principally an increase in permeability to sodium ions that causes the receptor potential. There is additional evidence for this conclusion (see *Photoreception* in *Bibliography*).

The depolarizing receptor potential is characteristic of the arthropod photoreceptor consisting of microvilli and the receptor cell. Species from most of the orders of Arthropoda have now been investigated, and exceptions have not been reported. Usually, the photoreceptor cell gives rise only to a receptor potential. The receptor potential is conducted passively a short distance to higher order nerve cells via a nerve fibre. The higher order nerve cells convert the receptor potential to propagated action potentials (nerve impulses) for transmission of the information over longer distances.

The proximal sense cell of the double retina of *Pecten* is similar to the above in that illumination gives rise to a depolarizing receptor potential and a decrease in resistance. The morphology of the proximal sense cell is also similar to that of the arthropod photoreceptors in that it has coarse microvilli. But the proximal sense cell of *Pecten* has a long nerve fibre that carries nerve impulses. The depolarizing receptor potential activates the mechanism that initiates an action potential, and steady depolarization of the axon gives rise to a series of nerve impulses; the stronger the intensity of illumination, the greater the extent of the receptor potential and the higher the frequency of nerve impulses. Nerve activity of this nature is rarely found in arthropod photoreceptor cells.

In addition to the arthropods and *Pecten,* a proximal sense cell in the leech (*Hirudo*), an annelid, has the depolarizing type of receptor potential.

The off response

Hyperpolarizing receptor potential with decrease in resistance. As noted above, the proximal sense cells of the *Pecten* eye have a depolarizing receptor potential, and the nerve responds to illumination with a series of nerve impulses. The distal sense cells, which also have nerve fibres, respond to a decrease in illumination with nerve impulses. This is called an off response. Recent studies indicate that the receptor potential of the distal sense cell is a hyperpolarizing potential, a potential change that is negative in sign. When the cell is depolarized, the extent of the receptor potential increases; when the cell is sufficiently hyperpolarized, the receptor potential reverses polarity and becomes depolarizing. The receptor potential of the distal sense cell is accompanied by a decrease in resistance. The hyperpolarizing potential is believed to be caused by an increase in permeability to potassium and perhaps to other ions.

The photoreceptor of the *Pecten* distal sense cell is a ciliary derivative. A primitive chordate, the tunicate *Salpa,* which has a microvillar photoreceptor, has the same type of hyperpolarizing photoreceptor potential. There thus does not seem to be any association of physiological potential types with structurally or embryologically similar photoreceptors.

Hyperpolarizing receptor potential with increase in resistance. The rod and cone photoreceptors of vertebrate eyes have a third type of receptor potential. In the dark, sodium ions flow steadily into the outer segment; light decreases this dark current—*i.e.,* the cell becomes hyperpolarized, and its resistance increases. If the cell is sufficiently hyperpolarized under experimental conditions, depolarization occurs; the receptor potential thus is reversed, as would be expected if the permeability change involves principally sodium ions.

Interpretation of the physiological response. As a nerve cell is depolarized, it discharges nerve impulses; as it becomes hyperpolarized, the membrane potential returns to the resting potential. The fact that certain photoreceptor cells are hyperpolarized by illumination is thus confusing and raises the possibility that light is inhibiting the receptors. In the generation and transmission of information about illumination, however, the sign of the response may actually be unimportant. It has been known for some time that there are several types of optic nerve fibres in vertebrate eyes; some discharge impulses during illumination, others discharge either only at the onset and cessation of illumination or merely at the cessation. In other words, information about the cessation of illumination is at least equal in importance to information about its onset and duration. Such observations of optic nerve activity prove that the hyperpolarizing response of the receptor, the only response of the receptor, is translated into a code that includes the discharge of nerve impulses both during and after illumination. Thus, the hyperpolarizing response of the receptor, although it may be important in emphasizing the off response, does not result in the loss of information about excitation. Because the hyperpolarizing response is definitely used to produce excitation at a later stage, it is misleading to consider it an inhibitory response. The same is true in reverse for the depolarizing response.

The way in which initial photoreception is processed by the retina and higher centres to cause visual perception is beyond the scope of this article. It may be mentioned, however, that the initial information used in this processing is the direct result of the physical and chemical properties of the photoreceptor. (W.H.M.)

Sound reception

Sound waves are a particular kind of mechanical activity consisting of vibrations in a gas, liquid, or solid medium. If an animal possessing an auditory mechanism comes in suitable contact with a medium vibrating at a frequency and intensity within its range of aural (hearing) sensitivity, it may hear the sound. For land animals, the usual vibrating medium is the air; for fishes and other aquatic creatures, it commonly is the water. Yet, under suitable conditions, all hearing animals can perceive sound waves transmitted by media other than the one in which they live; thus, humans can hear noise while underwater. (For a detailed discussion of human sound reception, see below *Human hearing*; additional information is contained in the article SOUND.)

In the course of evolution, animals have developed a variety of sense organs that respond to mechanical stimuli. There are at least 10 of these mechanoreceptors in vertebrates and perhaps as many in advanced invertebrates. Not all of these structures respond to sound, however, for among them are the simple touch endings of the skin and the motion receptors that serve (mediate) bodily equilibrium (see above *Mechanoreception*). Although the different ways of registering mechanical changes in the environment or within the body represent various structural specializations, it is not feasible to identify any one of them simply in terms of its structure; many different mechanisms, cells, or organs may perform similar functions. Ears, for example, take many forms in the lower animals and often have little resemblance to these organs in man and other higher vertebrates. Yet the service that they perform in sound reception is similar enough that they may be called ears.

Although there is no fossil record of the origin and development of auditory structures, in animals with ears the evolutionary process in every instance appears to have been a conversion to an auditory function of structures that previously mediated a simpler form of mechanoreception. Indeed, any mechanoreceptor, even though best adapted to respond to some other form of mechanical stimulation, will respond to vibrations within some region of the sound frequency range if the vibrations have a sufficiently high level of intensity.

Attempts to define hearing

Many attempts have been made to define hearing, often with indifferent success. The task is difficult, and in certain respects the lines of distinction are arbitrary. The ear cannot be identified by any standard structure, nor can it be identified in terms of the stimulus as simply a receiver of sound vibrations. As noted above, mechanical receptor organs will respond to sound vibrations within some region of the frequency range if a sufficiently high level of intensity is provided. Moreover, the ear cannot be characterized in terms of the physical principles by which it operates because these principles vary among the ears of different animal species.

A definition of hearing, therefore, must be sought in terms of the ear's specialization of function and the relative effectiveness with which it performs this function. Thus, hearing may be characterized as the reception of sound vibrations by an organ, the ear, that has developed for this particular purpose and that has reception of sound as its primary function. This definition excludes the reception of sound vibrations by touch (tactual) endings in the skin, for example, because these structures respond most readily to direct pressure. Before such receptors will respond to sound waves, the vibrational intensity of the sound must be relatively great. Also excluded are the hair sensilla, of which arthropods have many types, whenever it can be shown that these organs respond with greater sensitivity to another stimulus (most often a simple direct deflection of the central hair).

Principles of sound reception

Theoretically, several aspects of vibration might serve in its detection by an ear. These characteristics include the amplitude (extent) of the motion of particles (*e.g.*, molecules) in a medium, the velocity and acceleration of the motion, the pressure exerted upon an obstacle in the path of the sound waves, and temperature changes occasioned by the vibrations. All of these manifestations have been utilized in attempts to design microphones for the detection and measurement of sound, but only two (pressure and velocity effects) have proved to be of any practical value. Thus, those devices that employ these two effects are known as pressure and velocity microphones.

It seems more than coincidence that these same two aspects of sound, pressure and velocity, are the only stimulus characteristics on which the evolution of ears appears to have been based. Moreover, just as the pressure microphone is the most practical type designed by man, among ears the pressure type is the most widespread and the most highly developed. Ears that distinguish changes in velocity have appeared only in a few lower animals—as an elaborated hair organ in some insects and perhaps spiders and in two special forms among fishes. All other ears are pressure receptors that have taken two lines of evolutionary development, one in most of the insects and another in vertebrates above fishes.

Types of animals that have ears. Considering the usefulness of the sense of hearing to such highly organized animals as man, it may seem surprising that this sense is so limited in its appearance and development among animals. It is found only in two major groups of animals: arthropods (*e.g.*, insects and crabs) and vertebrates (*e.g.*, amphibians, birds, and mammals). The condition that probably limited the development of hearing in other species was the lack of sufficient advancement and flexibility of the nervous system.

Biological value of ears

In those animals with auditory structures, hearing serves purposes of great biological value: in its more primitive forms, it is used to sense danger and enemies, to detect prey, and to identify prospective mates; at a more complex level, hearing is involved in communication within social groups and in emotional expressions of various kinds. The cry of an infant mouse that has strayed from the nest elicits a response by the mother to retrieve it. The singing of a male thrush asserts a claim to its territory, attracts a female to the area, and warns off other males. Among higher mammals (*e.g.*, monkeys and apes) vocalizations show even greater variety and express a range of meanings that may be interpreted in human terms as expressions of such concepts as danger, aggression, love, and the availability of food. In man, the elaborations of auditory communication can be even more symbolically complex, extending to speech and music (see below *Human hearing*; see also SPEECH). The significant features in complicated sounds that people perceive and differentiate correspond to the physical dimensions of frequency (the number of waves, cycles, or vibrations per second), intensity, phase, complexity of wave form, and temporal pattern. The variety of distinguishable acoustic forms is enormous.

Among the most highly refined applications of the auditory sense are those found in such animals as bats and dolphins. These creatures are able to discern objects around them by a process called echolocation; the animal sends out a cry and, by the nature of the echo, is informed of the presence of obstacles or potential prey. For these animals, the sense of hearing provides a service in the dark that closely approaches the reliability of vision in the perception of objects and spatial relations.

ORGANS OF SOUND RECEPTION IN INVERTEBRATES

It has long been believed that at least some insects can hear. Chief attention has been given to those that make distinctive sounds (*e.g.*, katydids, crickets, and cicadas) because it was naturally assumed that these insects produce signals for communication purposes. Organs suitable for hearing have been found in insects at various locations on the thorax and abdomen and, in one group (mosquitoes), on the head.

Among the many orders of insects, hearing is known to exist in only a few: Orthoptera (crickets, grasshoppers, katydids), Homoptera (cicadas), Heteroptera (bugs), Lepidoptera (butterflies and moths), and Diptera (flies). In the Orthoptera, ears are present, and the ability to perceive sounds has been well established. The ears of katydids and crickets are found on the first walking legs; those of grasshoppers are on the first segment of the abdomen. Cicadas are noted for the intensity of sound produced by some species and for the elaborate development of the ears, which are located on the first segment of the abdomen. The waterboatman, a heteropteran, is a small aquatic insect with an ear on the first segment of the thorax. Moths have simple ears that are located in certain species on the posterior part of the thorax and in others on the first segment of the abdomen. Among the Diptera, only mosquitoes are known to possess ears; they are located on the head as a part of the antennae.

Location of insect ears

All the insects just mentioned have a pair of organs for which there is good evidence of auditory function. Other structures of simpler form that often have been considered to be sound receptors occur widely within these insect groups as well as in others. There is strong evidence that some kind of hearing exists in two other insect orders: the Coleoptera (beetles) and the Hymenoptera (ants, bees, and wasps). In these orders, however, receptive organs have not yet been positively identified.

Types of insect auditory structures. Four structures found in insects have been considered as possibly serving an auditory function: hair sensilla, antennae, cercal organs, and tympanal organs.

Hair sensilla. Many specialized structures on the bodies of insects seem to have a sensory function. Among these are hair sensilla, each of which consists of a hair with a base portion containing a nerve supply. Because the hairs have been seen to vibrate in response to tones of certain frequencies, it has been suggested that they are sound receptors. It seems more likely, however, that the sensilla primarily mediate the sense of touch and that their response to sound waves is only incidental to that function.

Antennae and antennal organs. Many sensory functions have been attributed to the antennae of insects, and it is believed that they serve both as tactual and as smell receptors (see above *Chemoreception*). In some species, the development of elaborate antennal plumes and brushlike terminations has led to the suggestion that they also serve for hearing. This suggestion is supported by positive evidence only in the case of the mosquito, especially the male, in which the base of the antenna is an expanded sac containing a large number of sensory units known as scolophores. These structures, found in many places in the bodies of insects, commonly occur across joints or body segments, where they probably serve as mechanoreceptors for movement. When the scolophores are associated with any structure that is set in motion by sound, however, the arrangement is that of a sound receptor.

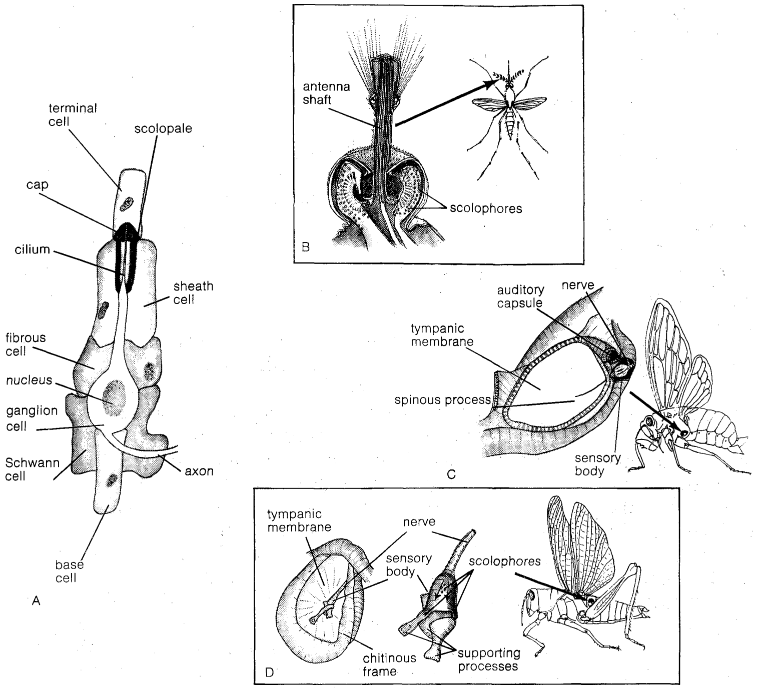


Figure 26: *Auditory mechanisms in insects.*
(A) A scolophore organ. (B) The mosquito ear. (C) The ear of the cicada *Magicicada septendecim*. (D) The ear of the grasshopper.

By courtesy of (B, right) the U.S. Department of Agriculture, and (B, left) *Acustica*, (1953) pages 335-343, and (C, right) H. Weber, *Grundriss der Insektenkunde* (4th ed. Ed. Gustav Fischer, Stuttgart,1966), and (D, right) reprinted from Anna Botsford Comstock: *Handbook of Nature Study*, Copyright, 1939, by Comstock Publishing Company, Inc. Used by permission of Cornell University Press, and (D, left and centre) E.G. Gray, *Philosophical Transactions* B. 243, 77 (1960)

Structure of the scolophore

The basic structure of the scolophore is shown in Figure 26. Four cells (base cell, ganglion cell, sheath cell, and terminal cell), together with an extracellular body called a cap, constitute a chain. Extending outward from the ganglion cell is the cilium, a hairlike projection that, because of its position, acts as a trigger in response to any relative motion between the two ends of the chain. The sheath cell with its scolopale provides support and protection for the delicate cilium. Two types of enclosing cells (fibrous cells and cells of Schwann) surround the ganglion and sheath cells. The ganglion cell has both a sensory and a neural function; it sends forth its own fibre (axon) that connects to the central nervous system.

In the mosquito ear (Figure 26) the scolophores are connected to the antenna and are stimulated by vibrations of the antennal shaft. Because the shaft vibrates in response to the oscillating air particles, this ear is of the velocity type. It is supposed that stimulation is greatest when the antenna is pointed toward the sound source, thereby enabling the insect to determine the direction of sounds. The male mosquito, sensitive only to the vibration frequencies of the hum made by the wings of the female in his own species, flies in the direction of the sound and finds the female for mating. For the male yellow fever mosquito, the most effective (*i.e.,* apparently best heard) frequency has been found to be 384 hertz, or cycles per second, which is in the middle of the frequency range of the hum of females of this species. The antennae of insects other than the mosquito and its relatives probably do not serve a true auditory function.

Cercal organs. The cercal organ, which is found at the posterior end of the abdomen in such insects as cockroaches and crickets, consists of a thick brush of several hundred fine hairs. When an electrode is placed on the nerve trunk of the organ, which has a rich nerve supply, a discharge of impulses can be detected when the brush is exposed to sound. Sensitivity extends over a fairly wide range of vibration frequencies, from below 100 to perhaps as high as 3,000 hertz. As observed in the cockroach, the responses to sound waves up to 400 hertz have the same frequency as that of the stimulus. Although the cercal organ is reported to be extremely sensitive, precise measurements remain to be carried out. It is possible, nevertheless, that this structure, which is another example of a velocity type of sound receptor, is primarily auditory in function.

Tympanal organs. The tympanal organ of insects consists of a group of scolophores associated with a thin, horny (chitinous) membrane at the surface of the body, one on each side. Usually the scolophores are attached at one end by a spinous process to the tympanic membrane (eardrum); the other ends rest on an immobile part of the body structure. When the membrane moves back and forth in response to the alternating pressures of sound waves, the nerve fibre from the ganglion cell of the scolophore transmits impulses to the central nervous system. Because the tympanic membrane is activated by the pressure of sound waves, this is a pressure type of ear.

Simple tympanal organs, such as those found in moths, contain only two or four elements, or scolophores. In cicadas, on the other hand, these organs are highly developed; they include a sensory body (a number of scolophores in a capsule) that may contain as many as 1,500 elements. Shown in Figure 26 is the structure of the tympanal organ in the 17-year cicada (*Magicicada septendecim*).

With 80 to 100 scolophores, the grasshopper ear, which has been studied more thoroughly than any other insect ear, is structurally between that of moths and cicadas. Ordinarily, the tympanic membrane is hidden beneath the base of the insect's wing cover. A bundle of auditory nerve fibres runs from one side of the sensory body, which lies on the inner surface of the membrane (Figure 26), and joins other nerve fibres of the region to form a large nerve extending to a ganglion (nerve centre) in the thorax.

Evidence of hearing and communication in insects. *Behavioral observations.* That the insect ear serves an auditory purpose has been proved by a large number of experimental observations, particularly those that have dealt most extensively with katydids and crickets. Males of these groups produce sounds by stridulation, which usually involves rubbing the covers of the wings together in a particular way. One wing has a serrated surface (a "file") that runs along an enlarged vein; the other wing has a sharp edge over which the file is scraped. The scraping causes the wing surfaces to vibrate; the natural resonances of the vibrations and the particular rhythm and repetition rate of the scraping movements determine the nature of the song, which varies with each species. Most females

Stridulation

are silent, but those of a few species have a poorly developed stridulatory apparatus, and weak sounds have been reported. Both males and females have tympanal organs for sound reception.

The observation that the males of many insect species produce repeated stridulatory sounds during the mating season led to the inference that the primary purpose of these noises was to attract a female. That this is indeed the case was first established by the extensive observations of the Yugoslavian entomologist Ivan Regen, who worked over the period 1902–30 mostly with a few species of katydids and crickets. In one of his earliest experiments, Regen proved (1913–14) that a male katydid of the species *Thamnotrizon apterus* responds to the sound of another male by chirping. The first male responds in turn to the second male's chirp, and the two insects then set up an alternating pattern of chirping. Although this pattern had been observed earlier, Regen was the first to prove by a series of experiments that it depends upon the sense of hearing. After removing the forelegs, on which the tympanal organs are located, of certain males, he found that even though these insects continued to stridulate, they did so only in individual rhythms that were not affected by the sounds of other males. Any alternation of chirping between deafened males, or between a deafened and a normal male, occurred only rarely, for brief times, and by chance.

A long series of check experiments by Regen showed that other stimuli, such as light, odours, and surface vibrations, did not affect the chirping behaviour. In these experiments the insects were placed in separate rooms, and their sounds were transmitted by telephone.

Chirping and mating behaviour

Further experiments carried out by Regen on field crickets (*Liogryllus campestris*) demonstrated the reactions of females to chirping males. In the most elaborate of these experiments, 1,600 sexually receptive females were released around the periphery of a large enclosed area in the middle of which had been placed a cage containing one or more chirping males. Precise data concerning the frequency with which the females moved toward the cage were obtained by surrounding the cage site with an array of traps in which the females were caught as they moved inward. The results were statistically significant. Normal females (those with intact tympanal organs) moved toward the cage and eventually reached it. The removal of one foreleg and its tympanal organ, however, caused difficulty; the movements were more random and the approaches fewer, although some females did succeed in reaching the cage. When both tympanal organs were removed or if the male failed to chirp, the performance of the females was reduced to chance. They also failed to exhibit the seeking performances if the male's stridulatory organ was modified, as by removing the file, so that little or no sound was produced.

In 1926 Regen returned to his study of the alternating chirping pattern of katydids and succeeded in having males react to an artificial sound, one that Regen himself produced. He also found that the alternation could be demonstrated with a suitably active male by using a variety of sounds—whistles, percussion noises, and sounds made with his mouth. It was never altogether clear, however, what changes Regen had made in his signals that finally brought success; probably the secret lay in the particular rhythm and timing of the signals. At any rate, this method made possible a study of the general nature of the auditory sensitivity of these insects and the range of sound frequencies to which they responded. It was shown that katydids are most sensitive to the very high frequencies, those that are beyond the limit of the human ear. The instruments available to Regen at the time, however, did not permit a precise measurement of intensity thresholds. (A threshold is the lowest point at which a particular stimulus will cause a response in an organism.)

Although the work of Regen and others established the basic character of sound reception in insects and its role in communication and mating, other details had to await the introduction of electrophysiological methods in this field as well as the development of electronic methods for the precise production, control, and measurement of sound stimuli.

Electrophysiological observations. When making electrophysiological observations of an auditory mechanism, an electrode (one terminal, generally a fine wire, in an electric circuit) is placed on a nerve or some other sensory structure in the mechanism. Sounds, presented at different frequencies and intensities, produce neural or sensory changes, which are actually electrical discharges or changes in electrical potential of extremely small magnitude. The impulses are picked up by the electrode and transmitted to an instrument with which they can be amplified, observed, and recorded. In both behavioral and electrophysiological observations, the auditory sensitivity of an animal to sounds of different frequencies can be illustrated by a curve.

The electrophysiological method was first used in research on the insect ear in 1933, with observations mainly on two katydid and one cricket species. The tympanal organ of these insects is located on one of the segments of the foreleg; its nerve goes to a ganglion in the thorax. When an electrode is placed on this nerve, its threshold sensitivity and overall frequency range can be determined by varying the intensity and frequency of the sounds applied to the tympanic membrane. It has been found that the tympanal organ of these insects responds poorly to low tones (those of low frequency) but improves rapidly as the frequency increases to a maximum sensitivity around 3,000 to 5,000 hertz. For higher frequencies the sensitivity declines, until a limit is reached at 30,000 hertz. It is likely that the insect's identification of its own species by means of song is primarily in terms of intensity and time patterns, with the rapid changes of intensity playing a prominent part. The possibility of frequency also entering into the pattern, however, cannot be ruled out.

A further question concerns the perception of the direction of a sound source. Clearly, if a female is to seek out and find a chirping male, the effectiveness of her performance depends upon an ability to localize the sound. Experiments indicate that the magnitude of electric responses from the tympanal nerve in katydids varies in a systematic manner when a given sound is presented at different angles while the distance is held constant. The insects continue to exhibit this directional pattern even after one of the tympanal organs has been removed. As was mentioned earlier, Regen found that female crickets deprived of one tympanal organ were still able to locate a chirping male, though less effectively than when both organs were intact.

Evidence of hearing and communication in spiders. Whether spiders have a sense of hearing has long been debated. Early anecdotal observations concerning this matter have now been reinforced with both behavioral and electrophysiological evidence showing without doubt that spiders are sensitive to mechanical vibrations and also to aerial sounds. Whether this sensitivity should be regarded as hearing is considered later in this section, after a review of the anatomical and behavioral evidence.

Lyriform organs

Anatomical evidence. The bodies of spiders contain many slitlike openings, called lyriform organs, that have been considered as sensory in nature. Most of these organs probably have a kinesthetic function and thus provide information on local movements of body parts. There is one type of lyriform organ, however, that differs from the others in its location and in certain structural details. It is found on the metatarsal (next to last) segment of each of the eight legs, close to the joint that this segment makes with the tarsus (the last segment, or foot), and consists of a number of slits—about 10 in the common house spider—that partially encircle the leg. Each slit contains a fluid chamber the inner wall of which is pierced by a tubule through which a thin filament runs to one of the two side walls (lamellae) that enclose the slit. This filament is evidently the termination of a ganglion cell that lies deeper in the leg. It has been suggested that an alternating compression of the lamellae stimulates the terminal filament.

The responsiveness of the common house spider to aerial sounds and mechanical vibrations includes a wide range, from below 20 to as high as 45,000 hertz. Within this range the sensitivity, as measured by electrical potentials, varies widely for aerial sounds; in some experiments narrow regions of frequency have been found in which no responses could be obtained at the highest intensities available. These variations of sensitivity are ascribed to mechanical resonances in the lyriform structure.

The tarsus evidently plays an important part in responses to sounds. Removal of portions of the tarsus reduces the responses about in proportion to the amount removed; immobilization of the tarsus greatly impairs the sensitivity. It appears, therefore, that the tarsus serves as a sensing element that transmits vibrations to the lyriform organ, which thus is a velocity type of ear.

Behavioral evidence. It has been reported that spiders react in characteristic ways to a buzzing insect caught in their web. The spider apparently locates the insect at once, runs to it, and attacks it. An inactive object, however, such as a small pebble enmeshed in the web, produces a different response: the spider manipulates the strands of the web, locates the object, and cuts away the filaments surrounding it so that the object drops to the ground. The reactions of a house spider to a mechanical vibrator applied to a point on the web have been observed. Such a stimulus elicits a response similar to that of an active insect if the vibratory frequency is between 400 and 700 hertz. For frequencies above 1,000 hertz, however, the spider reacts either by running to a secluded corner of the web or, if the intensity is too great, by abandoning the web altogether. From this and similar evidence it has been concluded that the spider has the ability of pitch (tone) discrimination between low and high ranges and perhaps can distinguish between tones of the lower range.

Spiders also react to aerial tones from an artificial source, such as a loudspeaker. These stimuli elicit an orientation response, in which the spider faces the source and reaches out with the two front legs. Thus, in view of the high level of sensitivity to both aerial and mechanical stimuli, the reception of sounds in the spider can probably be regarded as true hearing, and the lyriform organ as a form of ear. It is evidently a velocity type of ear, for there is no tympanic surface to respond to sound pressures, and the small leg segments seem to respond to the oscillatory motions of the air particles.

SOUND RECEPTION IN VERTEBRATES—AUDITORY MECHANISMS OF FISHES AND AMPHIBIANS

The labyrinth

The ear of vertebrates appears to have followed more than one line of evolutionary development, but always from the same basic type of mechanoreceptor, the labyrinth. All vertebrates have two labyrinths that lie deep in the side of the head, adjacent to the brain. They contain a number of sensory endings the primary functions of which are to regulate muscle tonus (a state of partial muscular contraction) and to determine the position and movements of the head and body.

Generalized sketches of vertebrate labyrinths are shown in Figure 27, with the usual locations of the sensory endings indicated for the different vertebrate classes. Two main divisions of these endings are distinguished: a superior division, which includes the three semicircular canals, the organs associated with the sense of balance, and the utricle, a small sac into which the semicircular canals open; and an inferior division, which includes the saccule (also a small sac) and its derivatives. Arising at or near the connection between the utricle and the saccule is the endolymphatic duct, which ends in an endolymphatic sac; this structure probably regulates fluid pressures in the labyrinth and aids in the disposal of waste materials.

The superior division of the labyrinth (Figure 27) is remarkably constant in form throughout the vertebrates except in the cyclostomes (*e.g.*, hagfishes and lampreys), in which the canals and endings are reduced in number. The utricle contains a macular ending, the macula utriculi, and each semicircular canal ends in a crista. In all vertebrate classes except the placental mammals and a few other scattered species, a papilla neglecta is present. It is usually located on the floor of the utricle or near the junction of the utricle and the saccule.

The inferior division of the labyrinth always contains a saccule with its macula, the macula sacculi, but the derivatives of the saccule vary greatly in the different vertebrate classes. In teleosts (bony fishes), amphibians, reptiles, and birds there is a lagena (a curved, flask-shaped structure), with its macula, the macula lagenae. Only the amphibians have a papilla amphibiorum, which is located near the junction of the utricle and the saccule. In some amphibians and in all reptiles, birds, and mammals, there is a papilla basilaris, which is usually called a cochlea in the higher forms, in which it is highly detailed. The elaborate sensory structure of higher types of ears, containing hair cells and supporting elements, is called the organ of Corti.

The macular endings consist of plates of ciliated cells (cells with short, hairlike projections) along with accessory cells, all surmounted by an otolith (a calcareous mass containing numerous particles of calcium carbonate embedded in a gelatinous matrix) or, in teleosts, by one large mass of calcium carbonate. The crista endings contain moundlike groups of sensory cells with supporting cells; the sensory cells have elongated cilia that are embedded in a gelatinous body, the cupula, which forms a sort of valve across an expanded portion of each semicircular canal. The papillae contain plates or ribbons of ciliated cells in a structural framework that lies on a movable membrane, except in amphibians, in which the papillae are on a solid base. These ciliated cells are not surmounted by an otolithic mass or a cupula, but some of the cilia are attached either directly or indirectly to a tectorial membrane (a membrane with one edge fixed to a stationary base, thus anchoring the cilia) or to an inertia body (a mass lying over the ciliated cells and restraining the movements of the cilia).

Functions of labyrinthine endings

The endings have different functions: the macular organs serve primarily as gravity receptors and detectors of sudden movements; the crista organs serve for the perception

From *Biological Review* (1936)

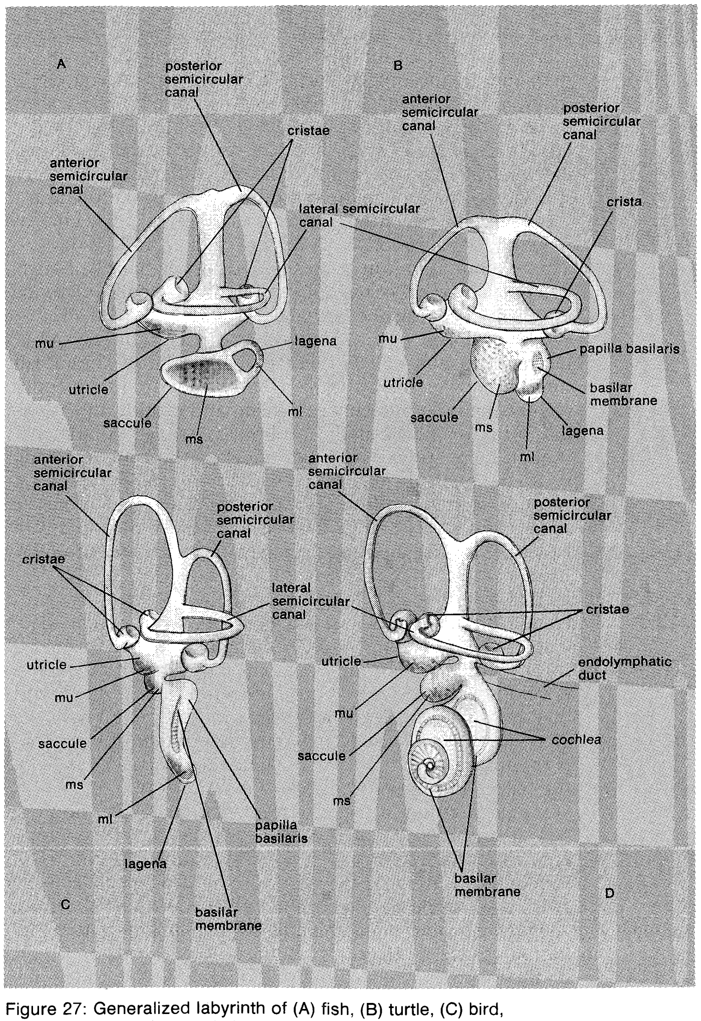

A
posterior semicircular canal
cristae
anterior semicircular canal
lateral semicircular canal
mu
lagena
utricle
ml
saccule
ms
B
anterior semicircular canal
posterior semicircular canal
crista
mu
papilla basilaris
utricle
basilar membrane
saccule
ms
lagena
ml
anterior semicircular canal
posterior semicircular canal
cristae
lateral semicircular canal
utricle
mu
saccule
ms
ml
papilla basilaris
lagena
basilar membrane
C
anterior semicircular canal
posterior semicircular canal
cristae
endolymphatic duct
utricle
mu
cochlea
saccule
ms
basilar membrane
D

Figure 27: Generalized labyrinth of (A) fish, (B) turtle, (C) bird, and (D) mammal. (Abbreviations: mu, macula utriculi; ms, macula sacculi; ml, macula lagenae.)

of rotational acceleration; and the papillae serve for hearing. As structural relations suggest, the auditory endings are derived either from the other labyrinthine receptors or from the primitive labyrinthine epithelium.

Hearing in fishes. The cyclostomes and the elasmobranchs (*e.g.,* sharks and rays) possess a labyrinth with maculae and cristae but have no auditory papillae. There are, nevertheless, two possible ways by which some of these cartilaginous fishes, especially the sharks, react to sounds in the water: by means of the macular organs and by means of the lateral-line apparatus. (For details on these mechanisms, see above *Mechanoreception: reception of external mechanical stimuli.*) It is in the bony fishes (teleosts) that a true ear whose function is hearing first appears among the vertebrates. This ear, which occurs in a number of forms, has varying degrees of effectiveness as a sound receiver; some fishes hear well, others poorly. The differences arise, at least in part, from the accessory mechanisms that aid in the utilization of sound energy.

The basic auditory mechanisms in teleosts. In most fishes, especially in many marine forms, the auditory mechanism is relatively simple, consisting of macular endings that evidently have been diverted from their primitive functions as detectors of gravity and motion. The important change is not in the structure of the end organ but in its innervation—the nerve supply has connections that transmit auditory information. It is thought that in most teleosts the change to an auditory function has occurred in the saccular macula, and probably the lagenar macula as well, and that the utricular macula continues as a receptor for gravity and motion.

Inertia principle in the teleost ear

The simple macular ending of the teleost ear is stimulated by sound through the operation of an inertia principle. Sound waves pass readily through the water and into the body of the fish, causing most of the tissues to vibrate in a uniform manner. The macular otolith, however, represents a discontinuity; because its density is greater than that of the other tissues, it exhibits an inertia effect (resistance to movement). Its motions not only lag behind those of the surrounding tissues but are probably of lesser amplitude as well. Accordingly, a sound creates a relative motion between the otoliths and the other tissues. More specifically, there is relative motion between the bodies of the hair cells, which rest on a tissue base, and the cilia of these cells, the tips of which are in contact with the otolith. This method of stimulating the auditory hair cells is inefficient, however, because of the relatively small difference in density between the body tissues and the otoliths.

Air sacs

Special stimulation mechanisms. In certain groups of teleosts the efficiency of hair-cell stimulation has been increased by a discontinuity that is nearly 1,000 times greater than the one between tissue and otolith; this is the discontinuity between the otolith and a gas bubble. Although there are varying anatomical methods of achieving it, the simplest arrangement, which is found in clupeids, mormyrids, labyrinthine fishes, and a few others, consists of a gas-filled sac that lies against one wall of the labyrinth. In clupeids (*e.g.,* herring), a group in which the utricular macula rather than the saccular or lagenar maculae has an auditory function, long anterior extensions of the swim bladder form air sacs, one adjacent to each utricular macula. In the mormyrids, which include the elephant-nosed fish, a similar condition exists in early life; during adult development, however, the connections with the swim bladder disappear, leaving the air sacs connected with the saccular and lagenar endings. The gas content of these sacs is then maintained by special glands that extract gas from the blood. Air sacs arise in various other ways.

One large group of fishes, referred to as the Ostariophysi (*e.g.,* catfishes, minnows, and carps), has no air sac adjacent to the labyrinth, but a possibly equivalent condition is achieved through a mechanical connection between the swim bladder and fluid chambers adjacent to the labyrinth. A chain of three or four small bones, known as the Weberian ossicles, extends from the anterior wall of a part of the swim bladder to a fluid-filled chamber called the atrium, which in turn connects by fluid passages with the two labyrinths in the region of the saccule-lagena complex (Figure 28). In this arrangement the discontinuity is between the air of the swim bladder and the chain of ossicles in contact with it; the relative motion arising from sound stimulation is communicated through the ossicular (bony) chain and the fluid channels to the macular endings.

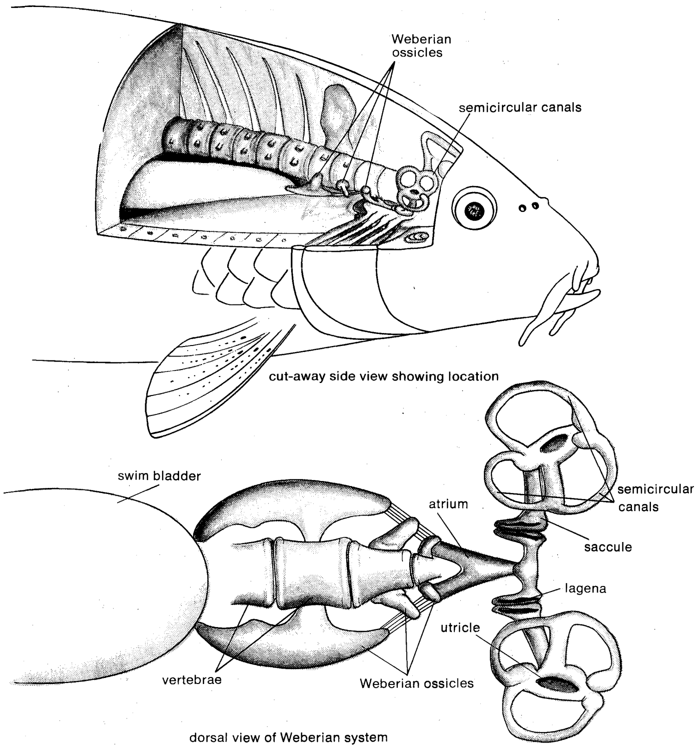


Figure 28: The Weberian ossicles and labyrinths of an ostariophysan fish.
Adapted from *Copeia* (1943) and (bottom) *Nature* (1938)

Regardless of the mechanism employed, however, the ear of all teleost fishes is basically a macular organ. Because it is stimulated by sound that is transmitted to tissues adjacent to the sensory cells and that acts differentially on these cells, this ear is of the velocity type.

Auditory sensitivity of fishes. Although only limited experimental data are available, it appears certain that, in general, fishes with the accessory mechanisms described above have greater sensitivity and a higher frequency range than do those lacking such mechanisms; while upper frequency limits are about 1,000 hertz for many fishes, they are about 3,000 hertz for the Ostariophysi and other specialized types.

Auditory response in goldfish

Many experiments have dealt with the problem of auditory sensitivity in fishes, but the species most extensively tested has been the goldfish, a variety of carp belonging to the Ostariophysi. In one well-controlled investigation, the sound intensities required to inhibit respiratory movements, after conditioning with electric shock, were studied. The greatest sensitivity was found to be around 350 hertz; above 1,000 hertz sensitivity declined rapidly.

In view of the simple anatomical character of the ear, the question of whether fishes can distinguish between tones of different frequencies is of special interest. Two studies dealing with this problem have shown that the frequency change just detectable is about four cycles for a tone of 50 hertz and increases regularly, slowly at first, then more rapidly as the frequency is raised.

Hearing in amphibians. There are three orders of living amphibians: the Apoda, which are legless, wormlike types such as caecilians; the Urodela, which are tailed forms such as mudpuppies, newts, and salamanders; and the Anura, which are tailless forms including frogs and toads. Although members of all three orders have ears, the structures vary greatly in the different groups, and little is known about them except in such advanced types as frogs.

The auditory mechanism in frogs. Although the frog has no external ear (structures on the outside that direct sound vibrations inward), the middle-ear mechanism is well developed. On each side of the head, flush with the surface, a disk of cartilage covered with skin serves as an eardrum. From the inner surface of this disk, a rod of cartilage and bone, called the columella, extends through an air-filled cavity to the inner ear. The columella ends in an expansion, the stapes, which makes contact with the fluids of the inner-ear (otic) capsule through an opening, the oval window. A second opening in the otic capsule, the round window, is covered by a thin, flexible membrane; it is bounded externally by a fluid-filled space that can expand into the air-filled cavity of the middle ear. When the alternating pressures of sound waves cause the eardrum to vibrate, the vibrations are transmitted along the columella and through the oval window to the inner ear, where they are relayed to the round window in a path across the otic capsule by movements of the inner-ear fluids. Along this path are two auditory endings, the amphibian and basilar papillae, the sensory hair cells of which are stimulated by the fluid movements. These movements are transmitted to the ciliary tufts of the sensory cells by a tectorial membrane, which is suspended from the hair cells in such a way that it can be moved by the oscillations of the inner-ear fluids.

Papillae

As sense organs for hearing, the papillae, which appear for the first time in amphibians, have cells like those in lower vertebrates that serve the same purpose. There are two types of papillae: the amphibian papilla, which is found in all amphibians, and the basilar papilla, which is found in some amphibians. Because they are located in different places in the inner ear, the papillae probably represent two distinct evolutionary developments. Moreover, they operate on a mechanical principle found in no other animal group: a tectorial membrane, moving in response to sound vibrations that have been transmitted to it by the inner-ear fluids, stimulates the sensory hair cells directly through connections to the cilia of these cells. In all higher types of ears, on the other hand, the sensory cells themselves are set in motion by the sound vibrations, while the tips of the ciliary tufts are restrained in one of several ways.

Auditory sensitivity of amphibians. Although it is presumed that all amphibians possess hearing of some kind, the evidence is sparse; only salamanders other than anurans have been studied experimentally. Salamanders trained to come for food at the sound of a tone responded only at low frequencies, up to 244 hertz in one specimen and to 218 hertz in three others.

From (bottom) H.W. Rand, *The Chordates*, copyright 1950; used with permission of McGraw-Hill Book Co.

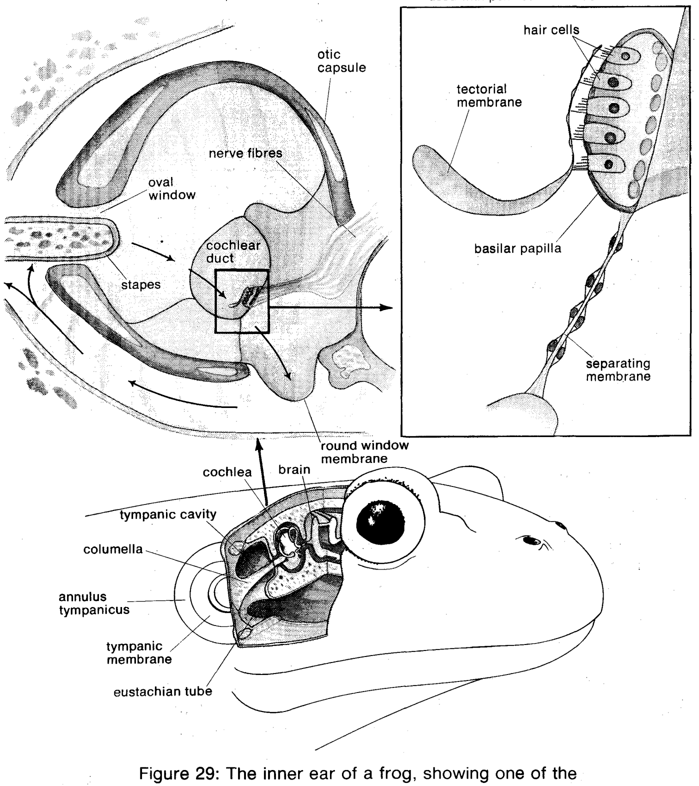


Figure 29: The inner ear of a frog, showing one of the papillae. The arrows indicate the path along which sound vibrations are transmitted by movements of the inner-ear fluids. At the right is an enlarged view of the separating membrane and the basilar papilla.

Frogs, which are of special interest because they first live in the water as tadpoles and then undergo a metamorphosis that equips them for life on land, have been studied more extensively. Considerable modifications of the middle-ear mechanism occur during metamorphosis. Presumably, the tadpole larva has an aquatic ear that is later transformed into an aerial type.

Croaking

Interest in the hearing of adult frogs has been stimulated by their active and often loud croaking during the breeding season. Evidently, their vocalizations assist in the location and selection of mates. The first experimental study of auditory sensitivity in frogs, carried out in 1905, showed that leg movements in response to strong tactual stimuli may be enhanced or even inhibited by sounds.

Somewhat later, following some unsuccessful attempts to train frogs to make behavioral responses to acoustic stimuli, two other methods were employed to determine the sensitivity and range of their hearing. One of these was the recording of changes in the electrical potentials of the inner ear and auditory nerve; the other was the observation of changes in the potentials of the skin (electrodermal responses) to acoustic stimuli. As a result of these investigations, inner-ear potentials and electrodermal responses in the bullfrog have been recorded over a range from 100 to 3,500 hertz. In the treefrog, these same responses have been found in a range that extended from 50 to 3,000 hertz, with the greatest sensitivity from 600 to 800 hertz, and again at 2,000 hertz.

The recording of impulses from single fibres in the auditory nerve of bullfrogs and the green frog indicates that two types of auditory nerve fibres are present. This has led to the suggestion that they represent the different characteristics of the amphibian and basilar papillae. It is believed that the amphibian papilla is more sensitive to low tones and that the basilar papilla is more sensitive to high tones.

AUDITORY STRUCTURES OF REPTILES

The living reptiles belong to four orders: the Squamata (lizards, snakes, and amphisbaenians), the Rhynchocephalia (one rare species, the tuatara of New Zealand), the Chelonia (turtles), and the Crocodylia (crocodiles and alligators). The reptile ear has many different forms, especially within the suborder Sauria (lizards), and variations occur in all elements of its structure—the external ear is often absent or may consist of an auditory meatus (passage) of varying length; the middle ear shows several forms in the different groups; and the inner ear varies in the degree of development of the auditory papilla and also in the ways by which the sensory cells are stimulated by sound.

Lizards. *Auditory structure.* There are about 20 families of lizards, ranging from the chameleon, a divergent type, to the gecko, certain species of which have the most highly developed ears found in the group. The chameleons, of those species studied thus far, have only a few sensory hair cells (40 to 50) in the auditory papilla. The geckos, on the other hand, have several hundred hair cells, and the *Gekko gecko* has about 1,600, the largest known number of hair cells in any saurian. Other lizard species fall between these two extremes in inner-ear development, with the iguanids, the most common lizards in the Western Hemisphere, having from 60 to 200 hair cells, according to the species.

Middle ear of lizards

What may be regarded as the standard type of middle-ear structure in the lizards consists of a tympanic membrane and a two-element ossicular chain that extends from the inner surface of this membrane to the oval window of the otic capsule. As shown in Figure 30, the ossicular chain is made up of two parts: the osseous (bony) columella, whose expanded innermost end (the stapes) fills the oval window, and the extracolumella, a cartilaginous extension that usually spreads out in two to four processes that are embedded in the fibrous layer of the tympanic membrane. Geckos have a single middle-ear muscle attached to the lateral part of the extracolumella; evidently, contractions

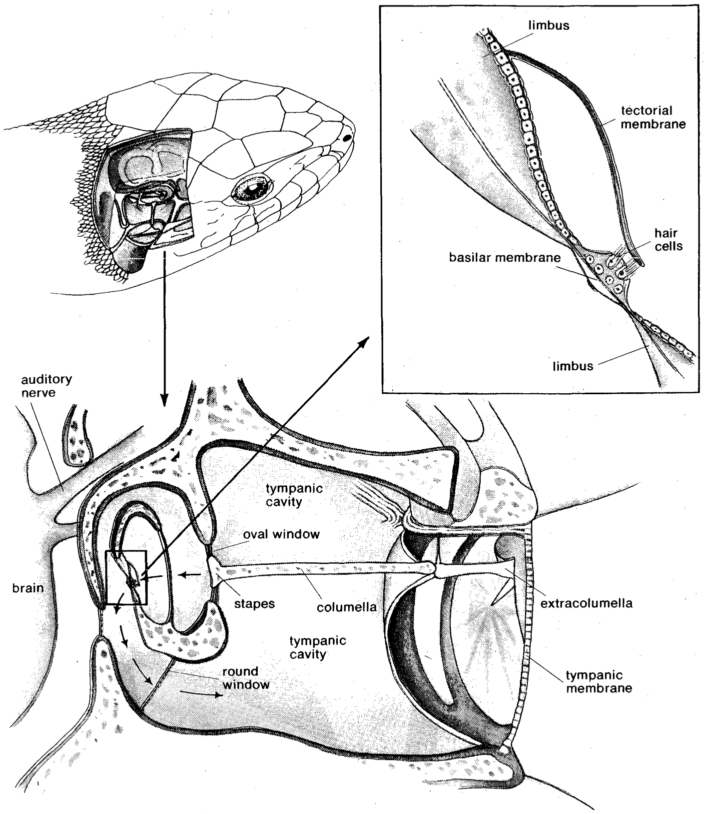


Figure 30: The auditory structures in the right lizard ear as seen from behind and from the left (see text).
From *Journal of Auditory Research* (1965)

of this muscle stiffen the extracolumella, thereby dampening the ossicular motions and protecting the ear against excessively intense sounds.

The auditory part (cochlea) of the inner ear (Figure 30) consists of a basilar membrane lying in an opening in the limbus, which is a plate of connective tissue. The form of the basilar membrane, which is unlike the structure of the same name in amphibians and is clearly of different origin, varies from a simple oval in iguanids to a long, tapered ribbon in gekkonids. In many species the middle portion of the basilar membrane is greatly thickened, especially in some regions of the cochlea. Over this thickening, which is called the fundus, lies the auditory papilla proper—*i.e.,* that part of the cochlea in which the sensory hair cells are held in a framework of supporting tissues and cells. The hair cells usually occur in regular transverse rows, with the number of cells in a row varying along the cochlea. They have a tuft of cilia, the so-called sensory hairs, of graduated lengths, the longest of which are usually attached either directly or indirectly to a tectorial membrane. This membrane arises from a region of the limbus that is usually elevated, often strikingly so, and runs as a thin web or sheet to the region of the hair cells. Only rarely does the free edge of the tectorial membrane connect directly with the cilia of the hair cells; usually there are intermediate connecting structures that take a variety of forms, from simple fibres to relatively massive plates.

Modes of auditory stimulation

The function of the tectorial membrane and its connections to the ciliary tuft of a hair cell is to immobilize the tuft when the body of the hair cell moves in unison with the basilar membrane on which it rests. This produces a relative motion between the ciliary tuft and the body of the cell and stimulates the cell. All auditory stimulation depends ultimately upon this relative motion, and the means just described for achieving it can be regarded as the most fundmental process by which sounds are perceived. Although it is employed in the great majority of ears, it is not the only mode of stimulation. Another mode is that in the ears of fishes, in which an otolith lies upon the ciliary tufts and, by its inertia, reduces and alters the motion of the tuft relative to the cell body. Still another method is the one in the frog papilla, in which the tectorial membrane is moved by the cochlear fluids while the body of the sensory cell remains at rest.

In some lizards the inertia principle has a form different from that found in fishes. In the former, a body called a sallet lies upon the ciliary tufts of a group of hair cells and, by its inertia (or by an equivalent means), restrains the movement of the cilia when the cell body is made to move. The result is a relative motion and a stimulation of the hair cells, like the more common restraint by a tectorial membrane.

The ears of two lizard families show only the inertial restraint method of stimulation; in several other families this method functions in some regions of the cochlea for certain hair cells. Hair-cell stimulation by two or more different arrangements within the same cochlea, however, is the rule rather than the exception because of its many advantages. Although the tectorial-restraint method provides great sensitivity for individual cells, the sallet system also attains good sensitivity, but in another way: by causing many cells—those in common contact with a given sallet—to work in parallel, thus producing a spatial summation. The sallet system has the advantage of being more resistant to damage by overstimulation from intense sounds. In such lizards as the geckos, for example, in which the hair cells are divided nearly equally between tectorial and sallet systems, an exposure to excessive sound has been observed to break all the tectorial connections to the hair cells while leaving the sallet connections intact. But even though the most sensitive hair cells are inoperative, the animal can respond to sounds, although with lesser acuity.

Hearing abilities of lizards. The lizards are the lowest vertebrates to have a well-developed spatial differentiation of the cochlea in which different regions respond to different frequencies of tone. The problem of tonal discrimination has been somewhat solved in frogs, in which the differential responses to tones by the two papillae may provide some information concerning the pitch of sounds. The mechanism in frogs, however, is a poor one, as it can give only crude and uncertain cues at best.

Tonal discrimination

In some lizards, such as iguanids and agamids, a minimum of structural variation occurs along the cochlea; in others (*e.g.,* geckos, which have very extensive differentiation along their extended basilar membranes) the differentiation is almost as great as that in higher vertebrates, including man. Most geckos are nocturnal in habit and use vocalizations to maintain individual territories and probably to find mates.

Although it has been possible to train two species of lizards (*Lacerta agilis* and *Lacerta vivipara*) to make feeding movements in response to a variety of sounds, including tones between 69 and 8,200 hertz, most attempts to train lizards to respond reliably to tonal stimuli have failed. The one useful method thus far developed to study the sensitivity of these animals to sounds involves recording electrical responses in the ear and in the auditory nervous system. Although such observations have provided information about peripheral response to sounds, they do not reveal anything about other processes in the nervous and behavioral systems.

Electrical responses in the cochlea of many lizard ears show considerable variations: in absolute sensitivity, in the tonal regions in which responsiveness is best, and in the extent of the frequency range. It has been concluded that most lizards have good auditory sensitivity over a range from 100 to 4,000 hertz and relatively poor hearing for lower and higher tones. This auditory range is not very different from that of man, although somewhat more restricted than that of most mammals.

Snakes. Without much doubt, snakes developed from some types of early lizards but lost their legs when they adopted habits of burrowing in the ground. Although some snakes burrow, others have taken up different habits: many species live on the surface of the ground, several are largely aquatic, and some live in trees. All, however, show drastic ear modifications that reflect their early history as burrowers; for example, there is no external ear—*i.e.,* no opening at the surface of the head for the entrance of

Auditory modifications of snakes

sound. This fact, together with a seeming indifference to airborne sounds, has led to the supposition that snakes are deaf or that they can perceive only such vibrations as reach them through the ground on which they crawl.

This supposition is incorrect; snakes are sensitive to some airborne sound waves and are able to receive them through a mechanism that serves as a substitute for the tympanic membrane. This mechanism consists of a thin plate of bone (the quadrate bone) that was once a part of the skull but that has become largely detached and is held loosely in place by ligaments. It lies beneath the surface of the face, covered by skin and muscle, and acts as a receiving surface for sound pressures. The columella, attached to the inner surface of the quadrate bone, conducts the received vibrations to its expanded inner end, which lies in the oval window of the cochlea. If the columella is severed, the sensitivity of the ear is significantly reduced.

Although the sensitivity of the snake ear varies with the species, it is appreciably sensitive only to tones in the low-frequency range, usually those in the region of 100 to 700 hertz. For this low range the large mass of the conducting mechanism and the presence of tissues lying over the quadrate bone are not of any great consequence. Moreover, while the sensitivity of most snakes to the middle of the low-tone range is below that of most other types of ears, it is not seriously so. In a few snakes, however, the sensitivity is about as keen as in the majority of lizards with conventional types of ear openings and middle-ear mechanisms.

That the ears of the snake receive some aerial sound waves instead of depending exclusively on vibrations transferred from the ground has been proved by recording the potentials in the cochlea of one ear while rotating the animal in front of a sound-wave source so that the ear being studied was sometimes facing the source and sometimes directed away from it. The recorded potentials were significantly greater when the ear was facing the source. There would have been no difference in the responses if the sound first set up vibrations in the ground and these were then transmitted to the body. This observation also shows that the ears of the snake can determine the direction of a sound in terms of its relative intensity in the two ears. Although snakes can perceive vibrations from the ground that are present at a sufficient intensity, this ability is not peculiar to them; all ears respond to vibrations transmitted to the head.

Amphisbaenians. The amphisbaenians form a little-known group of reptiles. Because they are burrowers and live almost entirely underground, they are seldom seen. The one species in the United States, *Rhineura floridana,* is found in some parts of Florida; a number of species occur in other regions of the world, especially in South America and Africa.

The animals construct a maze of underground tunnels, which they patrol in search of such food as grubs and worms. Although small eyes below the body surface can receive light through a transparent scale, amphisbaenians evidently make little use of vision. There is reason to believe, however, that they use hearing to locate their prey.

Auditory structures in the face of amphisbaenians

Amphisbaenians, like snakes, have no surface indication of an ear; a receptive mechanism below the surface and different from that in snakes conveys vibrations to the inner ear. In the oval window, which occupies the entire lateral surface of the otic capsule, is a stapes. The head of the stapes in most species is directed laterally and forward; it is united by a joint with a rod of cartilage (the extracolumella) that extends forward along the face, in the line of the lower jaw. The extracolumella lies below the surface, where it makes close contact with and finally enters a dense layer of the skin. When the facial region is exposed to sounds, the vibrations are transmitted through the dense layer of the skin to the extracolumellar rod and then through it to the stapes, finally reaching the fluid of the inner ear. That this is the route of sound conduction has been proved by cutting the extracolumella at different places and observing the reduction of recorded responses in the ear.

The auditory mechanism of amphisbaenians varies somewhat according to species but is substantially as described above. The sensitivity, which also varies with species, is surprisingly high in some species, considering the unusual nature of the mechanism involved. Studies similar to those described for snakes have proved that this ear receives aerial sounds and that it can determine the direction from which the sound originated. As expected, this ear also responds to mechanical vibrations communicated directly to the skull.

Turtles. It is sometimes supposed that the turtle's ear is a degenerate organ, largely or even completely unresponsive to sound. Although the turtle's ear is unusual in some respects, and can be regarded as specialized in its manner of receiving and utilizing sounds, it is not a degenerate organ. There is good evidence that turtles are sensitive to low-frequency airborne waves and that some species have excellent acuity in this range.

A plate of cartilage on each side of the head serves as a tympanic membrane. Leading inward from the middle of this plate is a two-element ossicular chain consisting of a peripheral extracolumella and a medial columella the expanded end (the stapes) of which lies in the oval window of the otic capsule. Within the otic capsule are the usual labyrinthine endings, including an auditory papilla. As shown in Figure 31, the auditory papilla lies in a path between the oval window and an opening (the round window) in the posterior wall of the otic capsule. Unlike the round window in most ears, that in turtles has no membranous covering for transmitting pressure changes to the air-filled cavity of the middle ear. Instead, the opening leads to a fluid-filled chamber, the pericapsular recess, that extends laterally and anteriorly to enclose the external

From *Proceedings of the National Academy of Sciences* (1956)

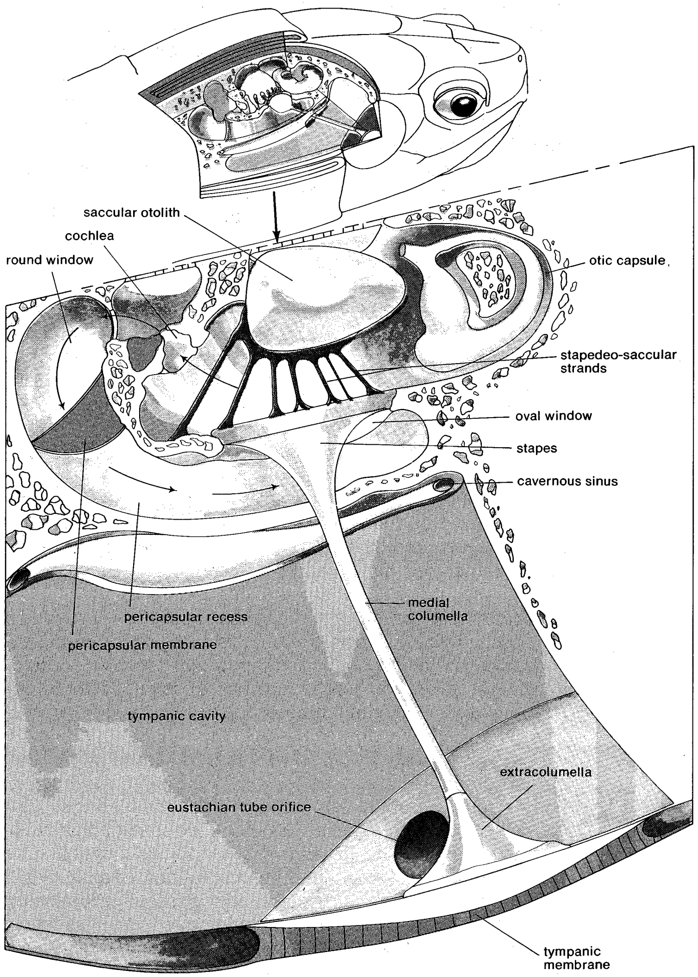


Figure 31: *Structure of the turtle ear.*
(Top) View of the head, showing the location of the otic capsule; (bottom) detail of the otic capsule. Arrows indicate the path along which vibrations are transmitted by movements of the cochlear fluids.

portion of the stapedial expansion of the columella. A pericapsular membrane separates the perilymph (fluid) of the otic capsule from the fluid of the recess. When the stapes is moved inward by the columella at one phase of a sound vibration, the fluid of the otic capsule is displaced, causing a pressure change that, after passing through the sac containing the auditory endings, continues in a circuitous course to the external surface of the stapes. This circuit is indicated by the arrows in Figure 31. When the columella moves outward, the fluid circuit reverses itself. Hence the result of a continuous sound wave is a surging back and forth of the fluids in the otic capsule and the pericapsular recess at the same frequency as that of the sound.

The special mechanical arrangement in the turtle ear is fully effective within the low-frequency range. Indeed, the relatively large mass of tissue and fluid involved in the response to sounds is in part responsible for the efficiency of the ear at low frequencies and also for the rapid loss of sensitivity as frequency increases.

This type of cochlear response to sounds is not peculiar to turtles; it is also found in snakes, through a structural arrangement of similar form. Although it also occurs in amphisbaenids, the fluid path in these animals is entirely different: it proceeds through the perilymphatic recess into the brain cavity and then by an anterior passage across the head to the lateral surface of the stapes.

Certain experiments involving the turtle's sensitivity to sounds have used training methods (conditioned responses); only a few have met with success. It has been found that turtles of the species *Pseudemys scripta,* trained to withdraw their head, respond to sound over the low-frequency range, with the greatest sensitivity in the region of 200 to 640 hertz. This result is in close agreement with electrophysiological observations in which it has been found that impulses could be obtained from the auditory nerve of *Chrysemys picta* for tones between 100 and 1,200 hertz, with highest sensitivity for tones below 500 hertz. Similar results have been obtained by additional observations of this kind with several other species of turtles, some of which are very sensitive to a narrow band of frequencies in the low-tone range. Evidently, the type of receptor mechanism in the turtle can achieve great sensitivity through mechanical resonance at a particular region of the low-frequency scale.

Low-frequency sensitivity in turtles

Evidence has also been obtained that these responses are to aerial waves and not to vibrations set up in the ground. The sensitivity to surface vibrations was considerably poorer than that to aerial sounds. In addition, cutting the columella seriously impaired the responses to aerial sounds but hardly affected responses to mechanical vibrations applied to the turtle's shell.

Crocodiles. The order Crocodylia includes four groups of closely related forms: crocodiles, alligators, caimans, and gavials. The crocodile ear, although clearly reptilian in general structure, has a number of peculiar features. Leading to a tympanic membrane on each side of the head is a shallow external passage the outside opening of which is protected by an earlid that is closed when the animal enters the water and dives. Beyond the tympanic membrane is a middle-ear cavity, with the one on the right connected to the one on the left by an air passage that runs across the head above the brain. A sound presented to one ear, therefore, reaches the other ear about equally well. A columellar system connects the tympanic membrane to the oval window of the otic capsule, as in other reptiles. The inner ear is highly developed and bears many similarities to the cochlea of birds, described in the next section. Elongated and slightly curved, the cochlea contains about 11,000 sensory hair cells, about seven times as many as found in that of the most advanced lizard (*Gekko gecko*).

In comparison to some lizards, the cochlea of *Caiman crocodilus,* which has been most extensively studied, exhibits only a moderate degree of structural differentiation. Yet in this cochlea fibre bundles that extend from the root portion of the tectorial membrane separate into fine fibres that form individual connections with the ciliary tuft of each hair cell. This arrangement is not a common one, though present in certain lizards, such as the chameleons, and also in some degree in birds. It probably provides a high level of specificity in the stimulation process or as much specificity as the overall mechanical pattern permits.

The hearing of crocodilians has not been studied very extensively. It has been noted that the breathing rate in a crocodile accelerates in response to loud sounds, such as the firing of a gun, and it has been observed that specimens of the Mississippi River alligator produce vocalizations of roaring or hissing when low-frequency sounds are made by blowing a horn or by plucking a metal rod. Studies of the electrical potentials in the ear of *Caiman crocodilus* show that it is sensitive to frequencies ranging from 20 to 15,000 hertz.

HEARING IN BIRDS

The avian auditory structure. Ears of birds show considerable uniformity in general structure and are similar in many respects to those of reptiles. The outer ear consists of a short external passage, or meatus, ordinarily hidden under the feathers at the side of the head. Most birds have a muscle in the skin around the meatus that can partially or completely close the opening.

The tympanic membrane of birds

The tympanic membrane bulges outward as in most lizards. In the songbirds, however, it consists of two separate membranes, with the outer one apparently serving to protect the inner one from injury. From the inner surface of the tympanic membrane an ossicular chain transmits vibrations of the cochlea. As in lizards, the chain consists of an osseous inner element, the columella, and a cartilaginous extracolumella that extends the columella peripherally and connects with the tympanic membrane.

The cochlea of birds is similar to that of crocodiles, consisting of a short, slightly curved bony tube within which lies the basilar membrane with its sensory structures. The length of the basilar membrane varies between 2.5 and 4.5 millimetres (0.1 and 0.2 inch) in most birds, but in the owls it may reach 10 millimetres (0.4 inch) or more. At the end of the cochlea is another ending with a different function, the lagena and its macula.

Auditory sensitivity in birds. Using the conditioned-response method to study auditory sensitivity in a small songbird, the bullfinch, responses over a frequency range from 100 to 12,800 hertz have been observed. The electrophysiological method was first applied to the study of hearing in birds in 1936. In this study impulses from the cochlea of pigeons were recorded for tones usually up to 10,000 hertz and occasionally as high as 11,500 hertz. Although this method has been used since 1936, few detailed and quantitative results have been obtained; nevertheless, one striking characteristic revealed by these studies has been the high degree of sensitivity in the low and middle range and the very rapid decrease in the high tones.

Uses of hearing in birds. Like other animals, birds use hearing to warn them of enemies and other kinds of danger. To a degree hardly equalled in lower species, they also use hearing in social relations and communication. Many male birds sing to hold their territories and to attract mates. Some birds also use vocalizations to identify their mates or group members. During the breeding period of the emperor penguin, for example, the male leaves his mate for a journey taking many days in order to obtain food. Upon returning to the general area where his mate has remained with a pack of hundreds of birds, the male is able to locate and to recognize his partner by an interchange of calls.

There is good reason to believe that certain birds, including the swiftlets (*Collocalia*) of Asia and Australia, the oilbirds (*Steatornis*) of tropical America, and possibly a few others, are able to use echolocation when flying in the dark caves that they inhabit. Moreover, it is well established that many owls locate and catch their prey by auditory cues. On a dark night, an owl perched in a tree can hear the rustling sounds made by a mouse in the grass and leaves on the ground below; by accurately localizing this signal, he can make his strike and capture the prey without any visual aid.

HEARING IN MAMMALS

Auditory structure of mammals. In the mammals the ear reaches its highest level of development, with well-differentiated divisions of outer ear, middle ear, and inner ear. Except in some of the sea mammals, in which certain modifications and degenerations have taken place, these structures carry out their functions in a remarkably regular manner.

The pinna

The outer ear consists of pinna (or auricle) located behind the ear opening and partially enclosing it and an auditory meatus that leads inward. The pinna varies greatly in size relative to the size of the animal, being large enough in many species to serve a useful purpose in the collection and reflection of sounds. Many mammals can move the pinna back and forth to regulate in some degree the entrance of sounds to the auditory meatus, which transmits the sounds inward to the tympanic membranes. In some mammals, such as many of the marine types, the external opening can be closed to keep out water when the animal dives, and in certain species of bats the tube itself contains a valve that can be closed to protect the ear against undesirable sounds.

The middle ear of mammals consists of a tympanic membrane, an ossicular chain of three elements, and two tympanic muscles. The tympanic membrane bulges inward, unlike the usually outward-bulging membrane of reptiles and birds. The elements in the ossicular chain are the malleus (hammer), incus (anvil), and stapes (stirrup), so named because of the resemblance of the bones to these objects. The malleus is attached to and partly embedded in the fibrous layer of the inner surface of the tympanic membrane. It connects to the incus, which connects in turn to the stapes, the footplate of which lies in the oval window of the cochlea.

One tympanic muscle extends from an attachment to the skull to an insertion on the malleus. Another muscle has its insertion on the neck of the stapes. By their contractions, both muscles add friction and stiffness to the ossicular chain, thereby reducing its mobility and protecting the inner ear from excessive sounds. The contraction of the muscles is a reflex action and occurs in both ears at the same time in response to loud sounds.

The coiled cochlea

The inner ear is called the cochlea because in man this structure is a complex tube coiled into about 2.5 turns, thus bearing some resemblance to a snail's shell, from which the term is derived. The name cochlea has now been extended to include the auditory portion of the labyrinth in all animals, even when the structure is not coiled, as in reptiles, birds, and egg-laying mammals. In the mammals in which it is coiled, the number of turns in the cochlea varies with species from a little less than two to as many as four. The guinea pig and its relatives have the largest number of cochlear turns. Extending along the inside of this coiled passage is the basilar membrane, bearing on its surface the sensory structure known as the organ of Corti, which contains the hair cells.

In mammals a uniform system is employed in the stimulation of the hair cells by sounds. A relatively thick tectorial membrane, anchored securely on one edge to the supporting structure (the limbus), lies with its free portion over the hair cells and with the cilia of these cells firmly attached to the lower surface of this portion. When vibratory movements of the basilar membrane cause the bodies of the hair cells to move, the tips of the cilia are restrained by their attachments to the tectorial membrane. Hence the relative motion between the bodies and cilia of the hair cells stimulates them.

The sizes, shapes, and spatial relations of many otic structures vary in the different mammalian species, but it is thought that the same basic principles of operation are involved. This uniformity contrasts with their situation in reptiles, in which different systems are present both in different species and sometimes within one ear.

Auditory sensitivity among mammals

A number of features are of particular significance in determining the sensitivity and frequency range, which vary with species. Because large masses involve great resistances when moved at high frequencies, the size and mass of the moving parts determine to some degree the variations of sensitivity with frequency and the frequency limits within which the ear operates. The ossicular chain is a mechanical lever, and its lever ratio and the difference in area between the tympanic membrane and the stapedial footplate determine the efficiency of sound transmission from air to the cochlear fluid. The mechanical characteristics of the cochlea and the degree of variation of these characteristics along its extent determine the frequency range of hearing and the degree to which different tones can produce different response patterns. Finally, the numbers and distribution of hair cells along the basilar membrane and the density and specificity of innervation of these cells determine the delicacy and precision with which their periodic activity and spatial patterns are registered by the central areas of the auditory nervous system.

These anatomical features have been studied in detail in a few animals: among the mammals, mainly in cats, guinea pigs, and to a lesser degree in man. The functional aspects, as shown in responses to sounds and to discriminations among different sounds, have been considered principally in man and to a much more limited extent in other mammals. (The characteristics of human hearing are treated at length below; see *Human hearing.*) Some of the auditory characteristics of mammals below man are described in the sections that follow.

Hearing in subhuman mammals. *Primates.* The hearing of other species in the division of mammals to which man belongs has always been of special interest. A number of species have been studied, including monkeys, marmosets, and chimpanzees among the primates considered as the most advanced, the anthropoids; and tree shrews, lemurs, and lorises among the more primitive.

By using a variety of training methods with chimpanzees, monkeys, and marmosets, behavioral thresholds have been recorded in response to sounds of different intensities and frequencies. When compared with each other and with man, it has been found that the hearing sensitivity of these animals and man is remarkably similar over a range of frequencies from 100 to 5,000 hertz, after which the sensitivity begins to differ. The differences observed at the higher frequencies, however, may be partly attributed to variations in experimental procedures. Thus, the results for the chimpanzee stop at 8,192 hertz because this was the highest tone used in the tests. Other observations have shown that chimpanzees can hear tones up to about 33,000 hertz and that young human subjects often hear tones as high as 24,000 hertz. It is also evident that monkeys and marmosets of the species studied can hear still higher tones.

Common laboratory animals. Certain mammals have long been favourite subjects for various kinds of biological studies in the laboratory, largely because of their convenient size, hardiness under caged conditions, and gentle temperament. Familiar among these are cats, dogs, guinea pigs, rats, mice, rabbits, and, more recently, hamsters, chinchillas, and gerbils. Auditory sensitivity functions have been obtained in these animals by a variety of behavioral and electrophysiological methods.

When measured behaviorally by conditioned responses and then plotted on a curve, the auditory threshold sensitivity of cats, guinea pigs, and chinchillas is much the same—a progressive improvement in sensitivity as the frequency is raised until the middle tones (about 500 to 5,000 hertz) are reached, at which point sensitivity tends to remain the same, and then shows a rapid loss in the upper frequencies. There are differences, however, in the maximum sensitivity attained in the middle region, with the guinea pig the least sensitive and the cat the most sensitive of the three species.

Differences between behavioral and electrophysiological observations

Sensory responses in the cochlea of mammals have been measured electrophysiologically by placing an electrode on the round window membrane. Unlike behavioral curves, however, the curves obtained by plotting the sound required to produce an arbitrary amount of electrical potential of the cochlea do not represent auditory thresholds. Instead, their usefulness is largely in their shapes, which indicate in a relative way the regions of good and poor sensitivity. In addition, these curves represent the performance of the peripheral portion of the auditory mechanism up to the point at which the sound stimulus activates the sensory hair cells in which the potentials are generated. Hence, unlike the curves obtained by behavioral responses, those obtained by cochlear potential methods do not indicate the performance of the central auditory nervous system (the nerve connections between the ear and brain and those parts of the brain in which neural impulses from the ear are processed to produce behavioral responses).

In the simpler animals, the two types of curves are much alike, judging from the very limited evidence available. In mammals, however, the behavioral curves differ from the cochlear potential curves in three ways. In the behavioral curves there is (1) an exaggerated gain in sensitivity to tones of low frequency, (2) a greater sensitivity to the medium-high tones, and (3) a more rapid loss of sensitivity to the extreme-high tones and a lower frequency of the upper limit. These differences are believed to arise mainly through the elaborate neural processing that takes place in the more highly developed mammalian nervous system, a processing that improves the sensitivitity to high-frequency tones but reaches a limit of effectiveness and finally fails above some frequency limit. With these conditions in mind, the electrophysiological curves can be used to predict reasonably well an animal's behavioral responses to sound waves.

Large mammals. Because most of the mammals in which hearing has been studied by laboratory methods are small, much less is known about the auditory capabilities of large ones, even of such domesticated animals as horses and cows; nevertheless, it is usually assumed that the auditory capabilities of these animals are much like those of humans. At least they hear sounds in man's vocal range because they seem to respond to verbal signals. Elephants, for example, trained as working animals, are said to obey as many as 30 different commands. A number of wild animals of medium and large size—raccoons, opossums, and several members of the cat and dog families—have been studied electrophysiologically by the cochlear-response method. Their sensitivity curves are fairly similar in form and in the upper limits attained.

Marine mammals. Of special interest are the sea mammals, which have been derived from early land species and which have undergone certain changes in order to adapt themselves to at least a partially aquatic existence. In the course of adapting to marine conditions, however, some sea mammals, such as seals and sea lions, seem to have made only limited alterations in their ear structures. In addition to being able to close the meatus when diving, their pinnas have been greatly reduced or essentially lost, a feature of streamlining for rapid progress through the water.

Adaptations for aquatic hearing

There are three possible ways that the hearing of marine mammals might be adapted to an aquatic environment: (1) unchanged aerial hearing, with no aquatic adaptation; (2) conversion to an aquatic type of hearing with loss of good hearing for aerial sounds; and (3) development of some kind of double system, with at least serviceable reception of both aerial and aquatic vibrations. In a study of hearing in the common seal, in which responses to aerial and aquatic stimuli were compared, it was found that this animal has a greater sensitivity to aquatic sounds, especially in the upper frequencies, which extended to the remarkably high frequency of 160,000 hertz. Yet, although the seal has made an adjustment for hearing in water, it has not sacrificed the quality of its aerial hearing, which remains at an excellent level, especially for one frequency around 2,000 hertz and another around 12,000 hertz. These differences in auditory senstivity suggest that the mechanisms in this animal for aerial and aquatic hearing are somehow different, but no complete explanation of the adaptations has yet been found.

Whales, on the other hand, have converted their ears to a truly aquatic form, apparently with some sacrifice of aerial reception. The study of their ears and hearing has been carried out in only a few species of the toothed whales, which produce sounds and use their ears in the process of echolocation (see next section).

The ear of whales has undergone extensive changes. The pinna is absent and the external ear opening has been reduced to such a minute size, almost a pinhole in some species, that it no longer serves as a path for the entrance of sound. The eardrum, although present in a modified form, seems to serve no useful purpose; it is connected to the malleus only by a ligament, and this connection can be cut without an ensuing loss of sound reception. The usual three ossicles of the middle ear are present, with the footplate of the stapes resting in the oval window. These ossicles are much more massive than the ordinary mammalian ossicles.

It appears that the whale ear has been converted to a true aquatic type, functioning according to principles similar to those found in the ears of fishes, as described earlier. Sound vibrations in the water readily pass through the tissues of the head and reach the deep-lying middle- and inner-ear structures. Probably the ossicles represent an inertial mass in somewhat the same way that the otolithic body does in fishes. Because of their inertia, the ossicles tend to move with smaller amplitudes and in different phase relations when the tissues of the head, including parts of the cochlea, are set in vibration. This difference in relative motion produces an alternating displacement of the cochlear fluid, which is in contact with the footplate of the stapes and which can be set in motion because of the presence of a pocket of gas in the region of the round window. The performance of the whale ear has been measured in an exact manner throughout the frequency range in one species, the bottle-nosed dolphin (*Tursiops truncatus*). By a conditioned-response method, it has been found that this animal possesses excellent auditory sensitivity that extends well into the high frequencies.

Echolocation in bats. Bats are divided into the large bats and the small bats. With one or two exceptions, the large bats live on fruits and find their way visually. The small bats feed mostly on insects, catching them on the wing by a process known as echolocation. As was mentioned earlier, echolocation is a process in which an animal produces sounds and listens for the echoes reflected from surfaces and objects in the environment. From the information contained in these echoes, the animal is able to perceive the objects and their spatial relations.

How echolocation works

Bats produce sounds with the larynx, an organ in the throat that has undergone certain adaptations that make it unusually effective in producing intense, high-frequency sounds. The character of the sounds varies with the species and also with the particular activity. On striking a small object such as a flying insect, the emitted sounds are reflected with only a small fraction of their original energy; the sound is further weakened before reaching the ears of the bat when it must travel some distance through the air.

Although the frequency of bat cries varies with species, their cries usually occur in a range between 80,000 and 30,000 hertz. In most species, such as *Myotis lucifugus* and *Eptesicus fuscus,* the cry is a frequency-modulated pulse of sound; it begins at a high frequency, say, of 70,000 hertz, and in about 0.2 second declines in frequency to about 33,000 hertz. The starting frequency may vary, even in successive cries; a second pulse might begin at 60,000 and end at 30,000 hertz. The greatest energy in the cry is usually in the middle of this frequency range, perhaps around 50,000 hertz in the species mentioned above.

The use of such high frequencies is an essential feature of the bat's sonar system. In order to determine the nature of objects by reflected sounds, it is necessary that the wavelengths of the sound be small in relation to the dimensions of the objects—indeed, as small as possible if fine details are to be represented.

An important problem of echolocation is how the bat is able to detect reflected sounds, often in the presence of disturbing noises, and to obtain the information necessary for tracking and catching an insect as well as discriminating between this object and others in the environment. This problem involves considering first the structure of the auditory mechanism in bats and then the nature of their hearing.

The external ear of bats is usually well developed. In most species the pinna is large relative to the size of the head, and in those species called the whispering bats, because they make such faint sounds, this structure is huge. With its large surface, the pinna acts as an efficient collector and resonator of high-frequency sounds. It is also freely movable and can be rotated and inclined in various ways. The meatus leads inward to the eardrum and, as already mentioned, contains a valve that can be closed to reduce the entrance of sounds. The middle ear of bats is of the usual mammalian pattern—a three-part ossicular chain—but its structure is impressive in the extraordinary delicacy of the moving parts. The two tympanic muscles, however, are relatively large.

The cochlea of bats also shows the general mammalian form, but there are variations that may be significant for the special functions that are performed by this ear. The basilar membrane is not particularly well developed; it is short in comparison with that of most mammals, and its structural variation from basal to apical ends is only moderate in extent. Whereas most basilar membranes are rather strongly tapered in width, being narrow at the basal end of the cochlea and several times broader near the apical end, in the bat there is only a slight taper, between twofold and threefold. Another curious feature in the cochlea of bats is the presence of local thickenings of the basilar membrane that may add to the stiffness of the cochlear structure.

The auditory portion of the nervous system has undergone extraordinary development in bats. The regions concerned with hearing are relatively enormous, which is in accord with the great predominance of hearing over the other senses in this animal.

Auditory sensitivity of bats

The hearing of bats has been studied by both electrophysiological and behavioral methods. In the species *Myotis lucifugus,* electrophysiological measurements of cochlear potentials indicate that response is poor in the low frequencies but improves fairly steadily until the range of 2,000 to 5,000 hertz is reached, at which it tends to level off. Beyond 15,000 hertz there are many irregularities but, in general, the sensitivity declines at a rapid rate. The results of similar studies on a specimen of *Eptesicus fuscus* are much the same as those for *Myotis,* though the observations were not extended into the lowest frequencies. The most sensitive range for this species is around 4,000 to 15,000 hertz, after which there is a fairly rapid decline in the upper frequencies.

The behavioral threshold curve for *Eptesicus* has a markedly different form. There is a rapid improvement in sensitivity from 2,500 to 10,000 hertz, but the greatest sensitivity is in two peak areas, from 10,000 to 30,000 hertz and from 50,000 to 70,000 hertz, with a separation by a moderate reduction around 40,000 hertz.

There are other peculiarities of the behavioral sensitivity of *Eptesicus* to sound stimuli that are of particular interest. The rapid loss of sensitivity to tones around 40,000 hertz may be caused by a failure of neural processing for these tones. The slope of the low-frequency end of the curve is unexpectedly steep, and this appears in a region where the cochlear response is passing through its maximum. Nothing like this has been observed in other animals; it seems to be a peculiarity of the bat.

When the cochlear responses of bats are compared with similar responses in other small mammals—as, for example, the rat—there is a general similarity in the results. The rat, however, has better sensitivity as measured by this method, reaching a level of especially good acuity in the range from 20,000 to 60,000 hertz, the range in which the bat sensitivity falls off rapidly. As mentioned previously, it must be kept in mind that the sensitivity indicated by the cochlear potentials is mediated in the peripheral mechanism, before involvement of the central auditory nervous system. When the behavioral response is considered, however, the contribution made by the bat's central auditory nervous system can be appreciated: the region of greatest auditory sensitivity, extending from 10,000 to 70,000 hertz, is the same region as the frequency of the echolocation cries and the one in which bats have the greatest need of acute hearing.

The failure of these bats to exhibit a behavioral response to tones below a frequency of 10,000 hertz can perhaps be explained also in relation to their peculiar use of hearing. This is a region of frequency that has little or no value for echolocation. More than that, it often contains noises of various kinds that, if heard, might be detrimental to this essential function. It has often been observed that bats are not easily disturbed by extraneous sounds of low frequencies, even extraordinarily intense ones. This peculiarity of hearing in bats may account for their resistance to masking sounds. The slight degree of structural differentiation found in the cochlea of bats may represent another aspect of the limitation of their hearing to that part of the sound spectrum most useful in echolocation. Therefore, it appears that the ear of the bat, which is a rather ordinary type of mammalian structure so far as level of auditory sensitivity and degree of tonal differentiation are concerned, has been developed for a particular purpose—namely, the reception of high-frequency sounds within a limited range.

Echolocation in other mammals. Among the mammals possessing echolocation are the toothed whales. These animals probably produce sounds in the water in two ways: with the larynx and with the complex system of passages connected to the blowhole, which is a nostril in the top of the head. Although many different types of sounds are possible, during echolocation they consist mainly of a rapid series of clicks. These clicks contain many components, but the principal energy is in the high frequencies, from perhaps 50,000 to as much as 200,000 hertz. The use of such high frequencies by these animals is a requirement for effective echolocation in water. Because the velocity of sound is greater in water than in air, the wavelengths are longer; therefore, in order for echolocation to attain the same effectiveness of object discrimination as that achieved by the bat with aerial sounds, an aquatic animal has to use frequencies at least five times as high.

High-frequency clicks of toothed whales

Whales have good vision when submerged, and apparently their eyes remain fairly serviceable when their heads are out of water. Dolphins can be trained to strike targets or leap over obstacles held several feet above the surface of the water. For many tasks, however, they use echolocation very effectively, such as when catching fish at night or when visibility is poor in murky water. Dolphins have been trained to make fine discriminations of objects when their vision has been completely excluded by blindfolding. Echolocation of some form and degree of effectiveness is suspected in still other animals, such as shrews and sea lions, but the evidence is meagre thus far. (E.G.W.)

HUMAN SENSORY RECEPTION

Ancient philosophers called the human senses "the windows of the soul," and Aristotle enumerated at least five senses—sight, hearing, smell, taste, and touch—and his influence has been so enduring that many people still speak of the five senses as if there were no others. Yet, the human skin alone is now regarded as participating in (mediating) a number of different modalities or senses (*e.g.,* hot, cold, pressure, and pain). The modern sensory catalogue also includes a kinesthetic sense (sense organs in muscles, tendons, and joints) and a sense of balance or equilibrium (so-called vestibular organs of the inner ear stimulated by gravity and acceleration). In addition, there are receptors within the circulatory system that are sensitive to carbon dioxide gas in the blood or to changes in blood pressure; and there are receptors in the digestive tract that appear to mediate such experiences as hunger and thirst.

Not all receptors give rise to direct sensory awareness; circulatory (cardiovascular) receptors function largely in reflexes that adjust blood pressure or heart rate without the person being conscious of them. Though perceptible as hunger pangs, feelings of hunger are not exclusively mediated by the gastric (stomach) receptors. Some brain cells may also participate as "hunger" receptors. This is especially true of cells in the lower parts of the brain (such as the hypothalamus) where some cells have been found to be sensitive to changes in blood chemistry (water and other products of digestion) and even to changes in temperature within the brain itself.

General considerations of sensation

BASIC FEATURES OF SENSORY STRUCTURES

Classification by stimuli

One way to classify sensory structures is by the stimuli to which they normally respond; thus, there are photoreceptors (for light), mechanoreceptors (for distortion or bending), thermoreceptors (for heat), chemoreceptors (*e.g.,* for chemical odours), and nociceptors (for painful stimuli). This classification is useful because it makes clear that various sense organs can share common features in the way they convert (transduce) stimulus energy into nerve impulses. Thus, auditory cells and vestibular (balance) receptors in the ear and some receptors in the skin all respond similarly to mechanical displacement (distortion). Because many of the same principles apply to other animals, their receptors can be studied as models of the human senses. In addition, many animals are endowed with specialized receptors that permit them to detect stimuli that man cannot sense. A snake (the pit viper) boasts a receptor of exquisite sensitivity to "invisible" infrared light; some insects have receptors for ultraviolet light and for pheromones (chemical sex attractants and aphrodisiacs unique to their own species) thereby also exceeding human sensory capabilities.

Regardless of their specific anatomical form, all sense organs share basic features. (1) They contain receptor cells which are specifically sensitive to one class of stimulus energies, usually within a restricted range of intensity. Such selectivity means that each receptor can be said to have its own "adequate" or proper or normal stimulus, as, for example, light is the adequate stimulus for visual experience. Nevertheless, other energies ("inadequate" stimuli) can also activate the receptor if they are sufficiently intense. Thus, one may "see" pressure when, for example, the thumb is placed on a closed eye and one experiences a bright spot (phosphene) seen in the visual field at a position opposite the touched place. (2) The sensitive mechanism for each modality is often localized in the body at a receiving membrane or surface (such as the retina of the eye) where transducer neurons (sense cells) are to be found. Often the sensory organ incorporates accessory structures to guide the stimulating energy to the receptor cells; thus, the normally transparent cornea and lens within the eye focus light on the retinal sensory neurons. In some cases, blindness can be cured by surgically removing a lens that has grown opaque from cataract in order to permit light once again to reach the retina. Additional postoperative optical correction in the form of a contact lens or eyeglasses is necessary to compensate for the missing lens. Retinal nerve cells themselves are more or less shielded from nonvisual sources of energy by the surrounding structure of the eye; but mild electrical currents delivered to most sense organs, including the eye, can produce sensory experiences appropriate to the specific organ.

The generalized electrical nature of neural function largely accounts for the effectiveness of such currents in evoking a full range of different sensations. (3) The primary transducers or sensory cells in any receptor structure normally connect (synapse) with secondary, ingoing (afferent) nerve cells that carry the nerve impulse along. In some receptors, such as the skin, the individual primary cells possess threadlike structures (axons) that may be yards long, winding from just beneath the skin surface through subcutaneous tissues until they reach the spinal cord. Here each axon from the skin terminates and synapses with the next (second-order) neuron in the chain. By contrast, each primary receptor cell in the eye has a very short axon that is contained entirely in the retina, making synaptic contact with a network of several types of second-order (internuncial) cells, which, in turn, make synaptic contact with third-order neurons called bipolar cells, all still in the retina. The bipolar-cell axons extend afferently beyond the retina, leaving the eyeball to form the optic nerve, which enters the brain to make further synaptic connections. If this visual system is considered as a whole, the retina may be said to be an extended part of the brain on which light can directly fall. (4) From such afferent nerves, still higher order neurons make increasingly complex connections with anatomically separate pathways of the brainstem and deeper parts of the brain (*e.g.,* the thalamus) that eventually end in specific receiving areas in the cerebral cortex (the convoluted outer shell of the brain). Different sensory receiving areas are localized in particular regions of the cortex; *e.g.,* occipital lobes in back for vision, temporal lobes on the sides for hearing, and parietal lobes toward the top of the brain for tactual function.

Sensory parts of the brain

APPROACHES TO THE STUDY OF SENSING

The science of the human senses is truly interdisciplinary. Philosophers, physicians, anatomists, physical scientists, physiologists, psychologists, and others have all joined in studying sensory activities. Some of their earliest work was anatomical, an approach that continues to be fruitful. Physical scientists, particularly physicists and chemists, made especially important contributions to an understanding of the nature of stimulus energies (*e.g.,* acoustic, photic, thermal, mechanical, chemical); in the process, they also carried out many fundamental measurements of human sensory function. Hermann L.F. von Helmholtz, a 19th century German scientist who was a physicist, physiologist, and psychologist, studied the way in which sound waves and light are received (sensed or detected) and also how they are interpreted (perceived) by people.

Modern studies of sensation have been enhanced by contemporary devices permitting the precise production and control of sensory stimuli. With other kinds of instruments, physiologists have been able to probe the electrical signals generated by sensory cells and afferent nerve fibres to provide a biophysical analysis of sensory mechanisms. Psychophysics embraces the study of the inner (private, subjective) aspects of sensation in terms of outer (public, objective) stimulus energies. One of the oldest and most classical approaches to the study of sensation, psychophysics includes the study of people's reports of their sensations when they are stimulated: of their ability, for example, to match tones of equal loudness, to detect stimulus differences, and to estimate sensory magnitude or intensity under conditions of controlled stimulation. Psychophysical research continues as an active enterprise particularly among modern psychologists.

An old philosophical notion that "mind" is but a clean slate or tablet (*tabula rasa*) until written on by impressions from the senses no longer seems fully tenable; human infants, for example, show inborn (innate) ways of sensing or perceiving at birth. In its modern form, the problem of learned versus innate factors in sensory experience is studied in terms of the extent to which the genetically determined structure and function of sense organs and brain depend upon stimulation and experience for their proper maturation. Poverty of stimulation (sensory deprivation) in an infant's early life increasingly is being documented as detrimental to the full flowering of mature perceptual and intellectual functions. Since this sort of evidence may indeed lend some support to the notion of the *tabula rasa,* modern investigators give credence both to nativistic (based on heredity) and empiricistic (based on learning) interpretations of human sensory function (see also LEARNING AND COGNITION).

Emotional aspects of sensation

A distinction between the discriminatory (epicritic) and emotional (protopathic) features of sensations was made by Sir Henry Head (1861–1940), a British neurologist, who noted, for example, that after a sensory nerve from the skin had been cut, the first sensations to recover as the nerve healed appeared to be diffuse, poorly localized, and extremely unpleasant. He theorized that this initial lack of sharp discrimination associated with unpleasant experience reflected the properties of a primitive protopathic (emotional) neural system which regenerated first. He held that this system subserves pain and the extremes of temperature and pressure sensation usually associated with an affective (emotional) tone. Because recovery of fine tactual discrimination, sensitivity to lightly graded stimuli, and the ability to localize points touched on the skin returned later, Head posited the existence of another discriminatory system. While later research has not confirmed his theory, the sequence of changes in the recovery following nerve injury is most typical.

Chemical-visceral sensations particularly have hedonic (pleasure–pain) properties. Most people tend to refer to odours and tastes as pleasant or unpleasant; thus, the chemical senses are closely tied to motivation, to preferences, or to aversions. Although reflex licking or sucking is stimulated by tactual stimulation of lips and mouth, newborn infants tend to suck longer and harder when the stimulus has clear hedonic value; *e.g.,* avidly turning their lips toward a nipple to get a sweet taste. (See also EMOTION AND MOTIVATION.) Apparently, one's sweet tooth is largely nativistic, in that it requires little prior learning. The craving for salt (especially heightened under conditions of salt deprivation) likewise appears to be widespread throughout the animal kingdom without prior learning. The role of taste and smell as innate factors in behaviour may not be quite so influential in man as in other animals. Man's food habits and preferences are strongly overladen with custom and tradition; that is, they are learned in large measure.

In the modern era, the language of communication engineering has been found to be useful in describing human senses. Thus, each sensory modality may be described as a channel that receives stimulus information (input), that processes and stores the information, and that retrieves it as needed for the effective behaviour (output) of the individual. In addition, modern engineering has provided devices (*e.g.,* radio, television, radar, the electron microscope) that serve to extend the range and power of man's senses; in the last analysis, however, all such devices convert (transduce) their information back to a form of stimulus energy that is directly perceptible to the unaided senses. Thus a television set is a transducer that converts imperceptible electromagnetic waves into visual and auditory signals. For some special purposes, people may employ alternative sensory channels, as when blind people use Braille or other tactual input as substitutes for the missing visual channels. While the chemical senses have little function in symbolic communication among people, the use of perfumes in romantic signalling is a notable exception. In general, however, the human chemical senses are more directly involved in physiological survival; *e.g.,* warning that a putrid fish is dangerous to eat. One's physical well-being also rests heavily on his proprioceptors (for sensing his own bodily position) and on his sense of balance. These structures, monitoring (feeding back information on) his bodily orientation in space, provide crucial sensory feedback for guiding one's movements (see also PERCEPTION).

Survey of some of the human senses

CUTANEOUS (SKIN) SENSES

It was observed above that studies of cutaneous (skin) sensitivity yield evidence that man's senses number more than five. There is evidence for two pressure senses (for light and for deep stimulation), for two kinds of temperature sensitivity (warm and cold), and for a pain sense. In the 1880s, experimental findings that the human skin is punctate (selectively sensitive at different points) gave clear indication of a dissociation among functions once lumped together as the sense of touch. Mapping the skin with a fine bristle or with a narrow-tipped (warm or cold) cylinder showed that there were different spots of maximum sensitivity to pressure, warm, and cold. When stimulated between the spots on the skin, people reported no such sensations. Pain spots also can be located with a finely pointed needle, but the punctate character is less striking since pain seems to be widespread when stimulus intensity is increased. The number of spots is greatest for pain, next for touch, then for cold, and least for warm. Efforts to identify specific receptor cells for each of these sensitive points have been the subject of much debate and still pose a problem that is not completely settled.

Mapping sensitive spots of skin

Nerve function. Microscopic examination of the skin reveals a variety of nerve terminals; there are free nerve endings (which are most common), so-called Ruffini endings, and encapsulated endings, such as the pacinian corpuscle, Meissner's corpuscle, or Krause end bulb, all named for investigators who discovered them (Figure 32). At one time it was thought that each of these specialized structures mediated one of the cutaneous modalities, but efforts to extirpate (surgically remove) the nerve endings under the spots have yielded only questionable data. Further, the cornea of the eye shows only free nerve endings; yet pain, pressure, and some temperature sensations can be elicited by stimulating the corneal surface.

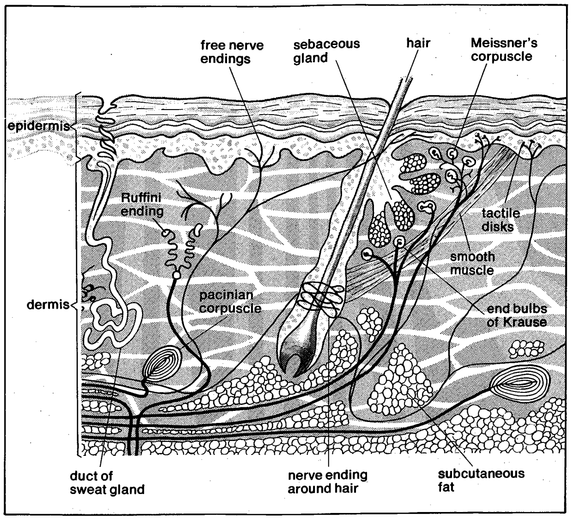


Figure 32: Microscopic view of human skin.
From E. Gardner, *Fundamentals of Neurology,* 5th ed. (1968); W.B. Saunders Company, Philadelphia

Electrical recordings from the cutaneous nerves of laboratory animals suggest a much wider variety of receptors than are encompassed by the reports people give of their sensations. Some nerve endings seem to respond only to one type of stimulus (*e.g.,* to pressure stimuli of very light weight or to slight temperature changes); others exhibit a broad range of sensitivity. There are some receptors that show combined sensitivity to both temperature and pressure. In some cases only special types of mechanical stimulation (such as rubbing) may be effective. Furthermore, there is an extensive overlap in the areas of skin (receptor fields) for the individual nerve fibres examined, suggesting that there is a neural integration of overlapping afferent inputs of skin nerves. A model of the sensory system that envisages only a single nerve fibre serving one tactual spot is clearly contradicted.

On the other hand, some tactual receptors (*e.g.,* Pacinian corpuscles) respond only to mechanical deformation. This corpuscle is an onion-shaped structure of non-neural (connective) tissue built up around the nerve ending; indeed, the distinctive corpuscle, if anything, reduces the mechanical sensitivity of the nerve terminal itself. If the onion-like capsule is entirely removed, mechanical sensitivity not only remains but is somewhat greater than when the capsule is present.

Variations in sensory nerve fibres

In addition to the differences in the sensory end structures of the skin, the afferent nerve fibres (axons) from them also show diversity. The nerve fibres range in size from large myelinated (sheathed) axons of 10 to 15 microns diameter down to extremely small unmyelinated fibres measuring only tenths of microns across. Fatter axons tend to conduct nerve impulses more rapidly than do small fibres; when axons of different diameters form a single bundle (a nerve), they constitute a so-called mixed nerve. Thus, electrical records from a mixed nerve show what are labelled A (fast), B (medium), and C (slow) components that reflect the typical speeds at which axons of different diameters conduct. Although such specialized capsules as Pacinian corpuscles tend to be associated with larger diameter axons, and temperature-sensitive endings tend to be associated with medium-size fibres, a unique relation of each of the skin modalities with one of the A, B, or C fibre groups cannot be supported. All of the cutaneous senses seem to be associated with some fibres of all diameters; furthermore, the C fibres (once thought to be restricted to the pain function) display quite specific sensitivities to nonpainful stimuli applied to the skin.

A major neural pathway for tactual impulses runs along the back (in the dorsal columns) of the spinal cord. Afferent fibres enter the cord from the cutaneous nerves and ascend without synaptic break in one (the ipsilateral) dorsal column. This is a very rapidly conducting pathway shared by fibres that mediate sensations of deep pressure and also kinesthesis. Other tactual, temperature, and pain information crosses the spinal cord close to the level of entry of the sensory fibres and ascends to the brain in contralateral pathways of the cord (the lateral and ventral spinothalamic tracts).

Each of the nerves distributed along the spinal cord contains a sensory bundle that serves a well-defined strip of skin (a dermatome) about 2.5 centimetres (one inch) or more wide on the body surface. Successive spinal nerves overlap, so that each place on the skin represents two and sometimes three dermatomes; this yields a segmented pattern of strips over the body from head to toe. All dermatomes feed into a single relay centre (the sensory thalamus) deep within the brain, where a precise three-dimensional layout of tactual sensitivity at the body surface can be found. The neurons in this part of the thalamus (the ventral posterolateral nucleus) are specific to particular skin senses (such as pressure) and form small and precise receptor fields. There is a second more diffuse thalamic system (in the posterior thalamic nuclei) where the receptor fields are large, perhaps bilateral, on the left and right sides, perhaps including one whole side of the body. The receptor fields here or the types of stimuli to which they respond are not clearly delineated.

The dissociation of cutaneous senses is dramatically demonstrated in the course of some diseases; for example, in a disorder called syringomyelia, degeneration of the central canal of the spinal cord leads to loss of pain and temperature sensitivity. Nevertheless, the sufferer still can experience pressure. In some instances there may be a complete absence of pain sensitivity with disastrous consequences for the welfare of the afflicted person. Such individuals are bruised and cut and even lose parts of the body because they are unable to sense the dangerous (painful) characteristics of stimuli. Among people born with total absence of the pain sense, there may not be demonstrable anatomic abnormality. Still other instances of dissociation of pain versus pressure occur in surgical procedures (such as tractotomy) in which spinal tracts or parts of the nerves leading into the brainstem are selectively cut. Such operations are designed specifically to relieve pain without unduly diminishing pressure sensitivity.

Pathways from the specific (ventral posterolateral) thalamus end (or project) in a narrow band of brain cortex (the posterior rolandic cortical sensory area in man) where there is a point-for-point representation of the body surface on the cortical surface. The cortical projection of the more diffuse (posterior) thalamic system is less well charted. There thus appears to be a dissociation between those tactual structures that are highly specific and those that are more generalized.

Tactual psychophysics. The mixture of sensitivities within a given patch of skin provides a ready basis for the concept of adequate stimulation. Sometimes, for example, a cold spot responds to a very warm stimulus, and the person experiences what is called paradoxical cold. The sensation of heat from a hot stimulus presumably arises from the adequate stimulation of warmth receptors combined with the inadequate or inappropriate (albeit effective) stimulation of cold and pain receptors.

Paradoxical cold

Human ability to barely detect pressure (*i.e.,* the human pressure threshold) generally appears when a tension of about 0.85 grams per square millimetre (equivalent to about 1.2 pounds per square inch) of skin surface is applied on the back of the hand. Thus a force of 85 milligrams applied to a stimulus hair (or bristle) of 0.1 square millimetre is just about enough to elicit the experience of pressure. The energy of impact at pressure threshold is very much greater than that required for hearing or seeing, the skin requiring on the order of 100,000,000 times more energy than the ear and 10,000,000,000 times more energy than the eye. Differential pressure discrimination (the ability to detect just noticeable differences in intensity) requires changes of roughly 14 percent at maximum sensitivity.

Adaptation to pressure is well known; one's awareness of a steadily applied bristle fades and ultimately disappears. As a result people are rarely aware of the steady pressure of their clothing unless movement brings about a change in stimulation. Most dramatic and perhaps best known among tactual experiences is adaptation to thermal stimulation. Continued presentation of a warm or cold stimulus leads to reduction or disappearance of the initial sensation and an increase in threshold values. Total obliteration of thermal sensation through adaptation occurs in the range from about 16° to 42° C (61° to 108° F). If one hand is placed in a bowl of hot (40° C [104° F]) water and adapted to that, and at the same time the other hand is adapted to cold (20° C [68° F]) water, then when both hands are simultaneously placed in lukewarm (30° C [86° F]) water, the previously cooled hand feels warm and the other hand feels cold. This effect was once interpreted as evidence for a single temperature sense, but careful study shows that there are indeed two kinds of temperature receptors, both of which show adaptation. Cold receptors are characterized by an electrical discharge on sudden cooling, normally showing no response to sudden warming; similarly appropriate electrical responses are made by warmth receptors. Both receptors show steady discharges selectively depending on temperature; maximum discharge typically occurs between 38° to 43° C (100° to 109° F) for individual warmth cells and between 15° and 34° C (59° and 93° F) for cold receptors. These temperature receptors show no electrical response to weak mechanical stimulation in either laboratory animals or human subjects.

Pain

Pain is least understood among all the human senses. The pattern of stimulation is more crucial in pain than in any other sense. A single brief electric shock to the skin or to an exposed nerve may not elicit the experience of pain; yet it tends to become painful upon repetitive stimulation. Cutaneous pain is often sensed more sharply than is pain associated with deep tissues of the body (*e.g.,* viscera). Certain areas of the body are relatively analgesic (free of pain); for example, the mucous lining of the cheek into which one can bite shallowly without discomfort. The organs of the abdominal cavity are usually insensitive to cutting or burning, but traction or stretching of hollow viscera is painful (as when the stomach is distended by gas). Pain displays sensory adaptation, although the process appears to be more complex than it is for other sensory modalities. Thus, the intensities of headaches, toothaches, and pains from injury often show cyclic fluctuations, possibly from such factors as changes in blood circulation or in degree of inflammation. The visceral pains, those of dental origin, or of diseased tissues can be reduced by analgesic drugs, which tend to be less effective on cutaneous pain. Pain has a strong emotional context. In certain cases, after frontal lobotomies (a type of brain surgery) have been performed, a person may report that he still feels the pain of a pin prick or other irritation but that he does not find it as disturbing or emotionally disruptive as he did before the lobotomy. Many phenomena indicate the powerful role of the brain and spinal cord in sensing potentially painful sensory input. Indeed, according to one theory, a so-called gate control system in the spinal cord modulates (increases or decreases) sensory input from the skin to determine whether the input is perceived as painful. This theoretical formulation also may account for moment-to-moment fluctuations in the intensity of perceived pain despite the absence of any stimulus change. Such brain-mediated factors as emotional tension or past psychological experience are held to influence pain perception by acting upon this spinal gate control system.

Itching

Itching seems to bear the same relation to pain as tickle does to pressure. The experience usually lasts long enough to demand attention and (like tickle) normally leads to a response such as rubbing or scratching the affected area. A number of skin disorders are accompanied by itching, presumably from a fairly low level of irritation in the affected area (which also may be produced in undiseased skin). While a single shock by a low-intensity electrical spark normally produces no sensation, a repetitive pattern of such shocks may induce an itch not unlike that produced by an insect bite. Itching also may occur as an aftereffect of the sharp pricking sensation produced by single strong shocks, presumably because the nerves continue to produce a patterned afterdischarge following the cessation of the stimulus. Nonpainful tactile pattern stimulation is exemplified by vibration. Different frequencies of vibration are readily discriminated and a tactual communication system employing vibrations on the skin has been devised, particularly for people who cannot see or hear. Further research will probably reveal other ways of utilizing the fine discriminatory capacities of the cutaneous senses as substitutes for other sensory avenues of communication.

KINESTHETIC (MOTION) SENSE

It is a common experience that, with the eyes closed, one is aware of the positions of his legs and arms and can perceive the active or passive movement of a limb and its direction. The term kinesthesis (literally "feeling of motion") has been coined for this sensibility.

Nerve function. Four types of sensory structures are widely distributed in muscles, tendons, and joints (Figure 33): (1) the neuromuscular spindle consists of small, fine muscle fibres around which sensory fibre endings are wrapped; (2) the Golgi tendon organ consists of sensory nerve fibres that terminate in a rich branching encapsulated within the tendon; (3) joint receptors (as in the knee) consist of "spray-type" endings (Ruffini) and Golgi-type and Pacinian corpuscles within the joints; and (4) free nerve endings. All these receptors combine to provide information on active contraction, passive stretch of muscle fibres, and tension whether produced actively or passively. In passive stretch both the muscle-spindle receptors and the tendon receptors send trains of impulses over their sensory (afferent) nerves; in active contraction the spindles show a silent period of neural activity when tension on the parallel fibres is unloaded, while the tendon receptors discharge just as when stretch is passive.

From E. Gardner, *Fundamentals of Neurology*, 5th ed. (1968); W.B. Saunders Company, Philadelphia

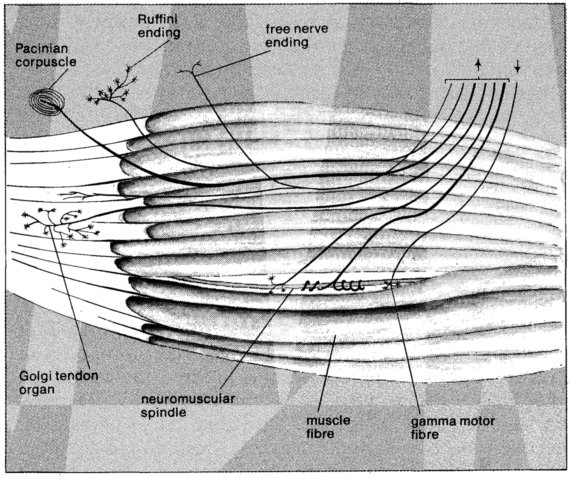


Figure 33: Sensory structures of muscle, tendon, and joint. Arrows indicate direction of conduction. (Motor nerves and motor nerve endings on contractile muscle fibres not shown.)

The muscle spindle is contractile in response to its own small-diameter, so-called gamma motor (efferent) fibre. The receptors and the gamma fibres of the muscle spindle form a neuromuscular loop that serves to insure that tension on the spindle is maintained within its efficient operating limits. The excitability of the muscle spindle also can be influenced through other neural pathways that control the general level of excitability of the central nervous system (brain and spinal cord). Activity of the descending reticular formation (a network of cells in the brainstem) may enhance the contraction of the spindle and hence influence its neural discharges.

The knee jerk

Muscle and tendon receptors combine to play an intimate and crucial role in the regulation of reflex and voluntary movement. Much of this control is automatic (involuntary) and not directly perceptible except in the aftereffects of movement or change of position. The knee jerk that follows a tap just below the kneecap of a freely hanging leg is one such involuntary reflex. Sensory (afferent) impulses from stretching the receptors (*e.g.*, in the muscles) relay to the spinal cord and activate a path to the motor (efferent) nerves leading back to the same muscle. The knee jerk is a purely spinal reflex response (the brain is not required) which is tested usually to determine nerve damage or other interference with the spinal-cord motor mechanisms. Besides producing loss of knee jerk, a disease such as syphilis may lead to locomotor ataxia (a clumsy and stumbling gait) when the germ (called a spirochete) attacks the sensory nerves of the cord's dorsal column. The result is that the sufferer has difficulty sensing the position of his limbs. Another general function of the muscle receptors is the maintenance of muscle tone (or tonus: partial contraction) to permit rapid response (fast reaction time) to stimulation. In normal conditions the muscle has tone and is ready to go; but, when it is without motor stimulation (deafferented), the muscle is flaccid, showing little tone. One's upright posture depends on the tonus of opposing (extensor and flexor) muscles in response to the effects of gravity.

The exact contribution of the muscle receptors to sensation is not entirely understood. It seems clear, however, that they are not essential to the sensation of bodily position. It has been suggested that the appreciation of passive movement of the limbs probably comes largely from the joints, since, after anesthetizing the overlying skin and muscles (*e.g.*, with cocaine), sensibility to the limb movement seems little affected. Evidence also shows that very few of the impulses arising from the muscle receptors themselves reach the cerebral cortex; instead, they ascend in the spinal pathways to another part of the brain (the cerebellum) where they interact in the automatic control of bodily movement. Impulses arising from the joint receptors, on the other hand, have been recorded in both the thalamus and brain cortex, the degree of angular displacement of a joint being reflected systematically in these structures by the frequency of nerve impulses. Symptoms of some diseases also emphasize the importance of joint sensitivity. When bone disease, for example, destroys only the joint receptors the person's ability to appreciate posture and movement is lost.

Cerebellar disorders

The feedback system. The feedback system leading to muscle tonus is a most delicately balanced mechanism. The gamma loop feeds back information that serves to maintain the muscle tonus and postural adjustments appropriate to the efficient performance of different voluntary actions. The afferent input from the muscle spindles via the spinal cord to the brain's cerebellum traverses an extensive circuit involving interactions of excitatory and inhibitory processes, the end result of which helps insure smooth and finely coordinated movements. Disease or other neurological damage in the cerebellum is characterized by distortions of movement and posture. Some sufferers of cerebellar disorders display crude, overactive motor activity (ballistic movement) that overshoots the mark in attempted voluntary coordination. Other victims of cerebellar disease display what resembles a drunken gait, jerkily stumbling and swaying along. By relying on other sensory cues (*e.g.*, visual and tactual), some people are able to compensate for awkward and uncoordinated movements associated with cerebellar damage.

VESTIBULAR SENSE (EQUILIBRIUM)

Function of the inner ear. The human inner ear contains parts (the nonauditory labyrinth or vestibular organ) that are sensitive to acceleration in space, rotation, and orientation in the gravitational field. Rotation is signalled by way of the semicircular canals, three bony tubes in each ear that lie embedded in the skull roughly at right angles to each other. These canals are filled with fluid (endolymph); in the ampulla of each canal are receptor cells with fine hairs that project up into the fluid to be displaced as the endolymph lags behind when rotation begins. When rotation is maintained at a steady velocity, the fluid catches up, and stimulation of the hair cells no longer occurs until rotation suddenly stops, again circulating the endolymph. Whenever the hair cells are thus stimulated, one normally experiences a sensation of rotation in space. During rotation one exhibits reflex nystagmus (back-and-forth movement) of the eyes. Slow displacement of the eye occurs against the direction of rotation and serves to maintain the gaze at a fixed point in space; this is followed by a quick return to the initial eye position in the direction of the rotation. Stimulation of the hair cells in the absence of actual rotation tends to produce an apparent swimming of the visual field, often associated with dizziness and nausea; compensatory postural adjustments commonly follow.

Two sacs or enlargements of the vestibule (the saccule and utricle) react to steady (static) pressures; *e.g.,* those of gravitational forces. Hair cells within these structures are covered by a gelatinous cap in which are embedded small granular particles of calcium carbonate (otoliths) that weigh against the hairs. When iron particles are implanted in the same structures of a fish, they may be displaced by externally applied magnets. In this way the fish can be made to assume inverted or other unusual positions in the water. In man, unusual stimulation of the vestibular receptors and semicircular canals also can give rise to sensory distortions in visual and motor activity. The resulting discord between one's visual and motor responses and the external space (as aboard ship in a heaving sea) often leads to nausea and disorientation (*e.g.,* seasickness). In space flight these sensory systems usually are not stimulated except as the weightless astronaut affects them by his own movements. Such abnormal gravitational and acceleratory forces apparently contribute to the nausea or disequilibrium sometimes reported by people in outer space. Training before space flight reduces the likelihood and severity of these symptoms.

Seasickness

Factors affecting equilibrium. In some diseases (*e.g.,* ear infections), irritation of vestibular nerve endings occurs, and the sufferer may be subject to falling as well as to spells of disorientation and vertigo (dizzy confusion). Similar symptoms may be induced by flushing hot and cold water into the outer opening of the ear, since the temperature changes produce currents in the endolymph of the semicircular canals. This caloric (temperature) effect is used in clinical tests for vestibular functions and in physiological experiments. Externally applied electrical currents may also stimulate the nerve endings of the vestibule. When current is applied to the right mastoid bone (just behind the ear), nystagmus to the right tends to occur, associated with a reflex right movement of the head; movement tends to the left for the opposite mastoid. In man, destruction of the labyrinth in only one ear causes vertigo and other vestibular symptoms, such as nystagmus, inaccurate pointing, and tendency to fall. (The vestibular functions of the ear are described in detail below; see *Human hearing*, with special attention to the section *Ear diseases and hearing disorders: the inner ear.*)

TASTE (GUSTATORY) SENSE

The sensory structures for taste in man are the taste buds, clusters of cells contained in goblet-shaped structures (papillae) that open by a small pore to the mouth cavity. A single bud contains about 50 to 75 slender cells, all arranged in a bananalike cluster pointed toward the gustatory pore (see also *Chemoreception* above). These are the taste receptor cells, which differentiate from the surrounding epithelium, grow to mature form, and then die out to be replaced by new cells in a turnover period as short as seven to 10 days. The various types of cells in the taste bud appear to be different stages in this turnover process. Slender nerve fibres entwine among and make contact usually with many cells. In man and other mammals, taste buds are located primarily in fungiform (mushroom-shaped), foliate, and circumvallate (walled-around) papillae of the tongue or in adjacent structures of the palate and throat. Many gustatory receptors in small papillae on the soft palate and back roof of the mouth in human adults are particularly sensitive to sour and bitter, whereas the tongue receptors are relatively more sensitive to sweet and salt. Some loss of taste sensitivity suffered among wearers of false teeth may be traceable to mechanical interference of the denture with taste receptors on the roof of the mouth.

Locations of taste buds

Nerve supply. There is no single sensory nerve for taste in vertebrates. In man, the anterior (front) two-thirds of the tongue is supplied by one nerve (the lingual nerve), the back of the tongue by another (the glossopharyngeal nerve), and the throat and larynx by certain branches of a third (the vagus nerve), all of which subserve touch, temperature, and pain sensitivity in the tongue as well as taste. The gustatory fibres of the anterior tongue leave the lingual nerve to form a slender nerve (the chorda tympani) that traverses the eardrum on the way to the brain stem. When the chorda tympani at one ear is cut or damaged (by injury to the eardrum), taste buds begin to disappear and gustatory sensitivity is lost on the anterior two-thirds of the tongue on the same side. Impulses have been recorded from the human chorda tympani, and good correlations have been found between the reports people give of their sensations of taste and of the occurrence of the afferent nerve discharge. The taste fibres from all the sensory nerves from the mouth come together in the brainstem (medulla oblongata). Here and at all levels of the brain, gustatory fibres run in distinct and separate pathways, lying close to the pathways for other modalities from the tongue and mouth cavity. From the brain's medulla, the gustatory fibres ascend by a pathway to a small cluster of cells in the thalamus and thence to a taste-receiving area in the anterior cerebral cortex.

Physiological basis of taste. No simple relation has been found between chemical composition of stimuli and the quality of gustatory experience except in the case of acids. The taste qualities of inorganic salts (such as potassium bromide, a sedative) are complex; epsom salt (magnesium sulfate) commonly is sensed as bitter, while table salt (sodium chloride) is typical of sodium salts, which usually yield the familiar saline taste. Experiences of sweet and bitter are elicited by many different classes of chemical compound.

Theorists of taste sensitivity classically posited only four basic or primary types of human taste receptors, one for each gustatory quality: salty, sour, bitter, and sweet. Yet, recordings of sensory impulses in the taste nerves of laboratory animals show that many individual nerve fibres from the tongue are of mixed sensitivity, responding to more than one of the basic taste stimuli, such as acid plus salt or acid plus salt plus sugar. Other individual nerve fibres respond to stimuli of only one basic gustatory quality. Most numerous, however, are taste fibres subserving two basic taste sensitivities; those subserving one or three qualities are about equal in number and next most frequent; fibres that respond to all four primary stimuli are least common. Mixed sensitivity may be only partly attributed to multiple branches of taste nerve endings. In man, experiences of sugars, synthetic sweeteners, weak salt solutions, and the taste of some unpleasant medicines are blocked by a drug (gymnemic acid) obtained from *Gymnema* bushes native to India. Among some laboratory animals, gymnemic acid blocks only the nerve response to sugar, even if the fibre mediates other taste qualities. Such a multiresponsive fibre still can transmit taste impulses (*e.g.,* for salt or sour), so that blockage by the drug can be attributed to chemically specific sites or cells in the taste bud.

Types of taste receptors

In some species of animals (*e.g.,* the cat), specific taste receptors appear to be activated by water; these so-called water receptors are inhibited by weak saline solutions. Water taste might be considered a fifth gustatory quality in addition to the basic four.

The qualities of taste. *Sour.* The hydrogen ions of acids (*e.g.,* hydrochloric acid, HCl) are largely responsible for the sour taste; but, although a stimulus grows more sour as its hydrogen ion (H^+) concentration increases, this factor alone does not determine sourness. Weak organic acids (*e.g.,* the acetic acid in vinegar) taste more sour than would be predicted from their hydrogen ion concentration alone; apparently the rest of the acid molecule affects the efficiency with which hydrogen ions stimulate.

Salt. Although the salty taste is often associated with water-soluble salts, most such compounds (except sodium chloride) have complex tastes such as bitter-salt or sour-salt. Salts of low molecular weight are predominantly salty, while those of higher molecular weight tend to be bitter. The salts of heavy metals such as mercury have a metallic taste, although some of the salts of lead (especially lead acetate) and beryllium are sweet. Both parts of the molecule (*e.g.,* lead and acetate) contribute to taste quality and to stimulating efficiency. In man the following series for degree of saltiness, in decreasing order, is found: ammonium (most salty), potassium, calcium, sodium, lithium, and magnesium salts (least salty). The order appears to vary for other animals.

Sweet. Except for some salts of lead or beryllium, the sweet taste is associated largely with organic compounds (such as alcohols, glycols, sugars, and sugar derivatives). Human sensitivity to synthetic sweeteners (*e.g.,* saccharin) is especially remarkable; the taste of saccharin can be detected in a dilution 700 times weaker than that required for cane sugar. The stereochemical (spatial) arrangement of atoms within a molecule may affect its taste; thus, slight changes within a sweet molecule will make it bitter or tasteless (Figure 34). Several theorists have proposed that the common feature for all of the sweet stimuli is the presence in the molecule of a so-called proton acceptor, such as the OH (hydroxyl) components of carbohydrates (*e.g.,* sugars) and many other sweet-tasting compounds. It has also been theorized that such molecules will not taste sweet unless they are of appropriate size.

From R.W. Moncrieff, *Chemical Senses* (1944); Leonard Hill Books

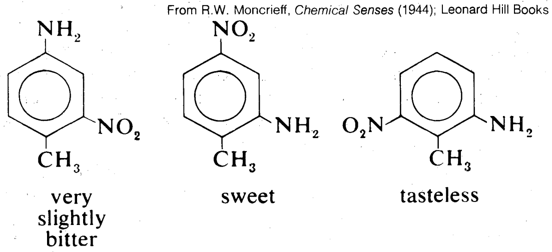


Figure 34: Effects of molecular arrangement on taste sensation (see text).

Bitter. The experience of a bitter taste is elicited by many classes of chemical compounds and often is found in association with sweet and other gustatory qualities. Among the best known bitter substances are such alkaloids (often toxic) as quinine, caffeine, and strychnine. Most of these substances have extremely low taste thresholds and are detectable in very weak concentrations. The size of such molecules is theoretically held to account for whether or not they will taste bitter. An increase in molecular weight of inorganic salts or an increase in length of chains of carbon atoms in organic molecules tends to be associated with increased bitterness.

Taste blindness

A substantial minority of people exhibit specific taste blindness, an inability to detect as bitter such chemicals as PTC (phenylthiocarbamide). Taste blindness for PTC and other carbamides appears to be hereditary (as a recessive trait), occurring least frequently among American Indians and Africans, in about a third of Europeans, and in roughly 40 percent of the people in Western India. Evidence for such taste blindness among other animals has been observed only in anthropoid apes; apparently it appeared at a relatively late stage of evolution. Taste blindness for carbamides is not correlated with insensitivity to other bitter stimuli, the reasons for this being poorly understood.

Factors affecting taste sensitivity. Fluids of extreme temperature, especially those that are cold, may produce temporary taste insensitivity. People generally seem to taste most acutely when the stimulus is at or slightly below body temperature. When the tongue and mouth are first adapted to the temperature of a taste solution, sugar sensitivity increases with temperature rise, salt and quinine sensitivity decrease, and acid sensitivity is relatively unchanged. Gustatory adaptation (partial or complete disappearance of taste sensitivity) may occur if a solution is held in the mouth for a period of time. The effect of one adapting stimulus on the person's sensitivity for another one (cross adaptation) is especially common with substances that are chemically similar and that elicit the same taste quality. Adaptation to sodium chloride will reduce one's ability to sense the saltiness of a variety of the inorganic salts but will leave undiminished or even enhance such qualities as bitterness, sweetness, or sourness that were part of the taste of the salt before adaptation. Likewise, adaptation by one acid may reduce sensitivity to the sourness of other acids.

Contrast effects

Adaptation studies often are complicated by so-called contrast effects; for example, people say that distilled water tastes sweet following their exposure to a weak acid. Water may take on other taste qualities as well; following one's adaptation to a sour-bitter chemical (urea), water may taste salty. Adaptation tends to diminish or enhance the effect of a subsequent stimulus depending on whether the two stimuli normally elicit the same or a contrasting taste. Thus the adapted sweetness of water and all normally sweet-tasting substances are enhanced after one has tasted acid (sour). The bitterness of tea and coffee or the sourness of lemon are masked or suppressed by sugar or saccharin.

The human gustatory difference threshold (for a just noticeable difference in intensity) is approximately a 20 percent change in concentration. For very weak taste stimuli, however, the threshold sensitivity is poorer.

Food choice. One's ability to taste is intimately concerned in his eating habits or in his rejection of noxious substances. One of the earliest reflex responses of the infant, that of sucking, can be controlled by gustatory stimuli. Sweet solutions are sucked more readily than is plain water; bitter, salty, or sour stimuli tend to stop the sucking reflex.

Among insects, a very specific feeding reaction (proboscis extension) is so automatic that it is widely used as an index of taste stimulation. If a common housefly is held relatively immobile in wax, different parts of the mouth, legs, and body readily may be stimulated by a drop of solution. Sugar solution will make the fly extend its proboscis when a drop is applied to the legs or mouth parts. A fly that has been starved will show a positive response to a weak sugar solution that ordinarily would not affect one that is satiated. Addition of salt or acid to the sugar solution inhibits this response.

Many animals provide clear examples of beneficially selective feeding behaviour. Laboratory rats, when given unhampered choice of carbohydrates, proteins, vitamins, and minerals (each in a separate container), show consistent patterns of self-selection that may be modified by physiological stresses and strains. A rat made salt-deficient by removal of its adrenal glands, for example, will increase its intake of sodium chloride sufficiently to maintain health and growth; normally, such gland removal is fatal in the absence of salt-replacement therapy. Histories of similar effects have been reported in human beings, one dramatic case being that of a child with adrenal disorder who kept himself alive by satisfying an intense salt craving.

Among human adults, past experience shows a strong influence on eating habits, sometimes to the point that physiological well-being suffers. Food habits and other factors play a significant role in eating behaviour.

Poisonous substances often are unpalatable, but not invariably. Lead acetate, sometimes called sugar of lead, once was used as a sweetening agent with disastrous results before its potentially fatal effects were appreciated. Many palatable substances, including some synthetic sweeteners, are toxic; taste alone is not a reliable guide to safety. A rat poison, alpha-naphthylthiourea (ANTU), was developed in a relatively insoluble and therefore tasteless form; soluble forms of ANTU had all been rejected by the animals. Taste aversions also may be readily established by conditioning, even for substances that have been normally preferred. In one study, a rat tasted saccharin solution three hours before being exposed to enough radiation to become sick. When the animal recovered, it was found to have a strong aversion to the taste of saccharin. Other aversions selectively can be produced by injecting the individual with a nauseating drug following specific taste experience. An unusual finding is that long delays of up to several hours in the time between the presentation of the taste stimulus and the induction of illness do not prevent the conditioning. In most other studies, only brief intervals (perhaps up to minutes in duration) have been found to result in successful conditioning. Bait shyness developed by wild rats that survive poisoning strongly suggests conditioned taste aversion. Positive preferences also are subject to conditioning, as when the tastes of drugs or vitamins become associated with the feelings of well-being they generate.

SMELL (OLFACTORY) SENSE

Olfactory receptors

In mammals the olfactory receptors are located high in the nasal cavity. The yellow-pigmented olfactory membrane in humans covers about 2.5 square centimetres (0.4 square inch) on each side of the inner nose. The olfactory sense receptor is a long thin cell ending in several delicate hairs (cilia) that project into and through the mucus that normally covers the nasal epithelium or lining (Figure 35). Electron microscope photographs show from six to 12 olfactory cilia per cell. The end of each receptor narrows to a fine nerve fibre, which, along with many others, enters the olfactory bulb of the brain through a fine channel in the bony roof of the nasal cavity. The olfactory membrane of a young rabbit contains about 100,000,000 ciliated receptor cells that provide as much area as total skin surface of the animal.

From *Physiological Psychology* by Peter M. Milner. Copyright © 1970 by Holt, Rinehart and Winston, Inc. Reprinted by permission of Holt, Rinehart and Winston, Inc. (After De Lorenzo, 1963)

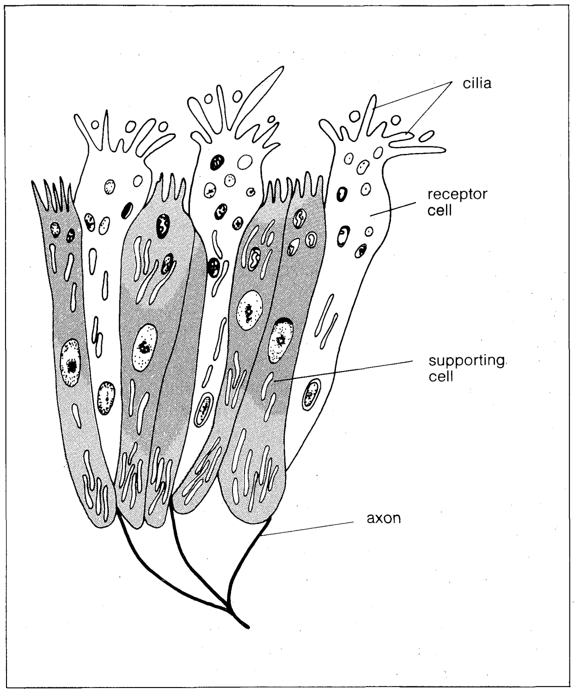


Figure 35: Olfactory lining of the nose.

Pain endings of the trigeminal nerve fibres are widely distributed throughout the human nasal cavity, including the olfactory region. Relatively mild odorants, such as orange oil, as well as the more obvious irritants, such as ammonia, stimulate such free nerve endings as well as the olfactory receptors.

Odorous molecules may be carried to the olfactory region by slight eddies in the air during quiet breathing, but vigorous sniffing brings a surge into the olfactory region. Odour sensitivity may be impaired by blocking the nasal passages mechanically, as when membranes are congested by infection.

It is generally agreed that the antennae of insects are their principal olfactory sites, but specific mouth parts may also bear endings for smell. The nonliving outer covering (cuticle) of such structures has been shown to be pierced by many ultramicroscopic holes through which odorous molecules enter, presumably to dissolve the fluid underneath. In vertebrates (such as man), olfactory stimulation occurs only after the odour molecule is dissolved in the mucus that covers the olfactory membrane. In spite of the wide biological gap between them, odour sensitivity of man and that of insect display certain similarities of mechanism.

Olfactory bulbs

In vertebrates the olfactory nerve fibres enter either of two specialized structures (olfactory bulbs), stemlike projections under the front part of the brain, to end in a microscopic series of intricate basket-like clusters called glomeruli. Each glomerulus receives impulses from about 26,000 receptors and sends them on through other cells, eventually to reach higher olfactory centres at the base of the brain. There are also fibres that cross over from one olfactory bulb to the other. When the olfactory bulbs are removed by surgery, the individual's ability to discriminate odours is lost; details of higher brain centres for smell are still unclear.

Olfactory qualities. The vocabulary of odour is rich with the names of substances that elicit a great variety of olfactory qualities. One of the best known published psychological attempts at classification was in 1916 on the basis of tests of more than 400 different scents on human observers. On the basis of the apparent similarities of perceived odour quality or confusions in naming, it was concluded that there were six main odour qualities: fruity, flowery, resinous, spicy, foul, and burned.

Electrical activity can be detected readily with fine insulated wires inserted into the olfactory bulb. Those portions of the bulb toward the anterior or oral (mouth) region in the rabbit are found to be more sensitive to water-soluble substances, whereas the more posterior or aboral (away from the mouth) parts of the olfactory bulb are more sensitive to fat-soluble substances. In addition, when very fine electrodes are used, individual cells (so-called mitral cells) are found to be sensitive to different groups of chemicals. Evidence for the existence of only a few primary receptors, however, does not emerge from such studies; a variety of different combinations of sensitivity has been found. Similarly, recordings from the primary receptor nerve fibres reveal different patterns of sensitivity. Electrical recording of this type also shows that olfactory sensitivity can be enhanced by a painful stimulus, such as a pinch on the foot of an experimental animal. This appears to be a reflex that serves to enhance the detection of dangerous stimuli in the environment. Different parts of the olfactory neural pathways seem to be selectively tuned to discriminate different classes of olfactory information. Thus, the third- and fourth-order olfactory neurons found beyond the olfactory bulb of the rat seem particularly concerned with distinguishing the odour of sexually receptive females. These neurons appear to be especially important in the preference the male rat shows for the smell of urine from a female in heat.

Characteristics that impart odour

Odorous substances. To be odorous, a substance must be sufficiently volatile for its molecules to be given off and carried into the nostrils by air currents. The solubility of the substance also seems to play a role; chemicals that are soluble in water or fat tend to be strong odorants, although many of them are inodorous. No unique chemical or physical property that can be said to elicit the experience of odour has yet been defined.

Only seven of the chemical elements are odorous: fluorine, chlorine, bromine, iodine, oxygen (as ozone), phosphorus, and arsenic. Most odorous substances are organic (carbon-containing) compounds in which both the arrangement of atoms within the molecule as well as the particular chemical groups that comprise the molecule influence odour. Stereoisomers (*i.e.,* different spatial arrangements of the same molecular components) may have different odours. On the other hand, a series of different molecules that derive from benzene all have a similar odour. It is of historic interest that the first benzene derivatives studied by chemists were found in pleasant-smelling substances from plants (such as oil of wintergreen or oil of anise), and so the entire class of these compounds was labelled aromatic. Subsequently, other so-called aromatic compounds were identified that have less attractive odours.

The scent of flowers and roots (such as ginger) depends upon the presence of minute quantities of highly odorous essential oils. Although the major odour constitutents can be identified by chemical analysis, some botanical essences are so complex that their odours can be duplicated only by adding them in small amounts to synthetic formulations.

Odour sensitivity. In spite of the relative inaccessibility of the human olfactory receptor cells, odour stimuli can be detected at extremely low concentrations. Olfaction is said to be 10,000 times more sensitive than taste. A human threshold value for such a well-known odorant as ethyl mercaptan (found in rotten meat) has been cited in the range of 1/400,000,000th of a milligram per litre of air. A just-noticeable difference in odour intensity may be apparent when there is a 20 percent increase in odorant strength, but at low concentrations as much as 100 percent increase in concentration may be required. Temperature influences the strength of an odour by affecting the volatility and hence the emission of odorous particles from the source; humidity also affects odour for the same reasons. Hunting dogs can follow a spoor (odour trail) most easily when high humidity retards evaporation and dissipation of the odour. Perfumes contain chemicals called fixatives, added to retard evaporation of the more volatile constituents. The temporary anosmia (absence of sense of smell) following colds in the nose may be complete or partial; in the latter case, only the odours of certain substances are affected. Paranosmia (change in perceived odour quality) also may occur during respiratory infections. Changes in sensitivity are reported to occur in women during the menstrual cycle, particularly in regard to certain odorants (steroids) related to sex hormones. Olfactory sensitivity also is said to become more acute during hunger.

Factors affecting odour sensitivity

Human adaptation to odours is so striking that the stench of a slaughterhouse or chemical laboratory ceases to be a nuisance after a few minutes have passed. Olfactory adaptation, as measured by a rise in threshold, is especially pronounced for stronger odours. Cross adaptation (between different odours) may take place; thus, eucalyptus oil may be difficult to detect after one becomes adapted to the smell of camphor.

Adaptation long was regarded solely as the result of changes in the olfactory receptor; however, electrical recordings show that the receptor cells in the nose seem to adapt only partially. Rhythmic discharges continue in the olfactory bulb long after the experimenter ceases to detect the odour that is stimulating the experimental animal. Apparently, some olfactory adaptation may occur in the brain as well as in the sense organ.

Odour blending and flavour. The ancient art of perfumery and the modern science of odour control depend on mixing and blending. Two different odours presented together may be readily identifiable. The more they resemble each other, the greater will be the tendency to blend; yet trained workers usually can discriminate the components of a successfully blended perfume. The substantially greater intensity of one odour may mask another. Masking, however, is less effective than chemical conversion or physical collection (removal) as a basis for odour control. Some odorants also may be removed by passing air through activated charcoal.

Flavours and the sense of smell

The distinctive flavours of food are known largely through the sense of smell. Flavour is the composite of experiences from many senses, but the aroma of roast beef and the delicate bouquet of wine are mainly olfactory in origin. Flavour technology has assumed a well-entrenched place in the food industry; it is common practice for manufacturers to use flavour panels of several trained members to judge the flavour of new food products. Using psychophysical methods with adequate statistical control, such panels are very reliable and can detect qualities so subtle that they defy the methods of physics and chemistry.

In using large, untrained consumer panels, the emphasis is on acceptability, often an emotional reaction to the product. Such untrained judgments are unstable and may vary among individuals because of idiosyncrasy or differences in experience. For example, the strong cheese odour so palatable to the gourmet can produce revulsion in the uninitiated.

Effects on behaviour. Recognition of friend or foe by social insects may depend upon olfactory cues: certain ants attack their own kind furiously if deprived of the sense of smell by amputation of their antennae; bees entering a strange hive are put to death because the scent of a foreign hive clings to them. Bees also have a specialized organ on the end of the abdomen that deposits a scent on a newly discovered food source to guide other foraging workers. The scents of flowers attest to the evolutionary importance of odour; those flowers that attract insects most efficiently are the likeliest to be pollinated and to reproduce their kind.

The effect of odours on the sexual behaviour of invertebrates is most striking; a female moth, for example, was observed to attract more than 100 males during relatively brief observation periods of about six hours. The physiological basis for the attraction can be traced to specific odour attractants (pheromones), molecules produced by the scent glands of the female. Pheromones also function in defense and as alarms of impending danger. The specific sex attractant of the female silk moth (*Bombyx*) has been identified and synthesized. It has been found that the antennae of the male silk moth contain olfactory receptors specifically sensitive to the female pheromone but that the females have no receptors to detect their own attractiveness.

Mammals in the wild state appear to utilize their odour glands for sexual attraction. Laboratory rats show a preference for the branch of a maze that has been scented with the odour of a sexually receptive female. It is likely that some rudiments of these effects operate in man. The most sexually provocative perfumes have a high proportion of musk or musklike odour. Genuine musk is derived from the sexual glands of the musk deer and is chemically related to human sex hormones; odour sensitivity in humans varies with the menstrual cycle.

Among laboratory animals the secretion of reproductive hormones can be markedly influenced by odour stimulation. This seems to be an innate physiological process rather than the result of learning. A most dramatic effect (pregnancy block) is observed when the odour of a strange male is presented to a recently mated female. The normal hormonal changes following copulation are blocked under these conditions, and the fertilized egg fails to survive. A related study of the periodicity and length of menstrual cycle in women exposed to the normal odours of men suggests there may be similar effects among people. Human behaviour, molded and shaped by custom and culture though it is, has many of its roots in his basic sensual appetites. (C.Pf.)

HUMAN VISION: STRUCTURE AND FUNCTION OF THE EYE

Anatomy of the visual apparatus

STRUCTURES AUXILIARY TO THE EYE

The orbit. The eye is protected from mechanical injury by being enclosed in a socket, or orbit, which is made up of portions of several of the bones of the skull to form a four-sided pyramid the apex of which points back into the head. Thus, the floor of the orbit is made up of parts of the maxilla, zygomatic, and palatine bones, while the roof is made up of the orbital plate of the frontal bone and, behind this, by the lesser wing of the sphenoid. The optic foramen, the opening through which the optic nerve runs back into the brain and the large ophthalmic artery enters the orbit, is at the nasal side of the apex; the superior orbital fissure is a larger hole through which pass large veins and nerves. These nerves may carry nonvisual sensory messages—*e.g.*, pain—or they may be motor nerves controlling the muscles of the eye. There are other fissures and canals transmitting nerves and blood vessels. The eyeball and its functional muscles are surrounded by a layer of orbital fat that acts much like a cushion, permitting a smooth rotation of the eyeball about a virtually fixed point, the centre of rotation. The protrusion of the eyeballs—proptosis—in exophthalmic goitre is caused by the collection of fluid in the orbital fatty tissue.

The eyelids. It is vitally important that the front surface of the eyeball, the cornea, remain moist. This is achieved by the eyelids, which during waking hours sweep

The layers in the lids

the secretions of the lacrimal apparatus and other glands over the surface at regular intervals and which during sleep cover the eyes and prevent evaporation. The lids have the additional function of preventing injuries from foreign bodies, through the operation of the blink reflex. The lids are essentially folds of flesh covering the front of the orbit and, when the eye is open, leaving an almond-shaped aperture. The points of the almond are called canthi; that nearest the nose is the inner canthus, and the other is the outer canthus (Figure 36). The lid may be divided into four layers: (1) the skin, containing glands that open onto the surface of the lid margin, and the eyelashes; (2) a muscular layer containing principally the orbicularis oculi muscle, responsible for lid closure; (3) a fibrous layer that gives the lid its mechanical stability, its principal portions being the tarsal plates, one in each lid, which border directly upon the opening between the lids, called the palpebral aperture; and (4) the innermost layer of the lid, a portion of the conjunctiva. The conjunctiva is a mucous membrane that serves to attach the eyeball to the orbit and lids but permits a considerable degree of rotation of the eyeball in the orbit.

From P. Kronfeld, *The Eye*, vol. 1 (1962); Academic Press

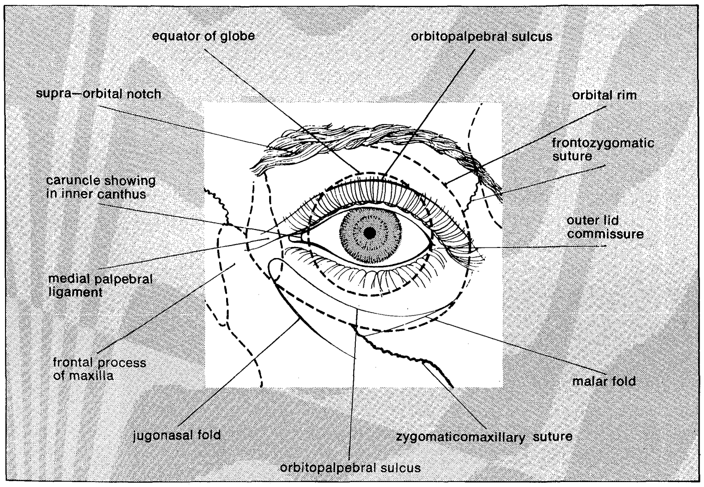


Figure 36: Frontal view of the eye and its related structures (see text).

The conjunctiva. The conjunctiva lines the lids and then bends back over the surface of the eyeball, constituting an outer covering to the forward part of this and terminating at the transparent region of the eye, the cornea. The portion that lines the lids is called the palpebral portion of the conjunctiva; the portion covering the white of the eyeball is called the bulbar conjunctiva. Between the bulbar and the palpebral conjunctiva there are two loose, redundant portions forming recesses that project back toward the equator of the globe. These recesses are called the upper and lower fornices, or conjunctival sacs; it is the looseness of the conjunctiva at these points that makes movements of lids and eyeball possible. (The ophthalmologist finds the lower conjunctival sac a useful cavity in which to place drops containing drugs. He accomplishes this by merely pulling the outer lid away from the globe. The drops are retained in the cavity long enough to act directly on the cornea and to diffuse through this into the internal structures of the eye.)

The fibrous layer. The fibrous layer, which gives the lid its mechanical stability, is made up of the thick, and relatively rigid, tarsal plates, bordering directly on the palpebral aperture, and the much thinner palpebral fascia, or sheet of connective tissue; the two together are called the septum orbitale. When the lids are closed, the whole opening of the orbit is covered by this septum. Two ligaments, the medial and lateral palpebral ligaments, attached to the orbit and to the septum orbitale, stabilize the position of the lids in relation to the globe. The medial ligament, by far the stronger, is well illustrated in Figure 36.

Horner's muscle and the muscle of Riolan

The muscles. Closure of the lids is achieved by contraction of the orbicularis muscle, a single oval sheet of muscle extending from the regions of the forehead and face and surrounding the orbit into the lids. It is divided into orbital and palpebral portions, and it is essentially the palpebral portion, within the lid, that causes lid closure. The palpebral portion passes across the lids from a ligament called the medial palpebral ligament and from the neighbouring bone of the orbit in a series of half ellipses that meet outside the outer corner of the eye, the lateral canthus, to form a band of fibres called the lateral palpebral raphe. Additional parts of the orbicularis have been given separate names—namely, Horner's muscle and the muscle of Riolan; they come into close relation with the lacrimal apparatus and assist in drainage of the tears. The muscle of Riolan, lying close to the lid margins, doubtless contributes to keeping the lids in close apposition, an important feature for maintaining the junction watertight. The orbital portion of the orbicularis is not normally concerned with blinking, which may be carried out entirely by the palpebral portion; however, it is concerned with closing the eyes tightly. The skin of the forehead, temple, and cheek is then drawn toward the medial (nose) side of the orbit, and the radiating furrows, formed by this action of the orbital portion, eventually lead to the so-called crow's feet of elderly persons. It must be appreciated that the two portions can be activated independently; thus, the orbital portion may contract, causing a furrowing of the brows that reduces the amount of light entering from above, while the palpebral portion remains relaxed and allows the eyes to remain open.

Opening of the eye is not just the result of passive relaxation of the orbicularis muscle but also is the effect of the contraction of the levator palpebrae superioris muscle of the upper lid. This muscle takes origin with the extraocular muscles at the apex of the orbit (the back of the eye socket) as a narrow tendon and runs forward into the upper lid as a broad tendon, the levator aponeurosis, which is attached to the forward surface of the tarsus and the skin covering the upper lid. Contraction of the muscle causes elevation of the upper eyelid. The nervous connections of this muscle are closely related to those of the extraocular muscle required to elevate the eye, so that when the eye looks upward the upper eyelid tends to move up in unison.

The orbicularis and levator are striped muscles under voluntary control. The lids contain, in addition, unstriped (involuntary) muscle fibres that are activated by the sympathetic division of the autonomic system and tend to widen the palpebral fissure (the eye opening) by elevation of the upper, and depression of the lower, lid.

In addition to the muscles already described, other facial muscles often cooperate in the act of lid closure or opening. Thus, the corrugator supercilii muscles pull the eyebrows toward the bridge of the nose, making a projecting "roof" over the medial angle of the eye and producing characteristic furrows in the forehead; the roof is used primarily to protect the eye from the glare of the sun. The pyramidalis, or procerus, muscles occupy the bridge of the nose; they arise from the lower portion of the nasal bones and are attached to the skin of the lower part of the forehead on either side of the midline; they pull the skin into transverse furrows. In lid opening, the frontalis

By permission from Eugene Wolff, *Anatomy of the Eye and Orbit*. London: H.K. Lewis & Co.

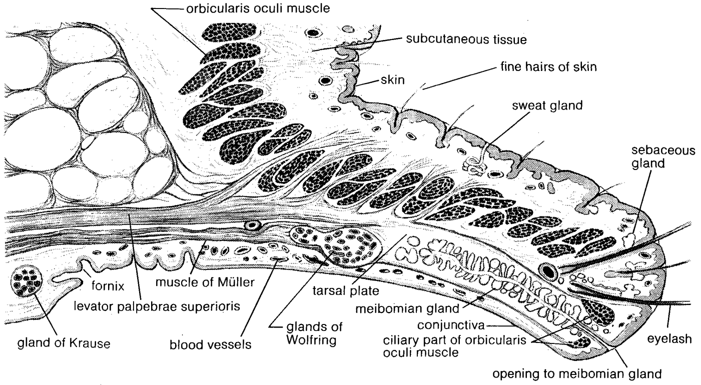


Figure 37: Vertical section through upper lid (see text).

muscle, arising high on the forehead, midway between the coronal suture, a seam across the top of the skull, and the orbital margin, is attached to the skin of the eyebrows. Contraction therefore causes the eyebrows to rise and opposes the action of the orbital portion of the orbicularis; the muscle is especially used when one gazes upward. It is also brought into action when vision is rendered difficult either by distance or the absence of sufficient light.

The skin. The outermost layer of the lid is the skin, with features not greatly different from skin on the rest of the body, with the possible exception of large pigment cells, which, although found elsewhere, are much more numerous in the skin of the lids. The cells may wander, and it is these movements of the pigment cells that determine the changes in coloration seen in some people with alterations in health. The skin has sweat glands and hairs. As the junction between skin and conjunctiva is approached, the hairs change their character to become eyelashes (Figure 37).

Tear glands

The glandular apparatus. The eye is kept moist by secretions of the lacrimal glands (tear glands). These almond-shaped glands under the upper lids extend inward from the outer corner of each eye. Each gland has two portions. One portion is in a shallow depression in the part of the eye socket formed by the frontal bone. The other portion projects into the back part of the upper lid. The ducts from each gland, three to 12 in number, open into the superior conjunctival fornix, or sac. From the fornix, the tears flow down across the eye and into the puncta lacrimalia, small openings at the margin of each eyelid near its inner corner. The puncta are openings into the lacrimal ducts; these carry the tears into the lacrimal sacs, the dilated upper ends of the nasolacrimal ducts, which carry the tears into the nose.

The evaporation of the tears as they flow across the eye is largely prevented by the secretion of oily and mucous material by other glands. Thus, the meibomian, or tarsal glands, consist of a row of elongated glands extending through the tarsal plates; they secrete an oil that emerges onto the surface of the lid margin and acts as a barrier for the tear fluid, which accumulates in the grooves between the eyeball and the lid barriers.

Extraocular muscles. Six muscles outside the eye govern its movements. These muscles are the four rectus muscles—the inferior, medial, lateral, and superior recti—and the superior and inferior oblique muscles. The rectus muscles arise from a fibrous ring that encircles the optic nerve at the optic foramen, the opening through which the nerve passes, and are attached to the sclera, the opaque portion of the eyeball, in front of the equator, or widest part, of the eye. The superior oblique muscle arises near the rim of the optic foramen and somewhat nearer the nose than the origin of the rectus medialis. It ends in a rounded tendon that passes through a fibrous ring, the trochlea, that is attached to the frontal bone. The trochlea acts as a pulley. The tendon is attached to the sclera back of the equator of the eye (Figure 38).

The inferior oblique muscle originates from the floor of the orbit, passes under the eyeball like a sling, and is attached to the sclera between the attachments of the superior and lateral rectus muscles. The rectus muscles direct the gaze upward and downward and from side to side. The inferior oblique muscle tends to direct the eye upward, and the superior oblique to depress the eye; because of the obliqueness of the pull, each causes the eye to roll, and in an opposite direction.

The oblique muscles are strictly antagonistic to each other, but they work with the vertical rectus muscles in so far as the superior rectus and inferior oblique both tend to elevate the gaze and the inferior rectus and superior oblique both tend to depress the gaze. The superior and inferior recti do not produce a pure action of elevation or depression because their plane of action is not exactly vertical; in consequence, as with the obliques, they cause some degree of rolling, but by no means so great as that caused by the obliques; the direction of rolling caused by the rectus muscle is opposite to that of its synergistic oblique; the superior rectus causes the eye to roll inward, and the inferior oblique outward.

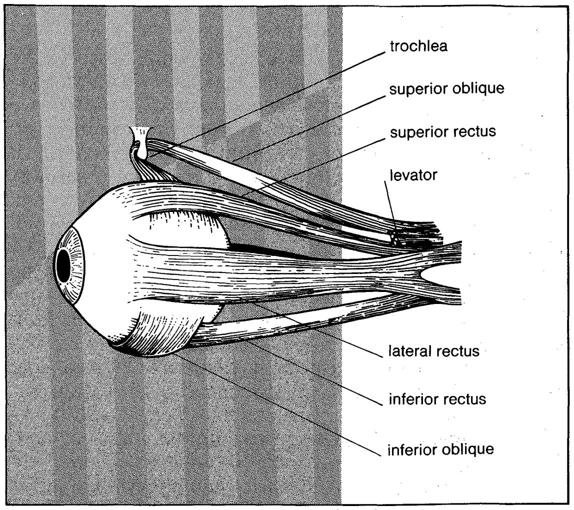


Figure 38: The extraocular muscles.

From P. Kronfeld, *The Eye*, vol. 1 (1962); Academic Press

THE EYE

Visible parts of the eye

General description. The eyeball is not a simple sphere but can be viewed as the result of fusing a small portion of a small, strongly curved sphere with a large portion of a large, not so strongly curved sphere (Figure 39). The small piece, occupying about one-sixth of the whole, has a radius of eight millimetres (0.3 inch); it is transparent and is called the cornea; the remainder, the scleral segment, is opaque and has a radius of 12 millimetres (0.5 inch). The ring where the two areas join is called the limbus. Thus, on looking directly into the eye from in front one sees the white sclera surrounding the cornea; because the latter is transparent one sees, instead of the cornea, a ring of tissue lying within the eye, the iris. The iris is the structure that determines the colour of the eye. The centre of this ring is called the pupil. It appears dark because the light passing into the eye is not reflected back to any great extent. By use of an ophthalmoscope, an instrument that permits the observer to illuminate the interior of the eyeball while observing through the pupil, the appearance of the interior lining of the globe can be made out; this is called the fundus; it is characterized by the large blood vessels that suppy blood to the retina; these are especially distinct as they cross over the pallid optic disk, or papilla, the region where the optic nerve fibres leave the globe.

From H. Davson, M.D., *Physiology of the Eye*

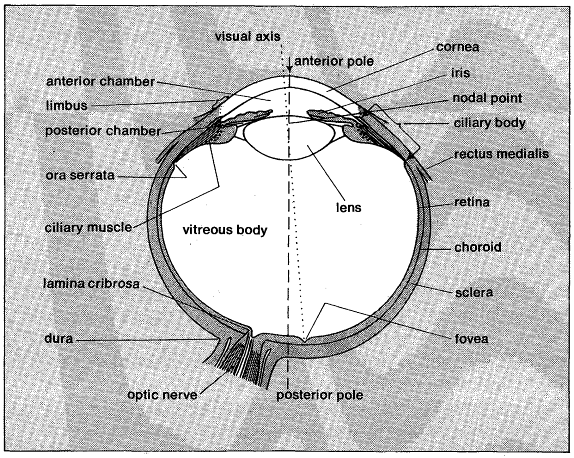


Figure 39: Horizontal section of the eye.

The dimensions of the eye are reasonably constant, varying among individuals by only a millimetre or two; the sagittal (vertical) diameter is about 24 millimetres (about one inch) and is usually less than the transverse diameter. At birth the sagittal diameter is about 16 to 17 millimetres (about 0.65 inch); it increases rapidly to about 22.5 to 23 millimetres (about 0.89 inch) by the age of three years; between three and 13 the globe attains its full size. The weight is about 7.5 grams (.25 ounce), and its volume 6.5 millilitres (0.4 cubic inch).

The eye is made up of three coats, which enclose the optically clear aqueous humour, lens, and vitreous body (Figure 39). The outermost coat consists of the cornea and the sclera; the middle coat contains the main blood supply to the eye and consists, from the back forward, of the choroid, the ciliary body, and the iris. The innermost layer is the retina, lying on the choroid and receiving most of its nourishment from the vessels within the choroid, the remainder of its nourishment being derived from the retinal vessels that lie on its surface and are visible in the ophthalmoscope. The ciliary body and iris have a very thin covering, the ciliary epithelium and posterior epithelium of the iris, which is continuous with the retina.

Within the cavities formed by this triple-layered coat there are the crystalline lens, suspended by fine transparent fibres—the suspensory ligament or zonule of Zinn—from the ciliary body; the aqueous humour, a clear fluid filling the spaces between the cornea and the lens and iris; and the vitreous body, a clear jelly filling the much larger cavity enclosed by the sclera, the ciliary body, and the lens. The anterior chamber of the eye is defined as the space between the cornea and the forward surfaces of the iris and lens, while the posterior chamber is the much smaller space between the rear surface of the iris and the ciliary body, zonule, and lens; the two chambers both contain aqueous humour and are in connection through the pupil.

Outer and middle tunics of the globe. *The outermost coat.* The outermost coat is made up of the cornea and the sclera. The cornea is the transparent window of the eye. It contains five distinguishable layers; the epithelium, or outer covering; Bowman's membrane; the stroma, or supporting structure; Descemet's membrane; and the endothelium, or inner lining. Up to 90 percent of the thickness of the cornea is made up of the stroma. The epithelium, which is a continuation of the epithelium of the conjunctiva, is itself made up of about six layers of cells. The superficial layer is continuously being shed, and the layers are renewed by multiplication of the cells in the innermost, or basal, layer.

The layers of the cornea

The stroma appears as a set of lamellae, or plates, running parallel with the surface and superimposed on each other like the leaves of a book; between the lamellae lie the corneal corpuscles, cells that synthesize new collagen (connective tissue protein) essential for the repair and maintenance of this layer. The lamellae are made up of microscopically visible fibres that run parallel to form sheets; in successive lamellae the fibres make a large angle with each other. The lamellae in man are about 1.5 to 2.5 microns (one micron = 0.001 millimetre) thick, so that there are about 200 lamellae in the human cornea. The fibrous basis of the stroma is collagen.

Immediately above the stroma, adjacent to the epithelium, is Bowman's membrane, about eight to 14 microns thick; in the electron microscope it is evident that it is really stroma, but with the collagen fibrils not arranged in the orderly fashion seen in the rest of the stroma.

Beneath the stroma are Descemet's membrane and the endothelium. The former is about five to 10 microns thick and is made up of a different type of collagen from that in the stroma; it is secreted by the cells of the endothelium, which is a single layer of flattened cells. There is apparently no continuous renewal of these cells as with the epithelium, so that damage to this layer is a more serious matter.

The sclera is essentially the continuation backward of the cornea, the collagen fibres of the cornea being, in effect, continuous with those of the sclera. The sclera is pierced by numerous nerves and blood vessels; the largest of these holes is that formed by the optic nerve, the posterior scleral foramen. The outer two-thirds of the sclera in this region continue backward along the nerve to blend with its covering, or dural sheath—in fact, the sclera may be regarded as a continuation of the dura mater, the outer covering of the brain. The inner third of the sclera, combined with some choroidal tissue, stretches across the opening, and the sheet thus formed is perforated to permit the passage of fasciculi (bundles of fibres) of the optic nerve. This region is called the lamina cribrosa (Figure 39). The blood vessels of the sclera are largely confined to a superficial layer of tissue, and these, along with the conjunctival vessels, are responsible for the bright redness of the inflamed eye. As with the cornea, the innermost layer is a single layer of endothelial cells; above this is the lamina fusca, characterized by large numbers of pigment cells.

The most obvious difference between the opaque sclera and the transparent cornea is the irregularity in the sizes and arrangement of the collagen fibrils in the sclera by contrast with the almost uniform thickness and strictly parallel array in the cornea; in addition, the cornea has a much higher percentage of mucopolysaccharide (a carbohydrate that has among its repeating units a nitrogenous sugar, hexosamine) as embedding material for the collagen fibrils. It has been shown that the regular arrangement of the fibrils is, in fact, the essential factor leading to the transparency of the cornea.

Differences between sclera and cornea

When the cornea is damaged—*e.g.,* by a virus infection—the collagen laid down in the repair processes is not regularly arranged, with the result that an opaque patch called a leukoma, may occur.

When an eye is removed, or a man dies, the cornea soon loses its transparency, becoming hazy; this is due to the taking in of fluid from the aqueous humour, the cornea becoming thicker as it becomes hazier. The cornea can be made to reassume its transparency by maintaining it in a warm, well-aerated chamber, at about 31° C (88° F, its normal temperature); associated with this return of transparency is a loss of fluid.

Modern studies have shown that, under normal conditions, the cornea tends to take in fluid, mainly from the aqueous humour and from the small blood vessels at the limbus, but this is counteracted by a pump that expels the fluid as fast as it enters. This pumping action depends on an adequate supply of energy, and any situation that prejudices this supply causes the cornea to swell—the pump fails, or works so slowly that it cannot keep pace with the leak. Death is one cause of the failure of the pump, but this is primarily because of the loss of temperature; place the dead eye in a warm chamber and the reserves of metabolic energy it contains in the form of sugar and glycogen are adequate to keep the cornea transparent for 24 hours or more. When it is required to store corneas for grafting, as in an eye bank, it is best to remove the cornea from the globe to prevent it from absorbing fluid from the aqueous humour. The structure responsible for the pumping action is almost certainly the endothelium, so that damage to this lining can lead to a loss of transparency with swelling.

The cornea is exquisitely sensitive to pain. This is mediated by sensory nerve fibres, called ciliary nerves, that run just underneath the endothelium; they belong to the ophthalmic branch of the fifth cranial nerve, the large sensory nerve of the head. The ciliary nerves leave the globe through holes in the sclera, not in company with the optic nerve, which is concerned exclusively with responses of the retina to light.

The uvea. The middle coat of the eye is called the uvea (from the Latin for "grape") because the eye looks like a reddish-blue grape when the outer coat has been dissected away. The posterior part of the uvea, the choroid, is essentially a layer of blood vessels and connective tissue sandwiched between the sclera and the retina. The forward portion of the uvea, the ciliary body and iris, is more complex, containing as it does the ciliary muscle and the sphincter and dilator of the pupil.

Ciliary body and iris

The blood supply to the human eye is twofold, consisting of the retinal and uveal circulations, both of which derive from branches of the ophthalmic artery. The two systems of blood vessels differ in that the retinal vessels, which supply nutrition to the innermost layers of the retina, derive from a branch of the ophthalmic artery, called the central artery of the retina, that enters the eye with the optic nerve, while the uveal circulation, which supplies the middle and outer layers of the retina as well as the uvea, is derived from branches of the ophthalmic artery that penetrate the globe independently of the optic nerve.

The ciliary body is the forward continuation of the choroid. It is a muscular ring, triangular in horizontal section, beginning at the region called the ora serrata and ending, in front, as the root of the iris (Figure 40). The surface is thrown into folds, called ciliary processes, the whole being covered by the ciliary epithelium, which is a double layer of cells; the layer next to the vitreous body (see below), called the inner layer, is transparent, while the outer layer, which is continuous with the pigment epithelium of the retina, is heavily pigmented. These two layers are to be regarded embryologically as the forward continuation of the retina, which terminates at the ora serrata. Their function is to secrete the aqueous humour.

From P. Kronfeld, *The Eye*, vol. 1 (1962); Academic Press

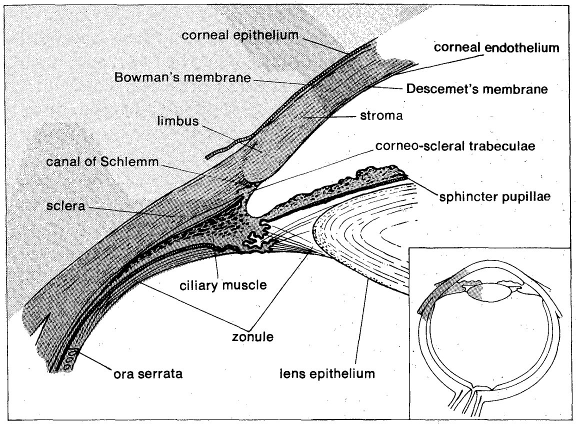


Figure 40: The anteronasal portion of a horizontal (meridional) section through a right eye. Shaded area of inset locates magnified portion. (See text.)

The ciliary muscle is an unstriped, involuntary, muscle concerned with alterations in the adjustments of focus—accommodation—of the optical system; the fibres run both across the muscle ring and circularly, and the effect of their contraction is to cause the whole body to move forward and to become fatter, so that the suspensory ligament that holds the lens in place is loosened.

The most forward portion of the uvea is the iris. This is the only portion that is visible to superficial inspection, appearing as a perforated disc, the central perforation, or pupil, varying in size according to the surrounding illumination and other factors. A prominent feature is the collarette at the inner edge, representing the place of attachment of the embryonic pupillary membrane that, in embryonic life, covers the pupil. As with the ciliary body, with which it is anatomically continuous, the iris consists of several layers: namely, an anterior layer of endothelium, the stroma; and the posterior iris epithelium. The stroma contains the blood vessels and the sphincter and dilator muscles; in addition, the stroma contains pigment cells that determine the colour of the eye. Posteriorly, the stroma is covered by a double layer of epithelium, the continuation forward of the ciliary epithelium; here, however, both layers are heavily pigmented and serve to prevent light from passing through the iris tissue, confining the optical pathway to the pupil. The pink iris of the albino is the result of the absence of pigment in these layers. The cells of the anterior layer of the iris epithelium have projections that become the fibres of the dilator muscle; these projections run radially, so that when they contract they pull the iris into folds and widen the pupil; by contrast, the fibres of the sphincter pupillae muscle run in a circle around the pupil, so that when they contract the pupil becomes smaller.

Usually, a baby belonging to the white races is born with blue eyes because of the absence of pigment cells in the stroma; the light reflected back from the posterior epithelium, which is blue because of scattering and selective absorption, passes through the stroma to the eye of the observer. As time goes on, pigment is deposited, and the colour changes; if much pigment is laid down the eye becomes brown or black, if little, it remains blue or gray.

The inner tunic of the globe. The inner tunic of the rear portion of the globe, as far forward as the ciliary body, is the retina, including its epithelia or coverings. These epithelia continue forward to line the remainder of the globe.

The pigment epithelium

The epithelia. Separating the choroid (the middle tunic of the globe) from the retina proper is a layer of pigmented cells, the pigment epithelium of the retina; this acts as a restraining barrier to the indiscriminate diffusion of material from the blood in the choroid to the retina. The retina ends at the ora serrata, where the ciliary body begins (Figure 39). The pigment epithelium continues forward as a pigmented layer of cells covering the ciliary body; farther forward still, the epithelium covers the posterior surface of the iris and provides the cells that constitute the dilator muscle of this diaphragm. Next to the pigment epithelium of the retina is the neuroepithelium, or rods and cones (see below). Their continuation forward is represented by a second layer of epithelial cells covering the ciliary body, so that by the ciliary epithelium is meant the two layers of cells that are the embryological equivalent of the retinal pigment epithelium and the receptor layer (rods and cones) of the retina. This unpigmented layer of the ciliary epithelium is continued forward over the back of the iris, where it acquires pigment and is called the posterior iris epithelium.

The retina. The retina is the part of the eye that receives the light and converts it into chemical energy. The chemical energy activates nerves that conduct the messages out of the eye into the higher regions of the brain. The retina is a complex nervous structure, being, in essence, an outgrowth of the forebrain.

Ten layers of cells in the retina can be seen microscopically. In general, there are four main layers: (1) Next to the choroid is the pigment epithelium, already mentioned. (2) Beneath the epithelium is the layer of rods and cones, the light-sensitive cells. The changes induced in the rods and cones by light are transmitted to (3) a layer of neurons (nerve cells) called the bipolar cells, which are analogous to the sensory neurons that carry messages from the touch and heat receptors of the skin and transmit them to the cells of the spinal cord or the medulla (the part of the brain that is a continuation of the spinal cord). These bipolar cells connect with (4) the innermost layer of neurons, the ganglion cells; and the transmitted messages are carried out of the eye along their projections, or axons, which constitute the optic nerve fibres. Thus, the optic nerve is really a central tract, rather than a nerve, connecting two regions of the nervous system, namely, the layer of bipolar cells, and the cells of the lateral geniculate body, the latter being a visual relay station in the diencephalon (the rear portion of the forebrain).

The arrangement of the retinal cells in an orderly manner gives rise to the outer nuclear layer (layer 4 in Figure 41), containing the nuclei of the rods and cones; the inner nuclear layer (layer 6), containing the nuclei and perikarya (main cell bodies outside the nucleus) of the bipolar cells, and the ganglion cell layer (layer 8), containing the corresponding structures of the ganglion cells. The plexiform layers are regions in which the neurons make their interconnections. Thus, the outer plexiform layer (layer 5) contains the rod and cone projections terminating as the rod spherule and cone pedicle; these make connections with the dendritic processes of the bipolar cells, so that changes produced by light in the rods and cones are transmitted by way of these connections to the bipolar cells. (The dendritic process of a nerve cell is the projection that receives nerve impulses to the cell; the axon is the projection that carries impulses from the cell.) In the inner plexiform layer (layer 7) are the axons of the bipolar cells and the dendritic processes of the ganglion and amacrine cells (see below). The association is such as to allow messages in the bipolar cells to be transmitted to the ganglion cells, the messages then passing out along the axons of the ganglion cells as optic nerve messages.

The rods and cones

The photosensitive cells are, in the human and in most vertebrate retinas, of two kinds, called rods and cones, the rods being usually much thinner than the cones but both being built up on the same plan. The light-sensitive pigment is contained in the outer segment (layer 2), which rests on the pigment epithelium (layer 1). Through the other end, called the synaptic body, effects of light are

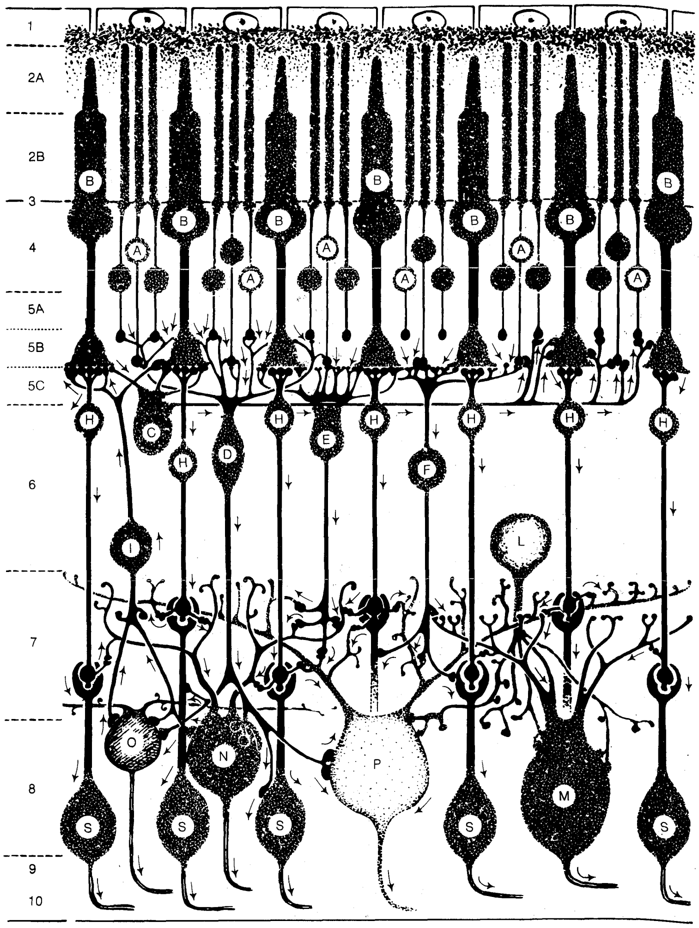


Figure 41: *The retinal cells directly involved in the visual process.* (A) Rod and (B) cone cells, the photoreceptors. (C) Horizontal cells. (D) Mop, (E) brush, (F) flat-topped, (H) midget, and (I) centrifugal varieties of bipolar cells. (L) Several varieties of amacrine cells. (M) Parasol, (N) shrub, (O) small diffuse, (P) garland, and (S) midget ganglion varieties of ganglion cells.

By courtesy of Stephen L. Polyak

transmitted to the bipolar and horizontal cells. When examined in the electron microscope, the outer segments of the rods and cones are seen to be composed of stacks of disks, apparently made by the infolding of the limiting membrane surrounding the outer segment; the visual pigment, located on the surfaces of these disks, is thus spread over a very wide area, and this contributes to the efficiency with which light is absorbed by the visual cell.

The arrangement of the retina makes it necessary for light to pass through the layers not sensitive to light first before it reaches the light-sensitive rods and cones. The optical disadvantages of this arrangement are largely overcome by the development of the fovea centralis, a localized region of the retina, close to the optic axis of the eye, where the inner layers of the retina are absent. The result is a depression, the foveal pit, where light has an almost unrestricted passage to the light-sensitive cells. It is essentially this region of the retina that is employed for accurate vision, the eyes being directed toward the objects of regard so that their images fall in this restricted region. If the object of interest is large, so as to subtend a large angle, then the eye must move rapidly from region to region so as to bring their images successively onto the fovea; this is typically seen during reading. In the central region of the fovea there are cones exclusively; toward its edges, rods also occur, and as successive zones are reached the proportion of rods increases while the absolute density of packing of the receptors tends to decrease. Thus, the central fovea is characterized by an exclusive population of very densely packed cones; here, also, the cones are very thin and in form very similar to rods. The region surrounding the fovea is called the parafovea; it stretches about 1,250 microns from the centre of the fovea, and it is here that the highest density of rods occurs. Surrounding the parafovea, in turn, is the perifovea, its outermost edge being 2,750 microns from the centre of the fovea; here the density of cones is still further diminished, the number being only 12 per hundred microns compared with 50 per hundred microns in the most central region of the fovea. In the whole human retina there are said to be about 7,000,000 cones and from 75,000,000 to 150,000,000 rods.

Retinal organization

The fovea is sometimes referred to as the macula lutea ("yellow spot"); actually this term defines a rather vague area, characterized by the presence of a yellow pigment in the nervous layers, stretching over the whole central retina; *i.e.,* the fovea, parafovea, and perifovea.

The blind spot in the retina corresponds to the optic papilla, the region on the nasal side of the retina through which the optic nerve fibres pass out of the eye.

Although the rods and cones may be said to form a mosaic, the retina is not organized in a simple mosaic fashion in the sense that each rod or cone is connected to a single bipolar cell that itself is connected to a single ganglion cell. There are only about 1,000,000 optic nerve fibres, while there are at least 150,000,000 receptors, so that there must be considerable convergence of receptors on the optic pathway. This means that there will be considerable mixing of messages. Furthermore, the retina contains additional nerve cells besides the bipolar and ganglion cells; these, the horizontal and amacrine cells, operate in the horizontal direction, allowing one area of the retina to influence the activity of another. In this way, for example, the messages from one part of the retina may be suppressed by a visual stimulus falling on another, an important element in the total of messages sent to the higher regions of the brain. Finally, it has been argued that some messages may be running the opposite way; they are called centrifugal and would allow one layer of the retina to affect another, or higher regions of the brain to control the responses of the retinal neurons. In primates the existence of these centrifugal fibres has been finally disproved, but in such lower vertebrates as the pigeon, their existence is quite certain.

The pathway of the retinal messages through the brain is described later in this article; it is sufficient to state here that most of the optic nerve fibres in primates carry their messages to the lateral geniculate body, a relay station specifically concerned with vision. Some of the fibres separate from the main stream and run to a midbrain centre called the pretectal nucleus, which is a relay centre for pupillary responses to light.

Secretion and course of aqueous humour

The transparent media. Within the cavities enclosed by the three layers of the globe described above there are the aqueous humour in the anterior and posterior chambers; the crystalline lens behind the iris; and the vitreous body, which fills the large cavity behind the lens and iris (Figure 39).

The aqueous humour. The aqueous humour is a clear colourless fluid with a chemical composition rather similar to that of blood plasma (the blood exclusive of its cells) but lacking the high protein content of the latter. Its main function is to keep the globe reasonably firm. It is secreted continuously by the ciliary body into the posterior chamber, and flows as a gentle stream through the pupil into the anterior chamber, from which it is drained by way of a channel at the limbus; that is, the juncture of the cornea and the sclera. This channel, the canal of Schlemm, encircles the cornea and connects by small connector channels to the blood vessels buried in the sclera and forming the intrascleral plexus or network. From this plexus the blood, containing the aqueous humour, passes into more superficial vessels; it finally leaves the eye in the anterior ciliary veins. The relation of the canal of Schlemm to the aqueous humour is clear from Figure 40. The wall of the canal that faces the aqueous humour is very delicate and allows the fluid to percolate through by virtue of the relatively high pressure of the fluid within the eye. Obstruction of this exit, for example, if the iris is pushed forward to cover the wall of the canal, causes a sharp rise in the pressure within the eye, a condition that is known as glaucoma. Often the obstruction is not obvious, but is caused perhaps by a hardening of the tissue just adjacent to the wall of the canal—the trabecular meshwork (Figure 40), in which case the rise of pressure is more gradual and insidious. Ultimately the abnormal pressure damages the retina and causes a variable degree of blindness. The normal intraocular pressure is about 15 millimetres (0.6 inch) of mercury above atmospheric pressure, so that if the anterior chamber is punctured by a hypodermic needle the aqueous humour flows out readily. Its function in maintaining the eye reasonably hard is seen by the collapse and wrinkling of the cornea when the fluid is allowed to escape. An additional function of the fluid is to provide nutrition for the crystalline lens and also for the cornea, both of which are devoid of blood vessels; the steady renewal and drainage serve to bring into the eye various nutrient substances, including glucose and amino acids, and to remove waste products of metabolism.

The vitreous body. The vitreous body is a jelly. It is remarkable for the small amount of solid matter required to give it this semisolid structure; the solid material is made up of a form of collagen, vitrosin, and a mucopolysaccharide, hyaluronic acid. Thus, its composition is rather similar to that of the cornea, but the proportion of water is much greater, about 98 percent or more, compared with about 75 percent for the cornea. The jelly is probably secreted by certain cells of the retina. In general, the vitreous body is devoid of cells, in contrast with the lens, which is packed tight with cells. Embedded in the surface of the vitreous body, however, there is a population of specialized cells, the hyalocytes of Balazs, which may contribute to the breakdown and renewal of the hyaluronic acid. The vitreous body serves to keep the underlying retina pressed against the choroid.

The crystalline lens. The lens is a transparent body, flatter on its anterior than on its posterior surface, and suspended within the eye by the zonular fibres of Zinn attached to its equator; its anterior surface is bathed by aqueous humour, and its posterior surface by the vitreous body. The lens is a mass of tightly packed transparent fibrous cells, the lens fibres, enclosed in an elastic collagenous capsule. The lens fibres are arranged in sheets that form successive layers; the fibres run from pole to pole of the lens, the middle of a given fibre being in the equatorial region. On meridional (horizontal) section, the fibres are cut longitudinally to give an onion-scale appearance, whereas a section at right-angles to this—an equatorial section—would cut all the fibres across, and the result would be to give a honeycomb appearance. The epithelium, covering the anterior surface of the lens under the capsule, serves as the origin of the lens fibres, both during embryonic and fetal development and during infant and adult life, the lens continuing to grow by the laying down of new fibres throughout life.

The visual process

THE WORK OF THE AUXILIARY STRUCTURES

The protective mechanisms. The first line of protection of the eyes is provided by the lids, which prevent access of foreign bodies and assist in the lubrication of the corneal surface. Lid closure and opening are accomplished by the orbicularis oculi and levator palpebri muscles; the orbicularis oculi operates on both lids, bringing their margins into close apposition in the act of lid closure. Opening results from relaxation of the orbicularis muscle and contraction of the levator palpebri of the upper lid; the smooth muscle of the upper lid, Müller's muscle, or the superior palpebral muscle, also assists in widening the lid aperture. The lower lid does not possess a muscle corresponding to the levator of the upper lid, and the only muscle available for causing an active lowering of the lid, required during the depression of the gaze, is the inferior palpebral muscle, which is analogous to the muscle of Müller of the upper lid (called the superior palpebral muscle). This inferior palpebral muscle is so directly fused with the sheaths of the ocular muscles that it provides cooperative action, opening of the lid on downward gaze being mediated, in effect, mainly by the inferior rectus.

Blinking

Innervation. The seventh cranial nerve—the facial nerve—supplies the motor fibres for the orbicularis muscle. The levator is innervated by the third cranial nerve—the oculomotor nerve—that also innervates some of the extraocular muscles concerned with rotation of the eyeball, including the superior rectus. The smooth muscle of the eyelids and orbit is activated by the sympathetic division of the autonomic system. The secretion of adrenaline during such states of excitement as fear would also presumably cause contraction of the smooth muscle, but it seems unlikely that this would lead to the protrusion of the eyes traditionally associated with extreme fear. It is possible that the widening of the lid aperture occurring in this excited state, and dilation of the pupil, create the illusion of eye protrusion.

Blinking is normally an involuntary act, but may be carried out voluntarily. The more vigorous "full closure" of the lids involves the orbital portion of the orbicularis muscle and may be accompanied by contraction of the facial muscles that have been described as accessory muscles of blinking: namely, the corrugator supercilii, which on contraction pulls the eyebrows toward the bridge of the nose; and the procerus or pyramidalis, which pulls the skin of the forehead into horizontal folds, acting as a protection when the eyes are exposed to bright light. The more vigorous full closure may be evoked as a reflex response.

Voluntary full closure of lids

Blink reflexes. Reflex blinking may be caused by practically any peripheral stimulus, but the two functionally significant reflexes are (1) that resulting from stimulation of the endings of the fifth cranial nerve in the cornea, lid, or conjunctiva—the sensory blink reflex, or corneal reflex—and (2) that caused by bright light—the optical blink reflex. The corneal reflex is rapid (0.1 second reflex time) and is the last to disappear in deepening anesthesia, impulses being relayed from the nucleus of the fifth nerve to the seventh cranial nerve, which transmits the motor impulses. The reflex is said to be under the control of a medullary centre. The optical reflex is slower; in man, the nervous pathway includes the visual cortex (the outer substance of the brain; the visual centre is located in the occipital—rear—lobe). The reflex is absent in children of less than nine months.

Normal rhythm. In the waking hours the eyes blink fairly regularly at intervals of two to 10 seconds, the actual rate being a characteristic of the individual. The function of this is to spread the lacrimal secretions over the cornea. It might be thought that each blink would be reflexly determined by a corneal stimulus—drying and irritation—but extensive studies indicate that this view is wrong; the normal blinking rate is apparently determined by the activity of a "blinking centre" in the globus pallidus of the caudate nucleus, a mass of nerve cells between the base and the outer substance of the brain. This is not to deny that the blink rate is modified by external stimuli.

There is a strong association between blinking and the action of the extraocular muscles. Eye movement is generally accompanied by a blink, and it is thought that this aids the eyes in changing their fixation point.

Secretion of tears. The exposed surface of the globe (eyeball) is kept moist by the tears secreted by the lacrimal apparatus, together with the mucous and oily secretions of the other secretory organs and cells of the lids and conjunctiva; these have been described earlier. The secretion produces what has been called the precorneal film, which consists of an inner layer of mucus, a middle layer of lacrimal secretion, and an outer oily film that reduces the rate of evaporation of the underlying watery layer. The normal daily (24-hour) rate of secretion has been estimated at about 0.75 to 1.1 grams (0.03–0.04 ounce avoirdupois); secretion tends to decrease with age. Chemical analysis of the tears reveals a typical body fluid with a salt concentration similar to that of blood plasma. An interesting component is lysozyme, an enzyme that has bactericidal action by virtue of its power of dissolving away the outer coats of many bacteria.

Tear reflexes

Tears are secreted reflexly in response to a variety of stimuli—*e.g.*, irritative stimuli to the cornea, conjunctiva, nasal mucosa; hot or peppery stimuli applied to the mouth and tongue; or bright lights. In addition, tear flow occurs in association with vomiting, coughing, and yawning. The secretion associated with emotional upset is called psychical weeping. Severing of the sensory root of the trigeminal (fifth cranial) nerve prevents all reflex weeping, leaving psychical weeping unaffected; similarly, the application of cocaine to the surface of the eye, which paralyzes the sensory nerve endings, inhibits reflex weeping, even when the eye is exposed to potent tear gases. The afferent (sensory) pathway in the reflex is thus by way of the fifth cranial, or the trigeminal nerve. The motor innervation is by way of the autonomic (involuntary) division; the parasympathetic supply derived from the facial nerve (the seventh cranial nerve) seems to have the dominant motor influence. Thus, drugs that mimic the parasympathetic, such as acetylcholine, provoke secretion, and secretion may be blocked by such typical anticholinergic drugs as atropine. Innervation of the lacrimal gland is not always complete at birth, so that the newborn infant is generally said to cry without weeping. Because absence of reflex tearing fails to produce any serious drying of the cornea, and surgical destruction of the main lacrimal gland is often without serious consequences, it seems likely that the subsidiary secretion from the accessory lacrimal glands is adequate to keep the cornea moist. The reflex secretion that produces abundant tears may be regarded as an emergency response.

A drainage mechanism for tears is necessary only during copious secretion. The mechanism, described as the lacrimal pump, consists of alternately negative and positive pressure in the lacrimal sac caused by the contraction of the orbicularis muscle during blinking.

Movements of the eyes. Because only a small portion of the retina, the fovea, is actually employed for distinct vision, it is vitally important that the motor apparatus governing the direction of gaze be extremely precise in its operation, and rapid. Thus, the gaze must shift swiftly and accurately during the process of reading. Again, if the gaze must remain fixed on a single small object—*e.g.,* a golf ball—the eyes must keep adjusting their gaze to compensate for the continuous small movements of the head and to maintain the image exactly on the fovea. The extraocular muscles that carry out these movements are under voluntary control; thus, the direction of regard can be changed deliberately. Most of the actual movements of the eyes are carried out without awareness, however, in response to movements of the objects in the environment, or in response to movements of the head or the rest of the body, and so on. In examining the mechanisms of the eye movements, then, one must resolve them into a number of reflex responses to changes in the environment or the individual, remembering, of course, that there is an overriding voluntary control.

Centring images on the fovea

The axes of the eye. It is worthwhile at this point to define certain axes of the eyes employed during different types of study. The optic axis of the eye is a line drawn through the centre of the cornea and the nodal (central) point of the eye; it actually does not intersect with the retina at the centre of the fovea as might be expected, but toward the nose from this, so that there is an angle of about five degrees between (1) the visual axis—the line joining the point fixated (the point toward which the gaze is directed) and the nodal point—and (2) the optic axis.

Actions of muscles. The general modes of action of the six extraocular muscles have been described in connection with their anatomy: rotation of the eye toward the nose is carried out by the medial rectus; outward movement is by the lateral rectus. Upward movements are carried out by the combined actions of the superior rectus and the inferior oblique muscles, and downward movements by the inferior rectus and the superior oblique. Intermediate directions of gaze are achieved by combined actions of several muscles. When the two eyes act together, as they normally do, and change their direction of gaze to the left, for example, the left eye rotates away from the nose by means of its lateral rectus, while the right eye turns toward the nose by means of its medial rectus. These muscles may be considered as a linked pair; that is, when they are activated by the central nervous system this occurs conjointly and virtually automatically. This linking of the muscles of the two eyes is an important physiological feature and has still more important pathological interest in the analysis of squint, when the two eyes fail to be directed at the same point.

Binocular movements. The binocular movements (the movements of the two eyes) fall into two classes, the conjugate movements, when both eyes move in the same direction, as in a change in the direction of gaze, and disjunctive movements, when the eyes move in opposite directions. Thus, during convergence onto a near object both eyes move toward the nose; the movement is horizontal, but disjunctive, by contrast with the conjugate movement when both eyes move, say, to the right. The disjunctive movement of convergence can be carried out voluntarily, but the act is usually brought about reflexly in response to the changed optical situation—*i.e.,* the nearness of the object of gaze. A seesaw movement of the eyes, whereby one eye looks upward and the other downward, is possible, but not voluntarily; to achieve this a prism is placed in front on one eye so that the object seen through it appears displaced upward or downward; the other eye sees the object where it is. The result of such an arrangement is that, unless the eye with the prism in front makes an upward or downward movement, independent of the other, the images will not fall on corresponding parts of the retinas in the two eyes. Such a noncorrespondence of the retinal images causes double vision; to avoid this, there is an adjustment in the alignment of the eyes so that a seesaw movement is actually executed. In a similar way, the eyes may be made to undergo torsion, or rolling. A conjugate torsion, in which both eyes rotate about their anteroposterior (fore-and-aft) axes in the same sense, occurs naturally; for example, when the head tips toward one shoulder the eyes tend to roll in the opposite direction, with the result that the image of the visual field on the retina tends to remain vertical in spite of the rotation of the head.

Conjugate and disjunctive movements

Nervous control. The nerves controlling the actions of the muscles are the third, fourth, and sixth cranial nerves, with their bodies (nuclei) in the brainstem; the third, or oculomotor nerve, controls the superior and inferior recti, the medial rectus, and inferior oblique; the fourth cranial nerve, the trochlear nerve, controls the superior oblique; and the sixth, the abducens nerve, controls the lateral rectus. The nuclei of these nerves are closely associated; especially, there are connections between the nuclei of the sixth cranial nerve, controlling the lateral rectus, and the nucleus of the third, controlling the medial rectus; it is through this close relationship that the linking of the lateral rectus of one eye and the medial rectus of the other, indicated above, is achieved. Another type of linking is concerned with reciprocal inhibition; that is, when there are two antagonistic muscles, such as the medial and the lateral rectus, contraction of one is accompanied by a simultaneous inhibition of the other. Muscles show a continuous slight activity even when at rest; this keeps them taut; this action, called tonic activity, is brought about by discharges in the motor nerve to the muscle. Hence, when the agonist muscle contracts its antagonist must be inhibited.

Reflex pathways. In examining any reflex movement one must look for the sensory input—*i.e.,* the way in which messages in sensory nerves bring about discharges in the motor nerves to the muscles; this study involves the connections of the motor nerves or nuclei with other centres of the brain.

When a subject is looking straight ahead and a bright light appears in the periphery of his field of vision, his eyes automatically turn to fix on the light; this is called the fixation reflex. The sensory pathway in the reflex arc leads as far as the cerebral cortex because removal of the occipital cortex (the outer brain substance at the back of the head) abolishes reflex eye movements in response to light stimuli. If the occipital cortex is stimulated electrically, movements of the eyes may be induced, and in fact one may draw a pattern of the visual field on the occipital cortex corresponding with the directions in which the gaze is turned when given points on the cortex are stimulated. This pattern corresponds with the pattern obtained by recording the visual responses to light stimuli from different parts of the visual field.

Fixation reflex

The remainder of the pathway—*i.e.,* from the occipital cortex to the motor neurons in the brainstem—has long been considered to involve the superior colliculi as relay stations, and they certainly have such a role in lower animals; but in human beings a pathway from the cortex to the eye-muscle nuclei independent of the superior colliculi of the midbrain is now generally assumed.

Continual movements of the eyes occur even when an effort is made to maintain steady fixation of an object. Some of these movements may be regarded as manifestations of the fixation reflex; thus, the eyes tend to drift off their target, and, because of this, the fixation reflex comes into play, bringing the eyes back on target.

Nystagmus

Experimentally, the fixation reflex can be studied by observation of the regular to-and-fro movements of the eyes as they follow a rotating drum striped in black and white. (Such movements of the eyes directed at a moving object are called optokinetic nystagmus; nystagmus itself is the involuntary movement of the eye back and forth, up and down, or in a rotatory or a mixed fashion.) While the eyes watch the moving drum, they involuntarily make a slow movement as a result of fixing their gaze on a particular stripe. At a certain point, fixation is broken off, and the eyes spring back to fix on a new stripe. Thus, the nystagmus consists of a slow movement with angular velocity equal to that of the rotation of the drum, then a fast saccade, or jump from one point of fixation to another, in the opposite direction; the process is repeated indefinitely.

Another type of nystagmus reveals the play of another set of reflexes. These are mediated by the semicircular canals—*i.e.,* the organs of balance or the vestibular apparatus. Such a reflex may be evoked by rotating the subject in a chair at a steady speed; the eyes move slowly in the opposite direction to that of rotation and, at the end of their excursion, jump back with a fast saccade in the direction of rotation. If rotation suddenly ceases, the eyes go into a nystagmus in the opposite direction, the postrotatory nystagmus.

During rotation, certain semicircular canals are being stimulated, and the important point is that any acceleration of the head that stimulates these canals will cause reflex movements of the eyes; thus, acceleration of the head to the right causes a movement of the eyes to the left, the function of the reflex being to enable the eyes to maintain steady fixation of an object despite movements of the head. The reflex occurs even when the eyes are shut, and, when the eyes are open, it obviously cooperates with the fixation reflex in maintaining steady fixation. In many lower animals this connection between organs of balance and eyes is very rigid; thus, one may move the tail of a fish, and its eyes will move reflexly. In man, not only do the semicircular canals function in close relation to the eye muscles but so also do the gravity organ—the utricle—and the stretch receptors in the muscles of the neck. Thus, when the head is turned upward, there is a reflex tendency for the eyes to move downward, even if the eyes are shut. The actual movement is probably initiated by the reflex from the semicircular canals, which respond to acceleration, but the maintenance of the position is brought about by a reflex through the stretch of the neck muscles and also through the pull of gravity on the utricle, or otolith organ, in the inner ear.

Voluntary centre. The eyes are under voluntary control, and it is thought that the cortical area subserving voluntary eye movements is in the frontal cortex. Stimulation of this in primates causes movements of the eyes that are well coordinated, and a movement induced by this region prevails over one induced by stimulation of the occipital cortex. The existence of a separate centre in man is revealed by certain neurological disorders in which the subject is unable to fixate voluntarily but can do so reflexly; *i.e.,* he can follow a moving light.

The nature of eye movements. So far, the relation of the movements of the eyes to the requirements of the visual apparatus and their control have been touched upon. To examine the character of the movements in some detail requires rapid, accurate measurement of the movements that the eyes undergo. Modern studies of this subject employ a contact lens fitting on to the globe; on the lens is a small plane mirror, and a parallel bundle of rays is reflected off this mirror onto a moving film.

By the use of refined methods of measuring the position of the eyes at any moment, it becomes immediately evident that the eyes are never stationary for more than a fraction of a second; the movements are of three types: (1) irregular movements of high frequency (30–70 per second) and small excursions of about 20 seconds of arc; (2) flicks, or saccades, of several minutes of arc occurring at regular intervals of about one second; and between these saccades there occur (3) slow irregular drifts extending up to six minutes of arc. The saccades are corrective, serving to bring the fixation axis on the point of regard after this has drifted away from it too far, and thus are a manifestation of the fixation reflex.

The significance of these small movements during fixation was revealed by studies on the stabilized retinal image: by a suitable optical device the image of an object could be held stationary on the retina in spite of the movements of the eye. It was found that under these conditions the image would disappear within a few seconds. Thus, the movements of the eye are apparently necessary to allow the contours of the image to fall on a new set of rods and cones at repeated intervals; if this does not occur, the retina adapts to their stimulus and ceases to send messages to the central nervous system. The small flicks mentioned above are essentially the same as the larger movement made when the two eyes fixate (fix on) a light when it suddenly appears in the peripheral field; this is given the general name of the saccade, to distinguish it from the slower movements occurring during convergence and smooth following. The dynamics of the saccade have been studied in some detail by several workers. There is a reaction time of about 120 to 180 milliseconds, after which both eyes move simultaneously; there is a definite overshoot and, with an excursion of 20°, the operation is completed in about 90 milliseconds. The maximum velocity increases with the extent of the movement, being 300° per second for 10° and 500° per second for 30°. A remarkable feature is the apparent absence of significant inertia in the eyeball, so that movement is halted, not by any checking action of antagonistic muscles but simply by cessation of contraction of the agonists; thus, the movement is not ballistic. Once under way, the saccade is determined in amount, so that the subject cannot voluntarily alter its direction and extent. The control mechanism for the saccadic type of movement can be described as a sampled data system, *i.e.,* the brain makes discontinuous samples of the position of the eyes in relation to the target and corrects the error, in contrast to a continuous feedback system that takes account of the error all the time.

The saccade, disjunctive movements, and tracking

The movements of the eyes when they converge onto a near object are in remarkable contrast to the saccade; the angular velocity is only about 25° per second, compared with values as high as 500° per second in the saccade. The great difference in speed suggested to two investigators that the two movements are executed by different muscle fibres. In fact, the extraocular muscles do contain two types of muscle fibre with characteristically different nerve supplies, and some recent studies tend to support this view of a dual mechanism.

If a moving light suddenly appears in the field of view, and if its rate of movement is less than about 30° per second, the response of the eyes is remarkably efficient; a saccade brings the eyes on target, and they follow the motion at almost exactly the same angular velocity as that of the target; inaccuracies in following lead to corrective saccades. When the rate of movement of the target is greater than about 30° per second, these corrective saccades become more obvious because now smooth following is not possible; the eyes make constant-velocity movements, but the velocity rarely matches that of the moving target, so that there must be frequent corrective saccades. Studies have shown that the following movements are highly integrated and must involve a continuous feedback system whereby errors are used to modify the performance. Thus, the systems for control of saccades and tracking movements are fundamentally different.

Vision suppression during a saccade. If one looks into a mirror and fixates one of one's eyes and then fixates the other, one does not see the eyes moving; and it has been argued that, during an eye movement, vision is suppressed; if vision were not suppressed, moreover, it seems likely that the images of the external world would appear smeared during a movement. Experimental studies have shown that there is, indeed, a suppression of vision during a saccade.

THE WORK OF THE OPTICAL LENS SYSTEM

Refraction by cornea and lens. The optical system of the eye is such as to produce a reduced inverted image of the visual field on the retina; the system behaves as a convex lens but is, in fact, much more complex, refraction taking place not at two surfaces, as in a lens, but at four separate surfaces—at the anterior and the posterior surfaces of the cornea and of the crystalline lens. Each of these surfaces is approximately spherical, and at each optical interface—*e.g.,* between air and the anterior surface of the cornea—the bending of a ray of light is toward the axis, so that, in effect, there are four surfaces tending to make rays of light converge on each other. If the rays of light falling on the cornea are parallel—*i.e.,* if they come from a distant point—the net effect of this series of refractions at the four surfaces is to bring these rays to a point focus of the optical system, which in the normal, or emmetropic, eye corresponds with the retina. The greatest change of direction, or bending of the rays, occurs where the difference of refractive index is greatest, and this is when light passes from air into the cornea, the refractive index of the corneal substance being 1.3376; the refractive indices of the cornea and aqueous humour are not greatly different, that of the aqueous humour being 1.336 (as is that of the vitreous); thus, the bending, as the rays meet the concave posterior surface of the cornea and emerge into a medium of slightly less refractive index, is small. The lens has a greater refractive index than that of its surrounding aqueous humour and vitreous body, 1.386 to 1.406, so that its two surfaces contribute to convergence, the posterior surface normally more than the anterior surface because of its greater curvature (smaller radius).

Four refractive surfaces

Normal sightedness and near- and farsightedness. In contrast to the focussing of the normal (emmetropic) eye, in which the image of the visual field is focussed on the retina, the image may be focussed in front of the retina (nearsightedness, or myopia), or behind the retina (farsightedness or hyperopia). In myopia the vision of distant objects is not distinct because the image of a distant point falls within the vitreous and the rays spread out to form a blur circle on the retina instead of a point. In this condition the eye is said to have too great dioptric (refractive) power for its length. When the focus falls behind the retina, the image of the distant point is again a circle on the retina; and the farsighted eye is said to have too little dioptric power. The important point to appreciate is that emmetropia, or normal sight, requires that the focal power of the dioptric system be matched to the axial length of the eye; it certainly is remarkable that emmetropia is indeed the most common condition when it is appreciated that just one millimetre of error in the matching of axial length with focal length would cause a person to require a spectacle correction. In general, however, the effects of variations in dimensions tend to compensate each other. Thus, for example, an unusually large eye might, at first thought, be expected to be myopic, but a large eye tends to be associated with a large radius of curvature of the cornea, and this would reduce the power—*i.e.,* increase the focal length—and so an unusually large eye is not necessarily a myopic one.

Accommodation. *Effects of accommodation.* The image of an object brought close to the eye would be formed behind the retina if there were no change in the focal length of the eye. This change to bring the image of an object upon the retina is called accommodation. The point nearer than which accommodation is no longer effective is called the near point of accommodation. In very young people, the near point of accommodation is quite close to the eye, namely about seven centimetres (about three inches) in front at 10 years old; at 40 years the distance has increased to about 16 centimetres (about 6 inches), and at 60 years it is 100 centimetres or one metre (39 inches). Thus, a 60-year-old would not be able to read a book held at the convenient distance of about 40 centimetres (16 inches), and the extra power required would have to be provided by convex lenses in front of the eye, an arrangement called the presbyopic correction.

The near point of accommodation

Mechanism of accommodation. It is essentially an increase in curvature of the anterior surface of the lens that is responsible for the increase in power involved in the process of accommodation. A clue to the way in which this change in shape takes place is given by the observation that a lens that has been taken out of the eye is much rounder and fatter than one within the eye; thus, its attachments by the zonular fibres to the ciliary muscle within the eye preserve the unaccommodated or flattened state of the lens; and modern investigations leave little doubt that it is the pull of the zonular fibres on the elastic capsule of the lens that holds the anterior surface relatively flat. When these zonular fibres are loosened, the elastic tension in the capsule comes into play and remolds the lens, making it smaller and thicker. Thus, the physiological problem is to find what loosens the zonular fibres during accommodation. The ciliary muscle has been described earlier, and it has been shown that the effect of contracting its fibres is, in general, to pull the whole ciliary body forward and to move the anterior region toward the axis of the eye by virtue of the sphincter action of the circular fibres. Both of these actions will slacken the zonular fibres and therefore allow the change in shape. As to why it is the anterior surface that changes most is not absolutely clear, but it is probably a characteristic of the capsule rather than of the underlying lens tissue. Defective accommodation in presbyopia is not due to a failure of the ciliary muscle but rather to a hardening of the substance of the lens with age to the point that readjustments of its shape become ever more difficult.

Nerve action. Accommodation is an involuntary reflex act, and the ciliary muscle belongs to the smooth involuntary class. Appropriate to this, the innervation is through the autonomic system, the parasympathetic nerve cells belonging to the oculomotor nerve (the third cranial nerve) occupying a special region of the nucleus in the midbrain called the Edinger-Westphal nucleus; the fibres have a relay point in the ciliary ganglion in the eye socket, and the postganglionic fibres enter the eye as the short ciliary nerves. The stimulus for accommodation is the nearness of the object, but the manner in which this nearness is translated into a stimulus is not clear. Thus, the fact that the image is blurred is not sufficient to induce accommodation; the eye has some power of discriminating whether the blurredness is due to an object being too far away or too close, so that something more than mere blurredness is required.

The pupil. The amount of light entering the eye is restricted by the aperture in the iris, the pupil.

When a person is in a dark room his pupil is large, perhaps eight millimetres (0.3 inch) in diameter, or more. When the room is lighted there is an immediate constriction of the pupil, the light reflex; this is bilateral, so that even if only one eye is exposed to the light both pupils contract to nearly the same extent. After a time the pupils expand even though the bright light is maintained, but the expansion is not large. The final state is determined by the actual degree of illumination; if this is high, then the final state may be a diameter of only about three to four millimetres (about 0.15 inch); if it is not so high, then the initial constriction may be nearly the same, but the final state may be with a pupil of four to five millimetres (about 0.18 inch). During this steady condition, the pupils do not remain at exactly constant size; there is a characteristic oscillation in size that, if exaggerated, is called hippus.

The light and near reflexes

A pupillary constriction will also occur when a person looks at a near object—the near reflex. Thus, accommodation and pupillary constriction occur together reflexly and are excited by the same stimulus. The function of the pupil is clearly that of controlling the amount of light entering the eye, and hence the light reflex. The constriction occurring during near vision suggests other functions, too; thus, the aberrations of the eye (failure of some refracted rays to focus on the retina) are decreased by reducing the aperture of its optical system. In the dark, aberrations are of negligible significance, so that a person is concerned only with allowing as much light into the eye as possible; in bright light high visual acuity is usually required, and this means reducing the aberrations. The depth of focus of the optical system is increased when the aperture is reduced, and the near reflex is probably concerned with increasing depth of focus under these conditions.

Dilation of the pupil occurs as a result of strong psychical stimuli and also when any sensory nerve is stimulated; dilation thus occurs in extreme fear and in pain.

Neuromuscular mechanisms. The muscles of the iris have been described earlier. It is clear from their general features that constriction of the pupil is brought about by shortening of the circular ring of fibres—the sphincter; dilation is brought about by shortening of the radially oriented fibres. The sphincter is innervated by parasympathetic fibres of the oculomotor nerve, with their cell bodies in the Edinger-Westphal nucleus, as are the nerve cells controlling accommodation; thus, the close association between the accommodation and pupillary reflexes is reflected in a close anatomical contiguity of their motor nerve cells.

The sensory pathway in the light reflex involves the rods and cones, bipolar cells, and ganglion cells. As indicated earlier, a relay centre for pupillary responses to light is the pretectal nucleus in the midbrain. There is a partial crossing-over of the fibres of the pretectal nerve cells so that some may run to the motor nerve cells in the Edinger-Westphal nucleus of both sides of the brain, and it is by this means that illumination of one eye affects the other. The Edinger-Westphal motor neurons have a relay point in the ciliary ganglion, a group of nerve cells in the eye socket, so that its electrical stimulation causes both accommodation and pupillary constriction; similarly, application of a drug, such as pilocarpine, to the cornea will cause a constriction of the pupil and also a spasm of accommodation; atropine, by paralyzing the nerve supply, causes dilation of the pupil and paralysis of accommodation (cycloplegia).

Dilator response

The dilator muscle of the iris is activated by sympathetic nerve fibres. Stimulation of the sympathetic nerve in the neck causes a powerful dilation of the iris; again, the influx of adrenalin into the blood from the adrenal glands during extreme excitement results in pupillary dilation.

Many involuntary muscles receive a double innervation, being activated by one type of nerve supply and inhibited by the other; modern experimentation indicates that the iris muscles are no exception, so that the sphincter has an inhibitory sympathetic nerve supply, while the dilator has a parasympathetic (cholinergic) inhibitor. Thus, a drug like pilocarpine not only activates the constrictor muscle but actively inhibits the dilator. A similar double innervation has been described for the ciliary muscle. In general, any change in pupillary size results from a reciprocal innervation of dilator and constrictor; thus, activation of the constrictor is associated with inhibition of the dilator and vice versa.

The near response. In general, as has been indicated, pupillary constriction and accommodation occur together, in response to the same stimulus; a third element in this near response is, of course, the convergence (turning in) of the eyes, mediated by voluntary muscles, the medial recti. Experimentally, it is often possible to separate these activities, in the sense that one may cause convergence without accommodation by placing appropriate prisms in front of the eyes; or one may cause accommodation without convergence by placing diverging lenses in front of the eyes. There are many experiments that show that accommodation and convergence are neurologically linked to some extent, however.

THE WORK OF THE RETINA

Some basic facts of vision. So far, attention has been directed to what are essentially the preliminaries to vision; it is now time to examine some of the elementary facts of vision and to relate them to the structure of the retina and, later, to chemically identifiable events.

Measurement of the threshold. An important means of measuring a sensation is to determine the threshold stimulus—*i.e.,* the minimum energy required to evoke the sensation. In the case of vision, this would be the minimum number of quanta of light entering the eye in unit time. If it is found that the threshold has altered because of a change of some sort, then this change can be said to have altered the subject's sensitivity to light, and a numerical value can be assigned to the sensitivity by use of the reciprocal of the threshold energy. Practically, a subject may be placed in the dark in front of a white screen, and the screen may be illuminated by flashes of light; for any given intensity of illumination of the screen, it is not difficult to calculate the flow of light energy entering the eye. One may begin with a low intensity of flash and increase this successively until the subject reports that he can see the flash. In fact, at this threshold level, he will not see every flash presented, even though the intensity of the light is kept constant; for this reason, a certain frequency of seeing—*e.g.,* four times out of six—must be selected as the arbitrary point at which to fix the threshold.

Decline in visual threshold

When measurements of this sort are carried out, it is found that the threshold falls progressively as the subject is maintained in the dark room. This is not due to dilation of the pupil because the same phenomenon occurs if the subject is made to look through an artificial pupil of fixed diameter. The eye, after about 30 minutes in the dark, may become about 10,000 times more sensitive to light. Vision under these conditions is, moreover, characteristically different from what it is under ordinary daylight conditions. Thus, in order to obtain best vision, the eye must look away from the screen so that the image of the screen does not fall on the fovea; if the screen is continuously illuminated at around this threshold level it will be found to disappear if its image is brought onto the fovea, and it will become immediately visible on looking away. The same phenomenon may be demonstrated on a moonless night if the gaze is fixed on a dim star; it disappears on fixation and reappears on looking away. This feature of vision under these near-threshold or scotopic conditions suggests that the cones are effectively blind to weak light stimuli, since they are the only receptors in the fovea.

Duplicity theory of vision

This is the basis of the duplicity theory of vision, which postulates that when the light stimulus is weak and the eye has been dark-adapted, it is the rods that are utilized because, under these conditions, their threshold is much lower than that of the cones. When the subject first enters the dark, the rods are the less sensitive type of receptor, and the threshold stimulus is the light energy required to stimulate the cones; during the first five or more minutes the threshold of the cones decreases; *i.e.,* they become more sensitive. The rods then increase their sensitivity to the point that they are the more sensitive, and it is they that now determine the sensitivity of the whole eye, the threshold stimuli obtained after 10 minutes in the dark, for example, being too weak to activate the cones.

Scotopic sensitivity curve. When different wavelengths of light are employed for measuring the threshold, it is found, for example, that the eye is much more sensitive to blue-green light than to orange. The interesting feature of this kind of study is that the subject reports only that the light is light; he distinguishes no colour. If the intensity of a given wavelength of light is increased step by step above the threshold, a point comes when the subject states that it is coloured, and the difference between the threshold for light appreciation and this, the chromatic threshold, is called the photochromatic interval. This suggests that the rods give only achromatic, or colourless, vision, and that it is the cones that permit wavelength discrimination. The photochromatic interval for long wavelengths (red light) is about zero, which means that the intensity required to reach the sensation of light is the same as that to reach the sensation of colour. This is because the rods are so insensitive to red light; if the dark-adaptation curve is plotted for a red stimulus it is found that it follows the cone path, like that for foveal vision at all wavelengths.

Loss of dark adaptation. If, when the subject has become completely dark-adapted, one eye is held shut and the other exposed to a bright light for a little while, it is found that, whereas the dark-adapted eye retains its high sensitivity, that of the light-exposed eye has decreased greatly; it requires another period of dark adaptation for the two eyes to become equally sensitive.

These simple experiments pose several problems, the answers to which throw a great deal of light on the whole mechanism of vision. Why, for example, does it require time for both rods and cones to reach their maximum sensitivity in the dark? Again, why is visual acuity so low under scotopic conditions compared with that in daylight, although sensitivity to light is so high? Finally, why do the rods not serve to discriminate different wavelengths?

Bleaching of rhodopsin. It may be assumed that a receptor is sensitive to light because it contains a substance that absorbs light and converts this vibrational type of energy into some other form that is eventually transmuted into electrical changes, and that these may be transmitted from the receptor to the bipolar cell with which it is immediately connected. When the retina of a dark-adapted animal is removed and submitted to extraction procedures, a pigment, originally called visual purple but now called rhodopsin, may be obtained. If the eye is exposed to a bright light for some time before extraction, little or no rhodopsin is obtained. When retinas from animals that had been progressively dark-adapted were studied, a gradual increase in the amount of rhodopsin that could be extracted was observed. Thus, rhodopsin, on absorption of light energy, is changed to some other compound, but new rhodopsin is formed, or rhodopsin is regenerated, during dark adaptation. The obvious inference is that rhodopsin is the visual pigment of the rods, and that when it is exposed to relatively intense lights it becomes useless for vision. When the eye is allowed to remain in the dark the rhodopsin regenerates and thus becomes available for vision. There is now conclusive proof that rhodopsin is, indeed, the visual pigment for the rods; it is obtained from retinas that have only rods and no cones—*e.g.,* the retinas of the rat or guinea pig, and it is not obtained from the pure cone retina of the chicken.

Rhodopsin as the photopigment

When the absorption spectrum is measured, it is found that its maximum absorption occurs at the point of maximum sensitivity of the dark-adapted eye. Similar measurements may be carried out on animals, but the threshold sensitivity must be determined by some objective means—*e.g.,* the response of the pupil, or, better still, the electrical changes occurring in the retina in response to light stimuli. Thus, the electroretinogram (ERG) is the record of changes in potential between an electrode placed on the surface of the cornea and an electrode placed on another part of the body, caused by illumination of the eye.

The high sensitivity of the rods by comparison with the cones may be a reflection of the greater concentration in them of pigment that would permit them to catch light more efficiently, or it may depend on other factors—*e.g.,* the efficiency of transformation of the light energy into electrical energy. The pigments responsible for cone vision are not easily extracted or identified, and the problem will be considered in the material on colour vision. An important factor, so far as sensitivity is concerned, is the actual organization of the receptors and neurons in the retina.

Synaptic organization of the retina. The basic structure of the retina has been indicated earlier. As in other parts of the nervous system, the messages initiated in one element are transmitted, or relayed, to others. The regions of transmission from one cell to another are areas of intimate contact known as synapses. An impulse conveyed from one cell to another travels from the first cell body along a projection called an axon, to a synapse, where the impulse is received by a projection, called a dendrite, of the second cell. The impulse is then conveyed to the second cell body, to be transmitted further, along the second cell's axon.

It will be recalled that the functioning cells of the retina are the receptor cells—the rods and cones; the ganglion cells, the axons of which form the optic nerve; and cells that act in a variety of ways as intermediaries between the receptors and the ganglion cells. These intermediaries are named bipolar cells, horizontal cells, and amacrine cells.

Plexiform layers. As was indicated earlier, the synapses occur in definite layers, the outer and inner plexiform layers. In the outer plexiform layer the bipolar cells make their contacts, by way of their dendrites, with the rods and cones, specifically the spherules of the rods and the pedicles of the cones. In this layer, too, the projections from horizontal cells make contacts with rods, cones, and bipolar cells, giving rise to a horizontal transmission and thereby allowing activity in one part of the retina to influence the behaviour of a neighbouring part. In the inner plexiform layer, the axons of the bipolar cells make connection with the dendrites of ganglion cells, once again at special synaptic regions. (The dendrites of a nerve cell carry impulses to the nerve cell; its axon, away from the cell.) Here, too, a horizontal interconnection between bipolar cells is brought about, in this case by way of the axons and dendrites of amacrine cells.

Inner plexiform layer

The bipolar cells are of two main types: namely, those that apparently make connection with only one receptor—a cone—and those that connect to several receptors. The type of bipolar cell that connects to a single cone is called the midget bipolar. The other type of bipolar cell is called diffuse; varieties of these include the rod bipolar, the dendritic projections of which spread over an area wide enough to allow contacts with as many as 50 rods; and the flat cone bipolar, which collects messages from up to seven cones.

Ganglion cells are of two main types: namely, the midget ganglion cell, which apparently makes a unique connection with a midget bipolar cell, which in turn is directly connected to a single cone; and a diffuse type, which collects messages from groups of bipolar cells.

Convergence of the messages. The presence of diffuse bipolar and ganglion cells collecting messages from groups of receptors and bipolar cells, and, what may be even more important, the presence of lateral connections of groups of receptors and bipolar cells through the horizontal and amacrine cells, means that messages from receptors over a rather large area of the retina may converge on a single ganglion cell. This convergence means that the effects of light falling on the receptive field may be cumulative, so that a weak light stimulus spread over about 1,000 rods is just as effective as a stronger stimulus spread over 100 or less; in other words, a large receptive field will have a lower threshold than a small one; and this is, in fact, the basis for the high sensitivity of the area immediately outside the fovea, where there is a high density of rods that converge on single bipolar cells. Thus, if it is postulated that the cones do not converge to anything like the same extent as the rods, the greater sensitivity of the latter may be explained; and the anatomical evidence favours this postulate.

It has been indicated above that the regeneration of visual pigment is a cause of the increased sensitivity of the rods that occurs during dark adaptation. This, apparently, is only part of the story. An important additional factor is the change in functional organization of the retina during adaptation. When the eye is light-adapted, functional convergence is small, and sensitivity of rods and cones is low; as dark adaptation proceeds, convergence of rods increases. The anatomical connections do not change, but the power of the bipolar cells and ganglion cells to collect impulses is increased, perhaps by the removal of an inhibition that prevents this during high illumination of the retina.

Absolute threshold and minimum stimulus for vision. As was indicated earlier, the threshold is best indicated in terms of frequency of seeing since, because of fluctuations in the threshold, there is no definite luminance of a test screen at which it is always seen by the observer, and there is no luminance just below this at which it is never seen. Experiments, in which 60 percent was arbitrarily taken as the frequency of seeing and in which the image of a patch of light covered an area of retina containing about 20,000,000 rods, led to the calculation that the mean threshold stimulus represents 2,500 quanta of light that is actually absorbed per square centimetre of retina. This calculation leads to two important conclusions: namely, that at the threshold only one rod out of thousands comes into operation, and that during the application of a short stimulus the chances are that no rod receives more than a single quantum.

A quantum, defined as the product of Planck's constant (6.63×10^{-27} erg-second) times the frequency of light, is the minimum amount of light energy that can be employed. A rod excited by a single quantum cannot excite a bipolar cell without the simultaneous assistance of one or more other rods. Experiments carried out in the 1940s indicated that a stimulus of about 11 quanta is required; thus it may require 11 excited rods, each receiving one quantum of light, to produce the sensation of light.

Uncertainty principle

Quantum fluctuations. With such small amounts of energy as those involved in the threshold stimulus, the uncertainty principle becomes important; according to this, there is no certainty that a given flash will have the expected number of quanta in it, but only a probability. Thus, one may speak of a certain average number of quanta and the actual number in any given flash, and one may compute on statistical grounds the shape of curve that is obtained by plotting frequency with which a flash contains, say, four quanta or more against the average number in the flash. One may also plot the frequency with which a flash is seen against the average number of quanta in the flash, and this frequency-of-seeing curve turns out to be similar to the frequency-of-containing-quanta curve when the number of quanta chosen is five to seven, depending on the observer. This congruence strongly suggests that the fluctuations in response to a flash of the same average intensity are caused by fluctuations in the energy content of the stimulus, and not by fluctuations in the sensitivity of the retina.

Spatial summation. In spatial summation two stimuli falling on nearby areas of the retina add their effects so that either alone may be inadequate to evoke the sensation of light, but, when presented simultaneously, they may do so. Thus, the threshold luminance of a test patch required to be just visible depends, within limits, on its size, a larger patch requiring a lower luminance, and vice versa. Within a small range of limiting area, namely that subtending about 10 to 15 minutes of arc, the relationship called Ricco's law holds; *i.e.,* threshold intensity multiplied by the area equals a constant. This means that over this area, which embraces several hundreds of rods, light falling on the individual rods summates, or accumulates, its effects completely so that 100 quanta falling on a single rod are as effective as one quantum falling simultaneously on 100 rods. The basis for this summation is clearly the convergence of receptors on ganglion cells, the chemical effects of the quanta of light falling on individual rods being converted into electrical changes that converge on a single bipolar cell through its branching dendritic processes. Again, the electrical effects induced in the bipolar cells may summate at the dendritic processes of a ganglion cell so that the receptive field of a ganglion cell may embrace many thousands of rods.

Temporal summation. In temporal summation, two stimuli, each being too weak to excite, cause a sensation of light if presented in rapid succession on the same spot of the retina; thus, over a certain range of times, up to 0.1 second, the Bunsen-Roscoe law holds: namely, that the intensity of light multiplied by the time of exposure equals a constant. Thus it was found that within this time interval (up to 0.1 second), the total number of quanta required to excite vision was 130, irrespective of the manner in which these were supplied. Beyond this time, summation was still evident, but it was not perfect, so that if the duration was increased to one second the total number of quanta required was 220. Temporal summation is consistent with quantum theory; it has been shown that fluctuations in the number of quanta actually in a light flash are responsible for the variable responsiveness of the eye; increasing the duration of a light stimulus increases the probability that it will contain a given number of quanta, and that it will excite.

Inhibition. In the central nervous system generally, the relay of impulses from one nerve cell or neuron to excite another is only one aspect of neuronal interaction. Just as important, if not more so, is the inhibition of one neuron by the discharge in another. So it is in the retina. Subjectively, the inhibitory activity is reflected in many of the phenomena associated with adaptation to light or its reverse. Thus, the decrease in sensitivity of the retina to light during exposure to light is only partially accounted for by bleaching of visual pigment, be it the pigment in rod or cone; an important factor is the onset of inhibitory processes that reduce the convergence of receptors on ganglion cells. Some of the rapidly occurring changes in sensitivity described as alpha adaptation are doubtless purely neural in origin.

Many so-called inductive phenomena indicate inhibitory processes; thus, the phenomenon of simultaneous contrast, whereby a patch of light appears much darker if surrounded by a bright background than by a black, is due to the inhibitory effect of the surrounding retina on the central region, induced by the bright surrounding. Many colour-contrast phenomena are similarly caused; thus, if a blue light is projected onto a large white screen, the white screen rapidly appears yellow; the blue stimulus falling on the central retina causes inhibition of blue sensitivity in the periphery; hence, the white background will appear to be missing its blue light—white minus blue is a mixture of red and green—*i.e.,* yellow. Particularly interesting from this viewpoint are the phenomena of metacontrast; by this is meant the inductive effect of a primary light stimulus on the sensitivity of the eye to a previously presented light stimulus on an adjoining area of retina. It is a combination of temporal and spatial induction. The effect is produced by illuminating the two halves of a circular patch consecutively for a brief duration. If the left half only, for example, is illuminated for 10 milliseconds it produces a definite sensation of brightness. If, now, both halves are illuminated for the same period, but the right half from 20 to 50 milliseconds later, the left half of the field appears much darker than before and, near the centre, may be completely extinguished. The left field has thus been inhibited by the succeeding, nearby, stimulus. The right field, moreover, appears darker than when illuminated alone—it has been inhibited by the earlier stimulus (paracontrast).

Meta- and paracontrast

Flicker. Another visual phenomenon that brings out the importance of inhibition is the sensation evoked when a visual stimulus is repeated rapidly; for example, one may view a screen that is illuminated by a source of light the rays from which may be intercepted at regular intervals by rotating a sector of a circular screen in front of it. If the sector rotates slowly, a sensation of black followed by white is aroused; as the speed increases the sensation becomes one of flicker—*i.e.,* rapid fluctuations in brightness; finally, at a certain speed, called the critical fusion frequency, the sensation becomes continuous and the subject is unaware of the alterations in the illumination of the screen.

At high levels of luminance, when cone vision is employed, the fusion frequency is high, increasing with increasing luminance in a logarithmic fashion—the Ferry-Porter law—so that at high levels it may require 60 flashes per second to reach a continuous sensation. Under conditions of night, or scotopic, vision, the frequencies may be as low as four per second. The difference between rod and cone vision in this respect probably resides in the power of the eye to inhibit activity in cones rapidly, so that the sensation evoked by a single flash is cut off immediately, and this leaves the eye ready to respond to the next stimulus. By contrast, the response in the rod lasts so much longer that, when a new stimulus falls even a quarter of a second later, the difference in the state of the rods is insufficient to evoke a change in intensity of sensation; it merely prolongs it. One interesting feature of an intermittent stimulus is that the intensity of the sensation of brightness, when fusion is achieved, is dependent on the relative periods of light and darkness in the cycle, and this gives one a method of grading the effective luminance of a screen; one may keep the intensity of the illuminating source constant and merely vary the period of blackness in a cycle of black and white. The effective luminance will be the average luminance during a cycle; this is known as the Talbot-Plateau law.

Visual acuity. As has been stated, the ability to perceive detail is restricted in the dark-adapted retina when the illumination is such as to excite only the scotopic type of vision; this is in spite of the high sensitivity of the retina to light under the same conditions. The power of distinguishing detail is essentially the power to resolve two stimuli separated in space, so that, if a grating of black lines on a white background is moved farther and farther away from an observer, a point is reached when he will be unable to distinguish this stimulus pattern from a uniformly gray sheet of paper. The angle subtended at the eye by the spacing between the lines at the point where they are just resolvable is called the resolving power of the eye; the reciprocal of this angle, in minutes of arc, is called the visual acuity. Thus, a visual acuity of unity indicates a power of resolving detail subtending one minute of arc at the eye; a visual acuity of two indicates a resolution of one-half minute, or 30 seconds of arc. The visual acuity depends strongly on the illumination of the test target, and this is true of both daylight (photopic) and night (scotopic) vision; thus, with a brightly illuminated target, with the surroundings equally brightly illuminated (the ideal condition), the visual acuity may be as high as two. When the illumination is reduced, the acuity falls so that, under ordinary conditions of daylight viewing, visual acuity is not much better than unity. Under scotopic conditions, the visual acuity may be only 0.04 so that lines would have to subtend about 25 minutes at the eye to be resolvable; this corresponds to a thickness of 4.4 centimetres (1.7 inches) at a distance of six metres (20 feet).

Resolving power and visual acuity

Measurement. In the laboratory, visual acuity is measured by the Landolt C, which is a circle with a break in it. The subject is asked to state where the break is when the figure is rotated to successive random positions. The size of the *C*, and thus of its break, is reduced until the subject makes more than an arbitrarily chosen percentage of mistakes. The angle subtended at the eye by the break in the *C* at this limit is taken as the resolving power of the eye. The testing of the eyes by the ophthalmologist or optometrist is essentially a determination of visual acuity; here the subject is presented with the Snellen chart, rows of letters whose details subtend progressively smaller angles at the eye. The row in which, say, five out of six letters are seen correctly is chosen as that which measures the visual acuity. If the details subtended one minute of arc, the visual acuity would be unity. The notation employed is somewhat obscure; a visual acuity of unity would be expressed as 6/6; an acuity of a half as 6/12, and so on; here the numerator is the viewing distance in metres from the chart and the denominator the distance at which details on the letters of the limiting row subtend one minute of arc at the eye.

The Snellen chart

Anatomical basis; the retinal mosaic. From an anatomical point of view one may expect the limit to resolving power to be imposed by the "grain" of the retinal mosaic in the same way that the size of the grains in a photographic emulsion imposes a limit to the accuracy with which detail may be photographed. Two white lines on a black ground, for example, could not be appreciated as distinct if their images fell on the same or adjacent sets of receptors. If a set of receptors intervened between the stimulated ones, there would be a basis for discrimination because the message sent to the central nervous system could be that two rows of receptors, separated by an unstimulated row, were sending messages to their bipolar cells. Thus, the limit to resolution, on this basis, should be the diameter of a foveal cone, or rather the angle subtended by this at the nodal point of the eye; this is about 30 seconds of arc and, in fact, corresponds with the best visual acuity attainable. If this grain of the retinal mosaic is to be the basis of resolution, however, one must postulate, in addition, a nervous mechanism that will transmit accurately the events taking place in the individual receptors, in this case the foveal cones; *i.e.*, there must be a one-to-one relationship between cones, bipolar cells, ganglion cells, and lateral geniculate cells so that what is called the local sign of the impulses from a given foveal cone may be obtained. It must be appreciated that restriction on convergence (or its reverse, spread) of messages may be achieved by inhibition; the anatomical connections may be there, but they may be made functionally inoperative by inhibition exerted by other neurons; thus, the horizontal and amacrine cells might well exert a restraining influence on certain junctions, thereby reducing the spread, or convergence, of messages, and it seems likely that the improvement in foveal visual acuity from one to two, brought about by increased luminance of the target and its surroundings, is achieved by an increase in inhibition that tends to make transmission one-to-one in the fovea.

It must be appreciated that true one-to-one connections in the retina do not exist; a cone, although making an exclusive type of synapse with a midget bipolar, may also make a less exclusive contact with a flat bipolar cell; furthermore, midget bipolars and cones are connected laterally by amacrine and horizontal cells so that it is most unlikely that a given optic nerve fibre carries messages from only a single cone. The one-to-one relationship may in fact exist under certain conditions, but that is because pathways from other receptors have been blocked or occluded by inhibitory processes that keep the line clear for a given cone.

Scotopic acuity

The low visual acuity obtained in night, or rod, vision is now understandable. It has been pointed out that a high sensitivity to light is achieved by the convergence of rods on the higher neurons to allow spatial summation, and it is this convergence that interferes with the resolution of detail. If hundreds of rods converge on a single bipolar cell and if many bipolar cells converge on a single ganglion cell, it is understandable that the unit responsible for resolution may be very large and thus that the visual acuity is very small.

The retinal image. It has been implied, in the comments on visual acuity, that the limiting factor is one of an anatomical arrangement of receptors and of their neural organization. A very important feature, however, must be the accuracy of the formation of an image of external objects by the optical system of the eye. It may be calculated, for example, that the image of a grating produces lines 0.5 micron wide on the retina, but this is on the basis of ideal geometrical optics; in fact, the optics of the eye are not perfect, while diffraction of light by its passage through the pupil further spoils the image. As a result of these defects, the image of a black and white grating on the retina is not sharp, the black lines being not completely black but gray because of spread of light from the white lines. (When the optical system of the eye is defective, moreover, as in nearsightedness, the imagery is worse, but this can be corrected by the use of appropriate lenses.) Physiologically, the eye effectively improves the retinal image by enhancing contrasts; thus, the image of a fine black line on a white background formed on the retina is not a sharply defined black line but a relatively wide band of varying degrees of grayness; yet to the observer the line appears sharply defined, and this is because of lateral inhibition, the receptors that receive most light tending to inhibit those that receive less; the result is a physiological "sharpening of the image," so that the eye often behaves as though the image were perfect. This applies to chromatic aberration, too, which should cause black and white objects to appear fringed with colour, yet, because of suppression of the chromatic responses, one is not aware of the coloured fringes that do in effect surround the images of objects in the external world.

The pupil

The iris behaves as a diaphragm, modifying the amount of light entering the eye; probably of greater significance than control of the light entering the eye is the influence on aberrations of the optical system; the smaller the pupil the less serious, in general, are the aberrations. The smaller the pupil, however, the more serious become the effects of diffraction, so that a balance must be struck. Experimentally, it is found that at high luminances with pupils below three millimetres (0.12 inch) in diameter the visual acuity is not improved by further reduction of the diameter; increasing the pupil size beyond this reduces acuity, presumably because of the greater optical aberrations. It is interesting that when a subject is placed in a room that is darkened steadily, the size of the pupil increases, and the size attained for any given level of luminance is, in fact, optimal for visual acuity at this particular luminance. The reason that visual acuity increases with the larger pupils is that the extra light admitted into the eye compensates for the increased aberrations. When the gaze is fixed intently on an object for a long time, peripheral images that tend to disappear reappear immediately when the eyes are moved. This effect is called the Troxler phenomenon. To study it reproducibly it is necessary to use an optical device that ensures that the image of any object upon which the gaze is fixed will remain on the same part of the retina however the eyes move. Two investigators found, when they did this, that the stabilized retinal image tended to fade within a few seconds. It may be assumed that in normal vision the normal involuntary movements—the microsaccades and drifts mentioned earlier—keep the retinal image in sufficient movement to prevent the fading, which is essentially an example of sensory adaptation, the tendency for any receptive system to cease responding to a maintained stimulus.

Electrophysiology of the retina. *Neurological basis.* Subjective studies on human beings can traverse only a certain distance in the interpretation of visual phenomena; beyond this the standard electrophysiological techniques, which have been successful in unravelling the mechanisms of the central nervous system, must be applied to the eye; this, as repeatedly emphasized, is an outgrowth of the brain.

Receptive fields

Records from single optic nerve fibres of the frog and from the ganglion cell of the mammalian retina indicated three types of response. In the frog there were fibres that gave a discharge when a light was switched on the "on-fibres." Another group, the "off-fibres," remained inactive during illumination of the retina but gave a powerful discharge when the light was switched off. A third group, the "on-off fibres," gave discharges at "on" and "off" but were inactive during the period of illumination. The responses in the mammal were similar, but more complex than in the frog. The mammalian retina shows a background of activity in the dark, so that on- and off-effects are manifest as accentuations or diminutions of this normal discharge. In general, on-elements gave an increased discharge when the light was switched on, and an inhibition of the background discharge when the light was switched off. An off-element showed inhibition of the background discharge during illumination and a powerful discharge at off; this off-discharge is thus a release of inhibition and reveals unmistakably the inhibitory character of the response to illumination that takes place in some ganglion cells. Each ganglion cell or optic nerve fibre tested had a receptive field; and the area of frog's retina from which a single fibre could be activated varied with the intensity of the light stimulus. The largest field was obtained with the strongest stimulus, so that, in order that a light stimulus, falling at some distance away from the centre of the field, might affect this particular fibre it had to be much more intense than a light stimulus falling on the centre of the field. This means that some synaptic pathways are more favoured than others.

The mammalian receptive field is more complex, the more peripheral part of the field giving the opposite type of response to that given by the centre. Thus, if, at the centre of the field, the response was "on" (an on-centre field) the response to a stimulus farther away in the same fibre was at "off," and in an intermediate zone it was often mixed to give an on-off element. In order to characterize an element, therefore, it must be called on-centre or off-centre, with the meaning thereby that at the centre of its receptive field its response was at "on" or at "off," respectively, while in the periphery it was opposite. By studying the effects of small spot stimuli on centre and periphery separately and together, one investigator demonstrated a mutual inhibition between the two. A striking feature was the effect of adaptation; after dark adaptation the surrounding area of opposite activity became ineffective. In this sense, therefore, the receptive field shrinks, but, as it is a reduction in inhibitory activity between centre and periphery, it means, in fact, that the effective field can actually increase during dark adaptation—*i.e.,* the regions over which summation can occur—and this is exactly what is found in psychophysical experiments on dark adaptation.

Inhibitory effect of periphery receptors

Anatomical basis. The receptive field is essentially a measure of the number of receptors—rods or cones or a mixture of these—that make nervous connections with a single ganglion cell. The organization of centre and periphery implies that the receptors in the periphery of an on-centre cell tend to inhibit it, while those in the centre of the field tend to excite it, so that the effects of a uniform illumination covering the whole field tend to cancel out. This has an important physiological value, as it means, in effect, that the brain is not bombarded with an enormous number of unnecessary messages, as would be the case were every ganglion cell to send discharges along its optic nerve fibre as long as it was illuminated. Instead, the cell tends to respond to change—*i.e.,* the movement of a light or dark spot over the receptive field—and to give an especially prominent response, often when the spot passes from the periphery to the centre, or vice versa. Thus, the centre-periphery organization favours the detection of movement; in a similar way it favours the detection of contours because these give rise to differences in the illumination of the parts of the receptive fields. The anatomical basis of the arrangement presumably is given by the organization of the bipolar and amacrine cells in relation to the dendrites of the ganglion cell; it is interesting that the actual diameter of the centre of the receptive field of a ganglion cell is frequently equal to the area over which its dendrites spread; the periphery exerts its effects presumably by means of amacrine cells that are capable of connecting with bipolars over a wide area. These amacrine cells could exert an inhibitory action on the bipolar cells connected to the receptors of the central zone of the field, preventing them from responding to these receptors; in this case, the ganglion cell related to these bipolars would be of an on-centre and off-periphery type.

Direction-sensitive ganglion cells. When examining the receptive fields of rabbit ganglion cells, investigators found some that gave a maximal response when a moving spot of light passed in a certain "preferred" direction, while they gave no response at all when the spot passed in the opposite direction; in fact, the spontaneous activity of the cell was usually inhibited by this movement in the "null" direction. It may be assumed that the receptors connected with this type of ganglion cell are organized in a linear fashion, so that the stimulation of one receptor causes inhibition of a receptor adjacent to it. This inhibition would prevent the excitatory effect of light on the adjacent receptor from having a response when the movement was in the null direction, but would arrive too late at the adjacent receptor if the light was moving in the preferred direction.

The electroretinogram. If an electrode is placed on the cornea and another, indifferent electrode, placed, for example, in the mouth, illumination of the retina is followed by a succession of electrical changes; the record of these is the electroretinogram or ERG. Modern analysis has shown that the electrode on the cornea picks up changes in potential occurring successively at different levels of the retina, so that it is now possible to recognize, for example, the electrical changes occurring in the rods and cones—the receptor potentials—those occurring in the horizontal cells, and so on. In general, the electrical changes caused by the different types of cell tend to overlap in time, so that the record in the electroretinogram is only a faint and attenuated index to the actual changes; nevertheless, it has, in the past, been a most valuable tool for the analysis of retinal mechanisms. Thus, the most prominent wave—called the *b*-wave—is closely associated with discharge in the optic nerve, so that in animals, or man, the height of the *b*-wave can be used as an objective measure of the response to light. Hence, the sensitivity of the dark-adapted frog's retina to different wavelengths, as indicated by the heights of the *b*-waves, can be plotted against wavelength to give a typical scotopic sensitivity curve with a maximum at 5000 angstroms (one angstrom = 1×10^{-4} micron) corresponding to the maximum for absorption of rhodopsin.

Flicker. Electrophysiology has been used as a tool for the examination of the basic mechanism of flicker and fusion. The classical studies based on the electroretinogram indicated that the important feature that determines fusion in the cone-dominated retina is the inhibition of the retina caused by each successive light flash, inhibition being indicated by the *a*-wave of the electroretinogram. In the rod-dominated retina—*e.g.,* in man under scotopic conditions— the *a*-wave is not prominent, and fusion depends simply on the tendency for the excitatory response to a flash to persist, the inhibitory effects of a succeeding stimulus being small. More modern methods of analysis, in which the discharges in single ganglion cells in response to repeated flashes are measured, have defined fairly precisely the nature of fusion, which, so far as the retinal message is concerned, is a condition in which the record from the ganglion cell becomes identical with the record observed in the ganglion cell during spontaneous discharge during constant illumination.

The nature of fusion

Visual acuity. Although the resolving power of the retina depends, in the last analysis, on the size and density of packing of the receptors in the retina, it is the neural organization of the receptors that determines whether the brain will be able to make use of this theoretical resolving power. It is therefore of interest to examine the responses of retinal ganglion cells to gratings, either projected as stationary images on to the receptive field or moved slowly across it. One group of investigators showed that ganglion cells of the cat differed in sensitivity to a given grating when the sensitivity was measured by the degree of contrast between the black and white lines of the grating necessary to evoke a measurable response in the ganglion cell. When the lines were made very fine (*i.e.,* the "grating-frequency" was high), a point was reached at which the ganglion cell failed to respond, however great the contrast; this measured the resolving power of the particular cell being investigated. The interesting feature of this work is that individual ganglion cells had a special sensitivity to particular grating-frequencies, as if the ganglion cells were "tuned" to particular frequencies, the frequencies being measured by the number of black and white lines in a given area of retina. When the same technique was applied to human subjects, the electrical changes recorded from the scalp being taken as a measure of the response, the same results were obtained.

Colour vision. The spectrum, obtained by refracting light through a prism, shows a number of characteristic regions of colour—red, orange, yellow, green, blue, indigo, and violet. These regions represent large numbers of individual wavelengths; thus, the red extends roughly from 7600 angstrom units to 6500; the yellow from 6300 to 5600; green from 5400 to 5000; blue from 5000 to 4200; and violet from 4200 to 4000. Thus, the limits of the visual spectrum are commonly given as 7600 to 4000 angstroms. In fact, however, the retina is sensitive to ultraviolet light to 3500 angstroms, the failure of the short wavelengths to stimulate vision being due to absorption by the ocular media. Again, if the infrared radiation is strong enough, wavelengths as long as 10,000–10,500 angstroms evoke a sensation of light.

Hue discriminations

Within the bands of the spectrum, subtle distinctions in hue may be appreciated. The power of the eye to discriminate light on the basis of its wavelength can be measured by projecting onto the two halves of a screen lights of different wavelengths. When the difference is very small—*e.g.,* five angstroms—no difference can be appreciated. As the difference is increased, a point is reached when the two halves of the screen appear differently coloured. The hue discrimination (hue is the quality of colour that is determined by wavelength) measured in this way varies with the region of the spectrum examined; thus, in the blue-green and yellow it is as low as 10 angstroms, but in the deep red and violet it may be 100 angstroms or more. Thus, the eye can discriminate several hundreds of different spectral bands, but the capacity is limited. If it is appreciated that there are a large number of nonspectral colours that may be made up by mixing the spectral wavelengths, and by diluting these with white light, the number of different colours that may be distinguished is high indeed.

Spectral sensitivity curve. At extremely low intensities of stimuli, when only rods are stimulated, the retina shows a variable sensitivity to light according to its wavelength, being most sensitive at about 5000 angstroms, the absorption maximum of the rod visual pigment, rhodopsin. In the light-adapted retina one may plot a similar type of curve, obtained by measuring the relative amounts of light energy of different wavelengths required to produce the same sensation of brightness; now the different stimuli appear coloured, but the subject is asked to ignore the colours and match them on the basis of their luminosity (brightness). This is carried out with a special instrument called the flicker-photometer. There is a characteristic shift in the maximum sensitivity from 5000 angstroms for scotopic (night) vision to 5550 angstroms for photopic (day) vision, the so-called Purkinje shift. It has been suggested that the cones have a pigment that shows a maximum of absorption at 5550 angstroms, but the phenomena of colour vision demand that there be three types of cone, with three separate pigments having maximum absorption in the red, green, and blue, so that it is more probable that the photopic luminosity curve is a reflection of the summated behaviour of the three types of cone rather than of one.

The Purkinje shift has an interesting psychophysical correlate; it may be observed, as evening draws on, that the luminosities of different colours of flowers in a garden change; the reds become much darker or black, while the blues become much brighter. What is happening is that, in this range of luminosities, called mesopic, both rods and cones are responding, and, as the rod responses become more pronounced—*i.e.,* as darkness increases—the rod luminosity scale prevails over that of the cones.

It may be assumed that the sensation of luminosity under any given condition is determined by certain ganglion cells that make connections to all three types of cone and also to rods; at extremely low levels of illumination their responses are determined by the activity aroused in the rods. As the luminance is increased, the ganglion cell is activated by both rods and cones, and so its luminosity curve is governed by both rod and cone activity. Finally, at extremely high luminances, when the rods are "saturated" and ceasing to respond, the luminosity curve is, in effect, compounded of the responses of all three types of cone.

Colour mixing. The fundamental principle of colour mixing was discovered by Isaac Newton when he found that white light separates spatially into its different component colours on passing through a prism. When the same light is passed through another prism, so that the individual bands of the spectrum are superimposed on each other, the sensation becomes one of white light. Thus, the retina, when white light falls on it, is really being exposed to all the wavelengths that make up the spectrum. Because these wavelengths fall simultaneously on the same receptors, the evoked sensation is one of white. If the wavelengths are spread out spatially, they evoke separate sensations, such as red or yellow, according to which receptors receive which bands of wavelengths. In fact, the sensation of white may be evoked by employing much fewer wavelengths than those in the spectrum: namely, by mixing three primary hues—red, green, and blue.

Three primary hues

Furthermore, any colour, be it a spectral hue or not, may be matched by a mixture of these three primaries, red, green, and blue, if their relative intensities are varied. Many of the colours of the spectrum can be matched by mixtures of only two of the primary colours, red and green; thus the sensations of red, orange, yellow, and green may be obtained by adding more and more green light to a red one.

To one accustomed to mixing pigments, and to mixing a blue pigment, for example, with yellow to obtain green, the statement that red plus green can give yellow or orange, or that blue plus yellow can give white, may sound strange. The mixing of pigments is essentially a subtractive process, however, as opposed to the additive process of throwing differently coloured lights on a white screen. Thus, a blue pigment is blue because it reflects mainly blue (and some green) light and absorbs red and yellow; and a yellow pigment reflects mainly yellow and some green and absorbs blue and red. When blue and yellow pigments are mixed, and white light falls on the mixture, all bands of colour are absorbed except for the green colour band.

Colour defectiveness. The colour-defective subject is one whose wavelength discrimination apparatus is not as good as that of the majority of people, so that he sees many colours as identical that normal people would see as different. About one percent of males are dichromats; they can mix all the colours of the spectrum, as they see them, with only two primaries instead of three. Thus, the protanope (red blind) requires only blue and green to make his matches; since, for the normal (trichromatic) subject the various reds, oranges, yellows, and many greens are the result of mixing red and green, the protanope matches all these with a green. In other words, he is unable to distinguish all these hues from each other on the basis of their colour; if he distinguishes them, it is because of their different luminosity (brightness). The protanope matches white with a mixture of blue and green and is, in fact, unable to distinguish between white and bluish-green. The deuteranope (green blind) matches all colours with a mixture of red and blue; thus, his white is a mixture of red and blue that appears purple to a person with normal vision. The deuteranope also is unable to discriminate reds, oranges, yellows, and many greens, so that both types of dichromat are classed as red-green-blind. For the protanope, however, the spectrum is more limited because he is unable to appreciate red. The tritanope (blue blind) is rare, constituting only one in 13,000 to 65,000 of the population; because he is blue blind, his colour discrimination is best in the region of red to green, where that of the protanope and deuteranope is worse.

Responses of uniform population of receptors. The scotopic (night) visual system, mediated by rods, is unable to discriminate between different wavelengths; thus, a threshold stimulus of light with a wavelength of 4800 angstroms gives a sensation of light that is indistinguishable from that evoked by a wavelength of 5300 angstroms. If the intensities are increased, however, the lights evoke sensations of blue and green, respectively. Rods are unable to mediate wavelength, or colour, discrimination while the cones can because the rods form a homogeneous population, all containing the same photopigment, rhodopsin. Thus, the response of a nerve cell connected with a rod or group of rods will vary with the wavelength of light, and probably in the manner indicated by Figure 42, in which the response, measured in frequency of discharge in the bipolar or ganglion cell, is plotted against the wavelength of the stimulating light. The curve is essentially similar to the absorption spectrum of rhodopsin when the same amount of energy is in each stimulus; thus, blue-green of 5000 angstroms has the most powerful effect because it is absorbed most efficiently, while violet and red have the smallest effects. In this sense, the rods behave as wavelength discriminators, but it is to be noted that there are pairs of wavelengths on each side of the peak to which the same response is obtained; thus, a blue of 4800 angstroms and a yellow of 6000 angstroms give the same discharge. Moreover, if the intensity of the stimulus is varied, a new curve is obtained (as in the dotted line of Figure 42), and now the same response is obtained with a high intensity of violet at 4000 angstroms as with blue at the lower intensity. In general, it is easy to show that, by varying the intensity of the stimulus of a single wavelength, all types of response may be obtained, so that the brain would never receive a message indicating, in a unique fashion, that the retina was stimulated with, say, green light of 5300 angstroms; the same message could be given by blue light of 4800 angstroms, red light of 6500 angstroms, and so on.

From H. Davson, M.D., *Physiology of the Eye*

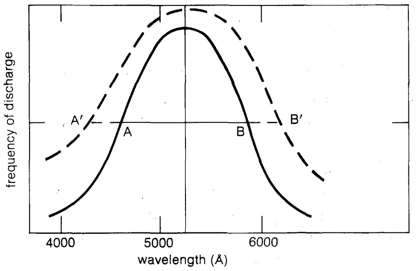


Figure 42: Theoretical wavelength response curve for a single receptor of *Limulus* retina. The maximum response will occur at 5200 Å, the absorptional maximum of the retinal pigment. With wavelengths of light corresponding to the points A and B, the responses will be identical, so that no discrimination between these two wavelengths is possible. When the intensity of the light is changed and a new curve is obtained (the dotted line), the responses at A′ and B′ are also identical.

Why rods fail to discriminate between colours

Ideally, colour discrimination would require a large number of receptors specifically sensitive to small bands of the spectrum, but the number would have to be extremely large because the capacity for hue discrimination is extremely great, as has been indicated. In fact, however, the phenomena of colour mixing suggest that the number of receptors may be limited.

Young-Helmholtz theory. It was the phenomena of colour mixing that led Thomas Young in 1802 to postulate that there are three receptors, each one especially sensitive to one part of the spectrum; these receptors were thought to convey messages to the brain, and, depending on how strongly they were stimulated by the coloured light, the combined message would be interpreted as that due to the actual colour. The theory was developed by Hermann Ludwig Ferdinand von Helmholtz, and is called the Young-Helmholtz trichromatic theory. As expressed in modern terms, it is postulated that there are three types of cone in the retina, characterized by the presence of one of three different pigments, one absorbing preferentially in the red part of the spectrum, another in the green, and another in the blue. A coloured stimulus—*e.g.,* a yellow light—would stimulate the red and green receptors, but would have little effect on the blue; the combined sensation would be that of yellow, which would be matched by stimulating the eye with red and green lights in correct proportions of relative intensity. A given coloured stimulus would, in general, evoke responses in all three receptors, and it would be the pattern of these responses—*e.g.,* blue strongly, green less strongly, and red weakest—that would determine the quality of the sensation. The intensity of the sensation would be determined by the average frequencies of discharge in the receptors. Thus, increasing the intensity of the stimulus would clearly change the responses in all the receptors, but if they maintained the same pattern, the sensation of hue might remain unaltered and only that of intensity would change; the observer would say that the light was brighter but still bluish green. Thus, with several receptors, the possibility is reduced of confusion between stimuli of different intensity but the same wavelength composition; the system is not perfect because the laws of colour mixing show the eye is incapable of certain types of discrimination, as, for example, between yellow and a mixture of red and green, but as a means of discriminating subtle changes in the environment the eye is a very satisfactory instrument.

The direct proof that the eye does contain three types of cone has been secured, but only relatively recently. This was done by examining the light emerging from the eye after reflection off the retina; in the dark-adapted eye the light emerging was deficient in blue light because this had been preferentially absorbed by the rhodopsin. In the light-adapted eye, when only cone pigments are absorbing light, the emerging light can be shown to be deficient in red and green light because of the absorption by pigments called erythrolabe and chlorolabe. Again, the light passing through individual cones of the excised human retina can be examined by a microscope device, and it was shown by such examination that cones were of three different kinds according to their preference for red, green, and blue lights.

The nervous messages. If the three types of cones respond differently to light stimuli, one may expect to find evidence for this difference in type of response by examining the electrophysiological changes taking place in the retina; ideally, one should like to place a microelectrode in or on a cone, then in or on its associated bipolar cell, and so on up the visual pathway. In the earliest studies, the optic nerve fibres of the frog were examined—*i.e.,* the axons of ganglion cells. The light-adapted retina was stimulated with wavelengths of light stretching across the

Dominator impulse

spectrum, and the responses in arbitrarily selected single fibres were examined. The responses to stimuli of the same energy but different wavelengths were plotted as frequency of discharge against wavelength, and the fibres fell into several categories, some giving what the investigator called a dominator response, the fibre responding to all wavelengths and giving a maximum response in the yellow-green at 5600 angstroms. Other fibres gave responses only over limited ranges of wavelengths, and their wavelengths of maximum response tended to be clustered in the red, green, and blue regions. The investigator called these modulators, and considered that the message in the dominator indicated to the brain the intensity of the stimulus—*i.e.,* it determined the sensation of brightness—while the modulators indicated the spectral composition of the stimulus, the combined messages in all the modulators resulting in a specific colour sensation. In the dark-adapted retina, when only rods were being stimulated, the response was of the dominator type, but this time the maximum response occurred with a wavelength of 5000 angstroms, the absorption maximum of rhodopsin.

A more careful examination of the responses in single fibres, especially in the fish, which has good colour vision, showed that things were not quite as simple as the original investigator had thought because, as has been seen, the response of a ganglion cell, when light falls on its receptive field in the retina, is not just a discharge of action potentials that ceases when the light is switched off. This type of response is rare; the most usual ganglion cell or optic nerve fibre has a receptive field organized in a concentric manner, so that a spot of light falling in the central part of the field produces a discharge, while a ring of light falling on the surrounding area has the opposite effect, giving an off-response—*i.e.,* giving a discharge only when the light is switched off. Such a ganglion cell would be called an on-centre-off-periphery unit; others behaved in the opposite way, being off-centre-on-periphery.

When these units are examined with coloured lights, and when care is taken to stimulate the centres and surrounding areas separately, an interesting feature emerges; the centre and surrounding areas usually have opposite or opponent responses. Thus, some may be found giving an on-response to red in the centre of the field and an off-response to green in the surrounding area, so that simultaneous stimulation of centre with red and surrounding area with green gives no response, the inhibitory effect of the off-type of response cancelling the excitatory effect of the on-type. With many other units the effects were more complex, the centre giving an on-response to red and an off-response to green, while the surrounding area gave an off-response to red and an on-response to green, and vice versa. This opponent organization probably subserves several functions. First, it enables the retina to emphasize differences of colour in adjacent parts of the field, especially when the boundary between them moves, as indeed it is continually doing in normal vision because of the small involuntary movements of the eyes. Second, it is useful in "keeping the retina quiet"; there are about one million optic nerve fibres, and if all these were discharging at once the problem of sorting out their messages, and making meaning of them, would be enormous; by this "opponence," diffuse white light falling on many of these chromatic units would have no effect because the inhibitory surrounding area cancelled the excitatory centre, or vice versa. When the light became coloured, however, the previously inactive units could come into activity.

S-potentials

These responses show that by the time the effect of light has passed out of the eye in the optic nerve the message is well colour-coded. Thus all the evidence points to the correctness of the Young-Helmholtz hypothesis with respect to the three-colour basis. The three types of receptor, responding to different regions of the spectrum in specific manners, transmit their effects to bipolar and horizontal cells. The latter neurons have been studied from the point of view of their colour-coding. The potentials recorded from them were called *S*-potentials; these were of two types, which classified them as responding to colour (*C*-units) and luminosity (*L*-units).

The *C*-type of cell gave an opponent type of response, in the sense that the electrical sign varied with the wavelength band, red and green having opponent effects on some cells, and blue and yellow on others. These responses reflect the connections of the horizontal cells to groups of different cones, the blue-yellow type, for example, having connections with blue and red and green cones, while the red-green would have connections only with red and green cones.

Lateral geniculate cells. As indicated above, the cells at the next stage, the ganglion cells, give a fairly precisely coded set of messages indicating the chromatic (colour) quality and the luminosity (brightness) of the stimulus, organized in such a way, however, as to facilitate the discrimination of contrast. At higher stages—*e.g.,* in the cells of the lateral geniculate body—this emphasis on opponence, or contrast, is maintained and extended; thus, several types of cell have been described that differ in accordance with the organization of their receptive fields from the colour aspect; some were very similar to ganglion cells, while others differed in certain respects. Some showed no opponence between colours when centre and periphery were compared, so that if a red light on the periphery caused inhibition, green and blue light would also do so. Others had no centre-periphery organization, the receptive field consisting of only a central spot; different colours had different effects on this spot; and so on.

In the cerebral cortex there is the same type of opponence with many units, but because cortical cells require stimuli of definite shape and often are not activated by simple spot stimuli, early studies carried out before these requirements were known probably failed to elucidate the true chromatic requirements of these high-order neurons. In general, the responses are what might be predicted on the basis of connections made to lateral geniculate neurons having the chromatic responses already known. Thus the final awareness of colour probably depends on the bombardment of certain higher-order cortical neurons by groups of primary cortical neurons, each group sending a different message by virtue of the connections it makes to groups of cones, connections mediated, of course, through the neurons of the retina and lateral geniculate body.

The photochemical process. For the energy of light to exert its effect it must be absorbed; it has been stated above that the action-spectrum for vision (the sensitivity of the eye to light) in the completely dark-adapted eye has a maximum in the region of 5000 angstroms, and that this corresponds with the maximum of absorption of light by the pigment, rhodopsin, extracted from the dark-adapted retina of the same species. The chemical nature of rhodopsin must now be examined, as well as its localization in the rod and the changes it undergoes in response to the absorption of light. It must be appreciated at the outset that the amount of light energy absorbed by a single rod at the threshold for vision is extremely small—namely, one quantum—and this is quite insufficient to provide the energy required to cause an electrical change in the membrane of the rod that will be propagated from the point of absorption of the light to the rod spherule (which takes part in the synapse between rod and bipolar cell). There must, therefore, be a chemical amplification process taking place within the rod, and the absorption of a quantum must be viewed as the trigger that sets off other changes, which in turn provide the required amount of energy.

Opsin and retinal

Rhodopsin. Visual purple, or rhodopsin, is a chromoprotein, a protein, opsin, with an attached chromatophore ("pigment-bearing") molecule that gives it its colour—*i.e.,* that allows it to absorb light in the visible part of the spectrum. In the absence of such a chromatophore, the protein would only absorb in the ultraviolet and so would appear colourless to the eye. The chromatophore group was identified as retinal, which is the substance formed by oxidation of vitamin A; on prolonged exposure of the eye to light, retinal can be found, free from the protein opsin, in the retina. When the eye is allowed to remain in the dark, the rhodopsin is regenerated by the joining up of retinal with opsin. Thus one may write:

$$\text{rhodopsin} \rightleftharpoons \text{retinal} + \text{opsin}.$$

The incidence of light on the retina causes the reaction to go to the right (that is, causes rhodopsin to form retinal plus opsin), and this photochemical change causes the sensation of light. The process is reversed by a thermal—*i.e.,* non-photochemical—reaction, so that for any given light intensity a steady state is reached with the regenerative process just keeping pace with the photochemical bleaching. Dark adaptation, or one element in it, is the regenerative process. The change in the rhodopsin molecule that leads to its bleaching—*i.e.,* the splitting off of the retinal molecule—takes place in a succession of steps; and there is reason to believe that the electrical change in the rod that eventually evokes the sensation of light occurs at a stage well before the splitting off of the retinal. One may describe as a transduction process the chemical events that take place between the absorption of light and the electrical event, whatever that may be; the rod behaves as a transducer in that it converts light into electrical or neural energy.

Prelumirhodopsin

The transduction process. Immediately after absorption of a quantum, the rhodopsin molecule is changed into a substance called prelumirhodopsin, recognized by its different colour from that of rhodopsin; this product is so highly unstable that at body temperature it is converted, without further absorption of light, into a series of products. These changes may be arrested by cooling the solution to −195° C (−319° F), at which temperature prelumirhodopsin remains stable; on warming to −140° C (−220° F) prelumirhodopsin becomes lumirhodopsin, with a slightly different colour; on warming further, successive changes are permitted until finally retinal is split off from the opsin to give a yellow solution. The important point to appreciate is that only at this stage is the chromatophore group split off; the earlier products have involved some change in the structure of the chromoprotein, but not so extreme as to break off the retinal. The precise nature of these changes is not yet completely elucidated, but the most fundamental one—namely, that occurring immediately after absorption of the quantum—has been shown to consist in a change in shape of the retinal molecule while it is still attached to opsin.

Photoisomerization

Thus retinal, like vitamin A, can exist in several forms because of the double bonds in its carbon chain—the socalled *cis-trans* isomerism. In other words, the same group of atoms constituting the retinal molecule can be twisted into a number of different shapes, although the sequence of the atoms is unaltered. While attached to the opsin molecule in the form of rhodopsin, the retinal has a shape called 11-*cis,* being somewhat folded, while on conversion to prelumirhodopsin the retinal has a straighter shape called all-*trans;* the process is called one of photoisomerization, the absorption of light energy causing the molecule to twist into a new shape. Having suffered this alteration in shape, the retinal presumably causes some instability in the opsin, making it, too, change its shape, and thereby exposing to the medium in which it is bathed chemical groupings that were previously shielded by being enveloped in the centre of the molecule. It may be assumed that these changes in shape induce alterations in the light-absorbing character of the molecule that permit the recognition of the new forms of molecule represented by lumirhodopsin, metarhodopsins I and II, and so on.

The final change is more drastic because it involves the complete splitting off of the retinal; an earlier stage—namely, the conversion of metarhodopsin I to metarhodopsin II—has been shown recently to involve a bodily change in position of the retinal, which in rhodopsin is linked to the lipid (fatty) portion of the molecule, whereas in metarhodopsin II it is found to have become attached to an amino acid in the backbone-chain of the protein (amino acids are subunits of proteins). Thus, in its native unilluminated state, retinal is attached to a lipid, which is presumably linked to the protein, so that rhodopsin is more properly called a chromolipoprotein rather than a chromoprotein. The outer segments of the rods are, as has been stated, constituted by membranous disks, and it is well established that the material from which these membranes are constructed is predominantly lipid, so that one may envisage the rhodopsin molecules as being, in fact, part of the membrane structure. The techniques used for extraction presumably tear the molecules from the main body of the lipid, but some of the lipid remains with the protein and retinal to constitute the link holding these two parts together.

Within the retina these chemical changes are all reversible, so that when a steady light is maintained on the retina the latter will contain a mixture of several or all of the intermediate compounds. In the dark, all will be gradually reconverted to rhodopsin. Because lack of vitamin A, from which retinal is derived, causes night blindness, some of the retinal must get lost from the eye to the general circulation; and it is actually replaced by the cells of the pigment epithelium, which are closely associated with the rods.

As to which of these chemical changes acts as the trigger for vision, there is some doubt. The discovery that the transition from metarhodopsin I to metarhodopsin II involves an actual shift of the retinal part of the molecule from linkage to lipid to linkage to protein reinforces the belief that this particular shift is sufficient to lead ultimately to electrical discharges in the optic nerve.

Cone pigments. So far as colour vision is concerned, the changes that take place in the three cone pigments have not been analyzed, simply because, so far, they have defied isolation, presumably because their concentrations are so much less than that of the rod pigment.

THE HIGHER VISUAL CENTRES

Partial decussation

The visual pathway. The axons of the ganglion cells converge on the region of the retina called the papilla or optic disk. They leave the globe as the optic nerve, in which they maintain an orderly arrangement in the sense that fibres from the macular zone of the retina occupy the central portion, the fibres from the temporal half of the retina take up a concentric position, and so on; when outside the orbit, there is a partial decussation (crossover). The fibres from the nasal halves of each retina cross to the opposite side of the brain, while those from the temporal halves remain uncrossed. This partial decussation is called the chiasma. The optic nerves after this point are called

By permission from Eugene Wolff, *Anatomy of the Eye and Orbit,* London: H. K. Lewis & Co. Ltd.

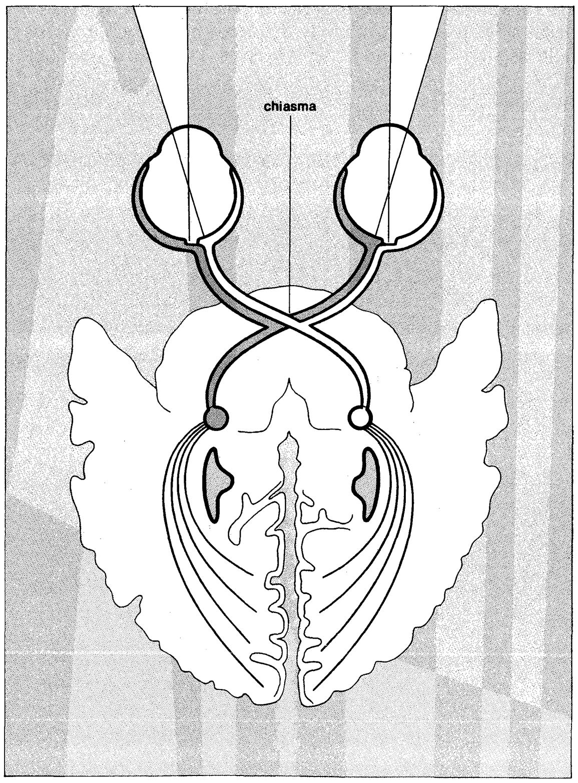


Figure 43: *Visual pathways.*
Fibres from the nasal side of the retina cross over in the chiasma to join the uncrossed fibres of the temporal half of the retina.

the optic tracts, containing nerve fibres from both retinas. The result of the partial decussation is that an object in, say, the right-hand visual field produces effects in the two eyes that are transmitted to the left-hand side of the brain only. With cutaneous (skin) sensation there is a complete crossing-over of the sensory pathway; thus, information from the right half of the body, and the right visual field, is all conveyed to the left-hand part of the brain by the time that it has reached the diencephalon (the posterior part of the forebrain, Figure 43).

Fusion of retinal images. Partial decussation is an arrangement that serves the needs of frontally directed eyes and permits binocular vision, which consists in the fusion of the responses of both eyes to a single object—more loosely, one speaks of the fusion of the retinal images. In many lower mammals, with laterally directed eyes and therefore limited binocular vision, the degree of decussation is much greater, so that in the rat, for example, practically all of the optic nerve fibres pass to the opposite side of the brain.

The fibres of the optic tracts relay their messages to nerve cells in those parts of the diencephalon called the lateral geniculate bodies, and from the lateral geniculate bodies the messages are relayed to nerve cells in the occipital cortex of the same side. (The occipital cortex is the outer substance in the posterior portion of the brain.)

The visual field. If one eye is fixed on a point in space, the visual field for this eye may be thought of as the part of a surface of a sphere on to which all visible objects are projected. The limits to this field will be determined by the sensitivity and extent of the retina and the accessibility of light rays from the environment. Experimentally or clinically, the field is measured on a perimeter, a device for ascertaining the point on a given meridian where a white spot just appears or disappears from vision when moved along this meridian. (A meridian is a curve on the surface of a sphere that is formed by the intersection of the sphere surface and a plane passing through the centre of the sphere.) The field is recorded on a chart, illustrated by Figure 44. On the nasal side, the field is restricted to about 60° from the midline. This is due to the obstruction caused by the nose, since the retina extends nearly as far forward on the temporal side of the globe as on the nasal side. It is customary to refer to the binocular visual field as that common to the two eyes, the uniocular field being the extreme temporal (outside) region peculiar to each eye. It will be clear from the field of the single eye shown in Figure 44 that the binocular field is determined in the horizontal meridian by the nasal field of each eye, and so will amount to about 60° to either side of the vertical meridian.

From H. Davson and M.G. Eggleton (eds.), *Principles of Human Physiology*

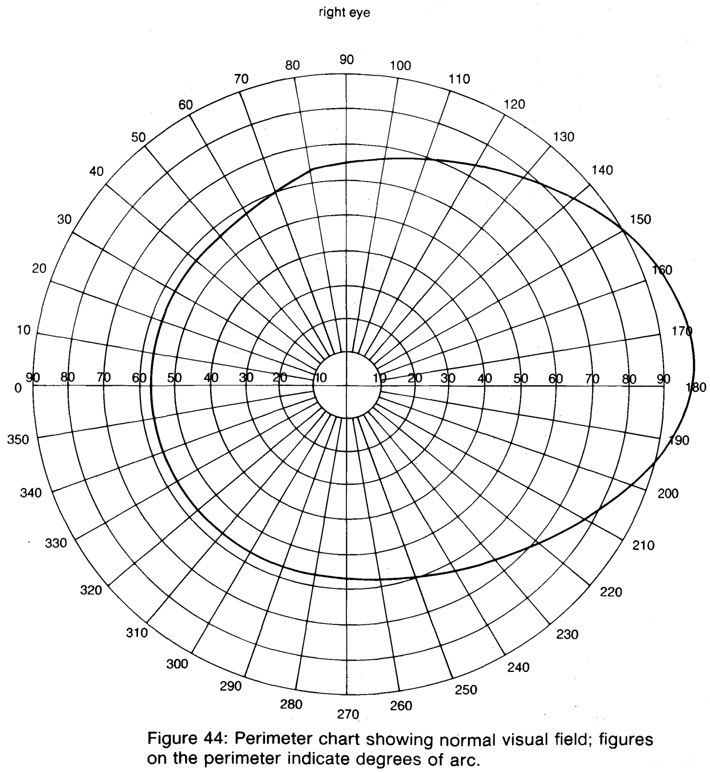


Figure 44: Perimeter chart showing normal visual field; figures on the perimeter indicate degrees of arc.

Lateral geniculate body. The dorsal (posterior) nucleus of the lateral geniculate body, where the optic tract fibres relay, has six layers, and the crossed fibres relay in layers 1, 4, and 6, while the uncrossed relay in layers 2, 3, and 5; thus, at this level, the impulses from the two eyes are kept separate, and when the discharges in geniculate neurons are recorded electrically it is rare to find any responding to stimuli in both eyes.

Point of awareness of visual stimulus

Striate area. The optic tract fibres make synapses with nerve cells in the respective layers of the lateral geniculate body, and the axons of these third-order nerve cells pass upward to the calcarine fissure (a furrow) in each occipital lobe of the cerebral cortex. This area is called the striate area because of bands of white fibres—axons from nerve cells in the retina—that run through it. It is also identified as Brodmann's area 17. It is at this level that the impulses from the separate eyes meet at common cortical neurons, or nerve cells, so that when the discharges in single cortical neurons are recorded it is usual to find that they respond to light falling in one or the other eye. It is probable that it is when the retinal messages have reached this level of the central nervous system, and not before, that the human subject becomes aware of the visual stimulus, since destruction of the area causes absolute blindness in man. Because of the partial decussation, however, the removal of only one striate cortex will not cause complete blindness in either eye, since only messages from two halves of the retinas will have been blocked; the same will be true if one optic tract is severed or one lateral geniculate body is destroyed. The result of such lesions will be half-blindness, or hemianopia, the messages from one half of the visual field being obliterated.

Pupillary pathways. Some of the fibres in the optic tracts do not relay in the lateral geniculate bodies but pass instead to a midbrain region—the pretectal centre—where they mediate (transmit) reflex alterations in the size of the pupil. Thus, in bright light, the pupils are constricted; this happens by virtue of the pupillary light reflex mediated by these special nerve fibres. Removal of the occipital cortex, although it causes blindness in the opposite visual field, does not destroy the reaction of the pupils to light; if the optic nerve is cut, however, the eye will be both completely blind and also unreactive to light falling on this eye. The pupil of the blind eye will react to light falling on the other eye by virtue of a decussation in the pupillary reflex pathway.

Point-to-point representation. Because of the ordered manner in which the optic tract fibres relay in the lateral geniculate bodies and from there pass in an orderly fashion to the striate area, when a given point on the retina is stimulated, the response recorded electrically in either the lateral geniculate body or the striate area is localized to a small region characteristic for that particular retinal spot. When the whole retinal field is stimulated in this point-to-point way, and the positions on the geniculate or striate gray matter on which the responses occur are plotted, it is possible to plot on these regions of the brain maps of the retinal fields or, more usually, maps of the visual fields.

Peristriate and parastriate areas

Visuopsychic or circumstriate areas. Area 17, the striate area, is the primary visual centre in the sense that, in primates at any rate, all of the geniculate fibres project onto it and none projects onto another region of the cortex. There are two other areas containing neurons that have close connections with the eye; these are the parastriate and peristriate areas, or Brodmann's areas 18 and 19, respectively, in close anatomical relationship to one another and to area 17. They are secondary visual areas in the sense that messages are relayed from area 17 to area 18 and from area 18 to area 19, and, because area 17 does not relay to regions beyond area 18, these circumstriate areas are the means whereby visual information is brought into relation with more remote parts of the cortex. Thus in writing, the eyes direct the activities of the fingers, which are controlled by a region of the frontal cortex, so that one may presume that visual information is relayed to this frontal region. In the monkey, bilateral destruction of the areas causes irrecoverable loss of a learned visual discrimination, but this can be relearned after the operation. In man, lesions in this region are said to cause disturbances in spatial orientation and stereoscopic vision, but much more knowledge is required before specific functions can be attributed to these circumstriate areas, if, indeed, this is possible.

Integration of the retinal halves. The two halves of the retina, and thus of the visual field, are represented on opposite cerebral hemispheres, but the visual field is perceived as a unity and hence one would expect an intimate connection between the two visual cortical areas.

Corpus callosum. The great bulk of the connections between the two sides of the cerebral mantle are made by the interhemispheric commissure (the point of union between the two hemispheres of the cerebrum) called the corpus callosum, which is made up of neurons and their axons and dendrites that make synapses with cortical neurons on symmetrically related points of the hemispheres. Thus, electrical stimulation of a point on one hemisphere usually gives rise to a response on a symmetrically related point on the other, by virtue of these callosal connections. The striate area is an exception, however, and it is by virtue of the connections of the striate neurons with the area 18 neurons that this integration occurs, the two areas 18 on opposite hemispheres being linked by the corpus callosum.

Stereopsis in the midline. Usually stereopsis, or perception of depth, is possible by the use of a single hemisphere because the images of the same object formed by right and left eyes are projected to the same hemisphere; however, if the gaze is fixed on a distant point and a pin is placed in line with this but closer to the observer, a stereoscopic perception of the distant point and the pin can be achieved by the fusion of disparate images of the pin, but the images of the pin actually fall on opposite retinal halves, so that this fusion must be brought about by way of the corpus callosum.

Callosal transfer. In experimental animals it is possible, by section of the chiasma, to ensure that visual impulses from one eye pass only to one hemisphere. If this is done, an animal trained to respond to a given pattern and permitted to use only one eye during the training is just as efficient, when fully trained, in making the discrimination with the other eye. There has thus been a callosal transfer of the learning so that the hemisphere that was not directly involved in the learning process can react as well as that directly involved. If the corpus callosum is also sectioned, this transfer is impossible, so that the animal, trained with one eye, must be trained again if it is to carry out the task with the other eye only.

Effect of cutting the corpus callosum

Superior colliculi. The visual pathway so far described is called the geniculostriate pathway, and in man it may well be the exclusive one from a functional aspect because lesions in this pathway lead to blindness. Nevertheless, many of the optic tract fibres, even in man, relay in the superior colliculi, a paired formation on the roof of the midbrain. From the colliculi there is no relay to the cortex, so that any responses brought about by this pathway do not involve the cortex. In man, as has been said, lesions in the striate area, which would of course leave the collicular centres intact, cause blindness, so that the visual fibres in these centres serve no obvious function. In lower animals, including primates, removal of the striate areas does not cause complete blindness; in fact, it is often difficult to determine any visual impairment from a study of the behaviour of the animals. Thus, in reptiles and birds, vision is barely affected, so that a pigeon that has been subjected to the operation can fly and avoid obstacles as well as a normal one. In rodents, such as the rabbit, removal of the occipital lobes causes some impairment of vision, but the animal can perform such feats as avoiding obstacles when running and recognizing food by sight. In the monkey, the effects are more serious, but the animal can be trained to discriminate lights of different intensity and even the shapes of objects, provided that these are kept in continual motion. It seems likely, then, that it is the visual pathway through the colliculi that permits the use of the eyes in the absence of visual cortex, although the connections of the optic tract fibres with the pulvinar of the thalamus (an area in the diencephalon), established in some animals, may well permit the use of regions of the cortex other than those denoted as visual.

SOME PERCEPTUAL ASPECTS OF VISION

So far, the visual process has been considered from rather elementary aspects; the ability to detect light and changes in its intensity, and to discriminate colour and form. It is now time to deal with more complex features, particularly some phenomena of binocular vision. It will then be in order to return to the electrophysiology of the visual pathway to see how some of the phenomena can be interpreted.

Projection of the retina. Objects are perceived in definite positions in space—positions definite in relation to each other and to the percipient. The first problem is to analyze the physiological basis for this spatial perception or, as it is expressed, the projection of the retina into space.

Relative positions of objects. The perception of the positions of objects in relation to each other is essentially a geometrical problem. Take, for the present, the perception of these relationships by one eye, monocular perception: a group of objects, as in Figure 45, produces images on the retina in a certain fixed geometrical relationship; for the perception of the fact that *C* is to the left of *D*, that *D* is to the left of *E*, and so on, it is necessary that the incidence of images at *c*, *d*, and *e* on the retina be interpreted in a similar, but, of course, inverted geometrical relationship. The neural requirements for this interpretation are (1) that the retina be built up of elements that behave as units throughout their conducting system to the visual cortex, and (2) that the retinal elements have "local signs." The local sign could represent an innate disposition or could result from experience—the association of the direction of objects in space, as determined by such evidence as that provided by touch, with the retinal pattern of stimulation. In neurophysiological terms, the retinal elements are said to be connected to cortical cells, each being specific for a given element, so that when a given cortical cell is excited the awareness is of a specific local sign. Studies of the projection of the retina on the cerebral cortex have confirmed this.

From H. Davson, M.D., *Physiology of the Eye*

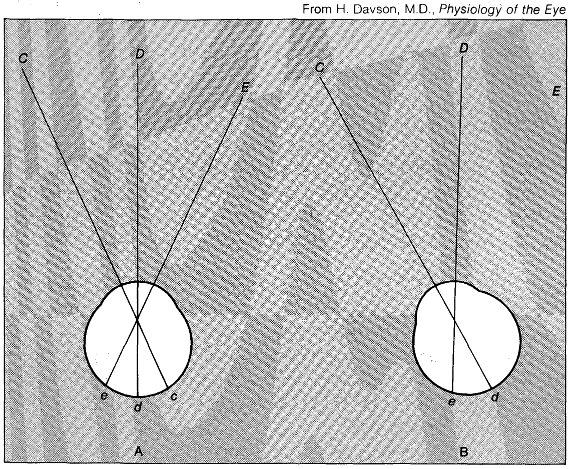


Figure 45: *Projection of retinal images into space.*
(A) A group of objects, *C,D,E,* produces images on the retina, *c,d,e;* the retinal images are projected outward in space toward the points evoking them. (B) The eye has moved to the left so that the image of *c* falls on *d,* previously projected to *D* but now projected to *C; E* no longer produces an image on the retina.

The retinal stimuli at *c*, *d*, and *e* in Figure 45 are appreciated as objects outside the eye, the retina is said to be projected into space, and the field of vision is thus the projection of the retina through the nodal point. (In Figure 45 the nodal point is the point of intersection of *Cc*, *Dd*, and *Ee*.) It will be seen that the geometrical relationship between objects and retinal stimuli is reversed; in the retina *c* is to the right of *d*, and so on.

Position in relation to observer. The recognition of the directions of objects in relation to the observer is more complex. If the eye (Figure 45B) is turned to the left, the image of *C* falls on the retinal point *d,* so that if *d* were always projected into the same direction in space, *C* would appear to be in *D*'s place. In practice, one knows that *C* is perceived as fixed in space in spite of the movements of the eye; hence, the direction of projection of a retinal point is constantly modified to take into account movements of the eye; this may be called psychological compensation. It will be seen that correct projection is achieved by projecting the stimulated retinal point through the nodal point of the eye. Movements of the eye caused by movements of the head must be similarly compensated. As a result, any point in space remains fixed in spite of movements of the eye and head. Given this system of compensated projection, the recognition of direction in relation to the individual is now feasible. *D* may be said to be due north or, more vaguely, "over there"; when the head is turned, since *D* is perceived to be in the same place, it is still due north or "over there." In some circumstances, the human subject makes an error in projecting his retinal image, so that the object giving rise to the image appears to be in a different place from its true one; the image is said to be falsely projected. If the eye is moved passively, for example, by pulling on the conjunctiva with forceps, the subject has the impression that objects in the outside world are moving in a direction opposite to that of the eye.

False projection

The apparent movement of an afterimage, when the eye moves, is an excellent illustration of psychological compensation. A retinal stimulus, being normally projected through the nodal point, is projected into different points in space as the eye moves; an afterimage can be considered to be the manifestation of a continued retinal impulse, and its projection changes as the eye moves. The afterimage thus appears to move in the same direction as that of the movement of the eye. Whether the drift of an afterimage across the field of view is entirely due to eye movements is difficult to say. One certainly has the impression that the eye is chasing the afterimage.

Visual estimates. *The directions of lines.* So far, consideration has been given to the problem of estimating the positions of points in relation to each other and to the percipient. The estimate of the directions of lines involves no really new principles, since, if two points, *A* and *B,* are exactly localized, the direction of the line *AB* can be appreciated. As will be seen, the organization of the neural connections of the retina and higher visual pathway is such as to favour the accurate recognition of direction; for the moment, the question of the maintenance of a frame of reference must be considered, in the sense that a map has vertical and horizontal lines with which to compare other directions. In fact, the vertical and horizontal meridians of the retina seem to be specialized as frames of reference; the accuracy with which a human subject can estimate whether a line is vertical or horizontal is very great.

An important point in this connection is that of the effects of eye movements on interpretation of the directions of lines because, when the eye moves to positions different from the primary straight-ahead position, the images of vertical lines will not necessarily fall on its vertical meridian. This can be due to an actual torsion of the eye about its anteroposterior (fore and aft) axis or to distortion of the retinal image. This means, then, that the line on the retina that corresponds to verticality in one position of the eye does not correspond to verticality in another, so that, once again, the space representation centre must take account not only of the retinal elements that have been stimulated but also of the corollary motor discharge.

Effects of eye movements

Comparison of lengths. The influence of the movements of the eyes in the estimation of length was emphasized by Helmholtz. An accurate comparison of the lengths of two parallel lines *AB* and *CD* can be made, whereas if an attempt is made to compare the nonparallel lines *A′B′* and *C′D′*, quite large errors occur. According to Helmholtz, the eye fixates first the point *A,* and the line *AB* falls along a definite row of receptors, thereby indicating its length. The eye is now moved to fixate *C,* and if the image of *CD* falls along the same set of receptors the length of *CD* is said to be the same as that of *AB.* Such a movement of the eye is not feasible with lines that are not parallel. Similarly, the parallelism, or otherwise, of pairs of lines can be perceived accurately because on moving the eye over the lines the distance between them must remain the same.

Fairly accurate estimates of relative size may be made, nevertheless, without movements of the eyes. If two equal lines are observed simultaneously, the one with direct fixation and the other with peripheral vision, their images fall, of course, on different parts of the retina; if the images were equally long it could be stated that a certain length of stimulated retina was interpreted as a certain length of line in space. It is probable that this is roughly the basis on which rapid estimates of length depend, although there are such complications as the fact that the retina is curved so that lines of equal length in different parts of the retina do not produce images of equal length on the retina.

Optical illusions. Many instances have been cited of well-defined and consistent errors in visual estimates under special conditions. There is probably no single factor by which the errors can be explained, but the tendency for distinctly perceptible differences to appear larger than those more vaguely perceived is important.

The perception of depth. *Monocular cues.* The image of the external world on the retina is essentially flat or two-dimensional, and yet it is possible to appreciate its three-dimensional character with remarkable precision; to a great extent this is by virtue of the simultaneous presentation of different aspects of the world to the two eyes, but even when the subject views the world with a single eye it does not appear flat to him and he can, in fact, make reasonable estimates of the relative positions of objects in all three dimensions. Examples of monocular cues are the apparent movements of objects in relation to each other when the head is moved. Objects nearer the observer move in relation to more distant points in the opposite direction to the movement of the head. Perspective, by which is meant the changed appearance of an object when it is viewed from different angles, is another important clue to depth. Thus the projected retinal image of an object in space may be represented as a series of lines on a plane—*e.g.,* a box—these lines, however, are not a unique representation of the box because the same lines could be used to convey the impression of a perfectly flat object with the lines drawn on it, or of a rectangular, but not cubical, box viewed at a different angle. In order that a three-dimensional object be correctly represented to the subject on a two-dimensional surface, he must know what the object is; *i.e.,* it must be familiar to him. Thus a bicycle is a familiar object. If it is viewed at an angle from the observer the wheels seem elliptical and apparently differ in size. Because the observer knows that the wheels are circular and of the same size, he perceives depth in a two-dimensional pattern of lines. The perception of depth in a two-dimensional pattern thus depends greatly on experience—the knowledge of the true shape of things when viewed in a certain way. Other cues are light and shade, overlapping of contours, and relative sizes of familiar objects.

Perspective

Binocular vision. The cues to depth mentioned above are essentially uniocular; they would permit the appreciation of three-dimensional space with a single eye. When two eyes are employed, two additional factors play a role, the one not very important—namely, the act of convergence or divergence of the eyes—and the other very important—namely, the stereoscopic perception of depth by virtue of the dissimilarity of the images presented by a three-dimensional object, or array of objects, to the separate eyes.

When a three-dimensional object or array is examined binocularly, the nearer points or objects require greater convergence for fixation than the more distant points or objects, so that this provides a cue to the three-dimensional character of the presentation. It is by no means a necessary cue, since presentation of the array for such a short time that movements of the eyes cannot occur still permits the three-dimensional perception, which is achieved under these conditions by virtue of the dissimilar images received by the two retinas.

A stereogram contains two drawings of a three-dimensional object taken from different angles, chosen such that the pictures are right- and left-eyed views of the object. When the stereogram is placed in a stereoscope, an optical device for enabling the two separate pictures to be fused and seen single, the impression created is one of a three-dimensional object. The perception is immediate, and is not a matter of interpretation. Clearly, with the stereoscope the situation is simulated as it normally occurs. To appreciate the full implications of the stereoscopic perceptual process, one must examine some simpler aspects of binocular vision.

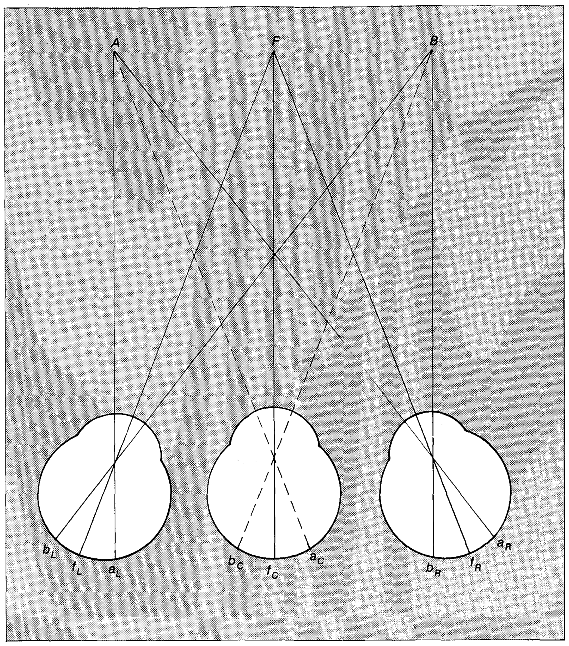


Figure 46: *The Cyclopean system of projection.* The images of the points *F, A,* and *B* on the two retinas are transposed to the retina of a hypothetical eye midway between the two. The pairs of images, a_L and a_R, b_L and b_R, and so on, coincide on the cyclopean retina indicating that the stimulated retina points are projected to a common direction (see text).

Figure 46 illustrates the situation in which a subject is fixating (fixing his gaze on) the point *F* so that the images of *F* fall on the foveal (retinal) points f_L and f_R, respectively. *F* is seen as a single point because the retinal points f_L and f_R are projected to the same point in space, and the projection is such that the subject says that the point *F* is straight in front of him, although it is to the right of his left eye and to the left of his right eye. The two eyes in this case are behaving as a single eye, "the cyclopean eye," situated in the centre of the forehead, and one may represent the projection of the two separate retinal points, f_L and f_R, as the single projection of the point f_C of the cyclopean eye. As will be seen, the cyclopean eye is a useful concept in consideration of certain aspects of stereoscopic vision.

Corre-
sponding
points

The points f_L and f_R may be defined as corresponding points because they have the same retinal direction values. The images formed by the points *A* and *B,* in the same frontal plane as *F,* fall on a_L and a_R and b_L and b_R; once again the pairs of retinal points are projected to the same points, namely, to *A* and *B,* and they are treated as being on the left and right of *F,* respectively. On the cyclopean projection, they may be said to be localized by the outward projections of a_C and b_C, respectively.

In Figure 47, the subject is once again fixing the point *F,* but the point *A* is now no longer in the same frontal plane as the point *F,* but closer to the observer. The images of *F* fall on corresponding points and are projected to a single point in front. The images of *A,* on a_L and a_R, do not fall on corresponding points and are, in fact, projected into space in different directions, as indicated by the cyclopean projection. This means that *A* is seen simultaneously at two different places, a phenomenon called physiological diplopia, and this in fact does happen, as can be seen by fixing one's gaze on a distant point and holding a pencil fairly close to the face; with a little practice the two images of the pencil can be distinguished. Thus, when the eyes are directed into the distance the objects closer to the observer are seen double, although one of the double images of any pair is usually suppressed. To return to Figures 46 and 47, *F* and *A* in Figure 46 are seen single and in the same plane because their images each fall on corresponding points. *F* is seen single and *A* double in Figure 47 because the images of *A* fall on noncorresponding, or disparate, points. *A* is appreciated as being closer to the observer than *F* in Figure 47 by virtue of these double images but, in general, although it is retinal disparity that creates the percept of three-dimensional space, it is not necessarily the formation of double images, since if the disparity is not large the point will be seen single, and this single point will appear to be in a different frontal plane from that containing the fixation point.

To appreciate the nature of this stereoscopic perception one must examine what is meant by corresponding points in a little more detail. In general, it seems that the two retinas are, indeed, organized in such a way that pairs of points are projected innately to the same point in space, and the horopter is defined as the outward projection of these pairs. One may represent this approximately by a sphere passing through the fixation point, or, if one confines attention to the fixation plane, it may be represented by the so-called Vieth-Müller horopter circle, as illustrated in Figure 48. On this basis, the corresponding points are arranged with strict symmetry, and each pair projects to a single point in space on the horopter circle. Theoretically, then, all points on the circle passing through the fixation point, *F,* will be seen single, and the point *X* will be seen double because it will be projected by the left eye to *F* and by the right eye to *A.* The actual situation is somewhat more complex than this, since experimentally the horopter turns out to have different shapes according to how close the fixation point is to the observer. The point to appreciate, however, is that the experimentally determined line, be it circular or straight or elliptical, is such that when points are placed on it they all appear to be in the same frontal plane—*i.e.,* there is no stereoscopic perception of depth when one views these points—and one may say that this is because the images of points on the horopter fall on corresponding points of the two retinas.

In Figure 49, to the left, are two eyes viewing an arrow

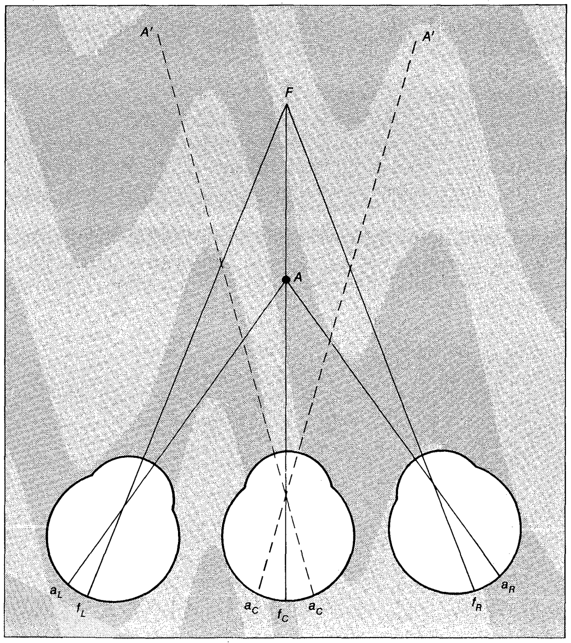


Figure 47: Physiological diplopia caused by an object, *A,* closer to the observer than the fixation point, *F.* The images of *A* fall on disparate or noncorresponding points on the two retinas, and these are projected to different points *A'* and *A'* (see text).

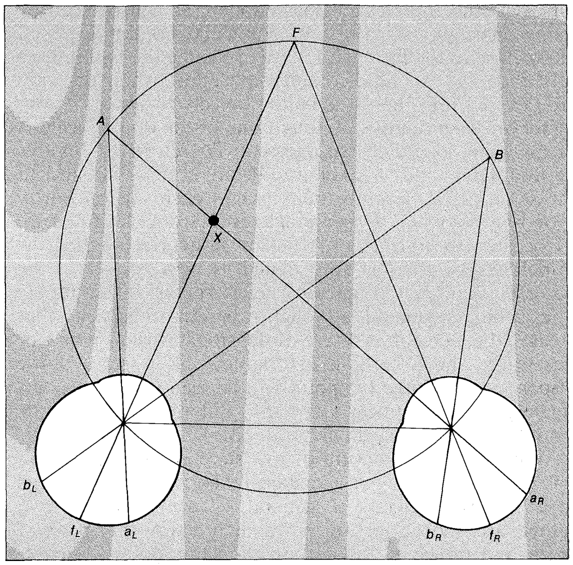


Figure 48: *The Vieth-Müller horopter circle.*
F is the fixation point. If corresponding points are symmetrically distributed about the foveas, the points in space in the fixation plane, whose images fall on the corresponding points, lie on the circle. The images of the point *X* lie on disparate points (see text).
From H. Davson, M.D., *Physiology of the Eye*

Fusion of disparate images

lying in the frontal plane—*i.e.*, with no stereopsis—and to the right the arrow is inclined into the third dimension—*i.e.*, it tends to point toward the observer. All points on the arrow are, in fact, seen single under both conditions, and yet it is clear from the right-hand figure that, if the gaze is fixed on *A*, the images of *B′* will fall on noncorresponding points. *B′* is not seen double but, instead, the noncorresponding points, b'_L and b'_R, are projected to a common point *B′* and a stereoscopic percept is achieved. Thus the noncorresponding, or disparate, points on the retinas can be projected to a single point, and it is essentially this fusion of disparate images by the brain that creates the impression of depth. If the point *B′* were brought much closer to the eyes, its images would fall on such disparate points that fusion would no longer be possible, and *B′* would be seen double, or one double image would be suppressed. There is thus a certain zone of disparity that, if not exceeded, allows fusion of disparate points. This is called Panum's fusional area; it is the area on one retina such that any point in it will fuse with a single point on the other retina.

To return to the stereoscopic perception of three-dimensional space, one may recapitulate that it is because the two eyes receive different images of the same object that the stereoscopic percept happens; when the two images of the object are identical, then, except under very special conditions, the object has no three-dimensionality. A special condition is given by a uniformly illuminated sphere; this is three-dimensional, but the observer would have to use special cues to discriminate this from a flat disk lying in the frontal plane. Such a cue might be the different degree of convergence of the eyes required to fixate the centre from that required to fixate the periphery, or the different degree of accommodation.

From H. Davson, M.D., *Physiology of the Eye*

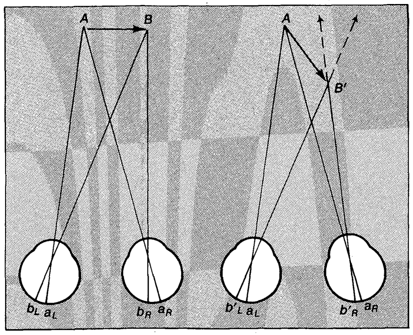


Figure 49: The distinction between corresponding points (a_L and a_R, b_L and b_R) and points that do not correspond (b'_L and b'_R; see text).

The difference in the two aspects of the same object (or group of objects), measured as the instantaneous parallax, is illustrated in Figure 50. *B* is closer to the observer than *A;* the fact is perceived stereoscopically because the line *AB* subtends different angles at the two eyes, and the instantaneous parallax is measured by the difference between the angles *a* and *b.* The binocular parallax of any point in space is given by the angle subtended at it by the line joining the nodal points of the two eyes; hence, the binocular parallax of *A* is *a;* that of *B* is *b;* the instantaneous parallax is thus the difference of binocular parallax of the two points considered.

Accuracy of stereoscopic perception

If one places three vertical wires in front of an observer in the frontal plane, one may move the middle one in front of, or behind, the plane containing the other two and ask the subject to say when he perceives that it is out of the plane; under correct experimental conditions the only cue will be the difference of binocular parallax, and it is found that the minimum difference is remarkably small, of the order of five seconds of arc, corresponding to a disparity of retinal images far smaller than the diameter of a single cone. With two editions of the same book, it is not possible, by mere inspection, to detect that a given line of print was not printed from the same type as the same line in the other book. If the two lines in question are placed in the stereoscope, it is found that some letters appear to float in space, a stereoscopic impression created by the minute differences in size, shape, and relative position of the letters in the two lines. The stereoscope may thus be used to detect whether a bank note has been forged, whether two coins have been stamped by the same die, and so on.

From Hugh Davson, M.D. *Physiology of the Eye*

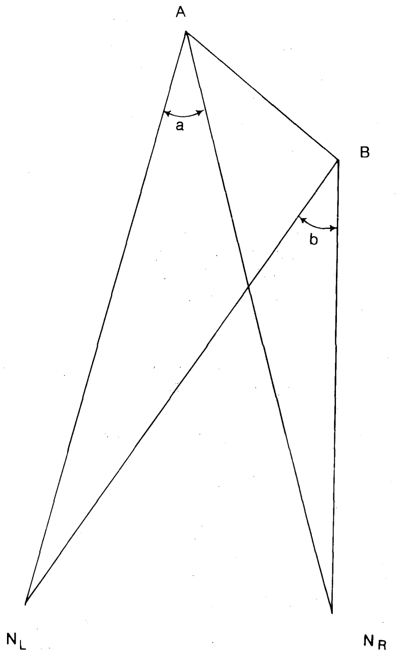


Figure 50: *Binocular and instantaneous parallax.*
N_L and N_R are the nodal points of the left and right eyes, respectively (see text).

The stereoscopic appearance obtained by regarding two differently coloured, but otherwise identical, plane pictures with the two eyes separately, is probably due to chromatic differences of magnification. If the left eye, for example, views a plane picture through a red glass and the right eye views the same picture through a blue glass, an illusion of solidity results. Chromatic difference in magnification causes the images on the two retinas to be slightly different in size, so that the images of any point on the picture do not fall on corresponding points; the conditions for a stereoscopic illusion are thus present.

Retinal rivalry. Stereoscopic perception results from the presentation to the two eyes of different images of the same object; if two pictures that cannot possibly be related as two aspects of the same three-dimensional object are presented to the two eyes, single vision may, under some conditions, be obtained, but the phenomenon of retinal rivalry enters. Thus, if the letter *F* occupies one side of a stereogram, and *L* the other, the two letters can be fused by the eyes to give the letter *E;* the letters *F* and *L* cannot, however, by any stretch of the imagination be regarded as left and right aspects of a real object in space, so that the final percept is not three-dimensional, and, moreover, it is not a unitary percept in the sense used in this discussion; great difficulty is experienced in retaining the appearance of the letter *E,* the two separate images, *F* and *L,* tending to float apart. This is a mode of binocular vision that may be more appropriately called simultaneous perception; the two images are seen simultaneously, and it is by superimposition, rather than fusion, that the illusion of the letter *E* is created. More frequent than superimposition is the situation in which one or the other image is completely suppressed; thus, if the right eye views a vertical black bar and the left eye a horizontal one, the binocular percept is not that of a cross; usually the subject is aware of the vertical bar alone or the horizontal bar alone. Moreover, there is a fairly characteristic rhythm of suppression, or alternation of dominance, as it is called.

Ocular dominance. Retinal rivalry may be viewed as the competition of the retinal fields for attention; such a notion leads to the concept of ocular dominance—the condition when one retinal image habitually compels attention at the expense of the other. While there seems little doubt that a person may use one eye in preference to the other in acts requiring monocular vision—*e.g.,* in aiming a rifle—it seems doubtful whether, in the normal individual, ocular dominance is really an important factor in the final awareness of the two retinal images. Where the retinal images overlap, stereoscopic perception is possible and the two fields, in this region, are combined into a single three-dimensional percept. In the extreme temporal fields (*i.e.,* at the outside of the fields of vision), entirely different objects are seen by the two eyes, and the selection of what is to dominate the awareness at any moment depends largely on the interest it arouses; as a result, the complete field of view is filled in and one is not aware of what objects are seen by only one eye. Where the fields overlap, and different objects are seen by the two eyes—*e.g.,* on looking through a window the bars may obscure some objects as seen by one eye but not as seen by the other—the final percept is determined by the need to make something intelligible out of the combined fields. Thus, the left eye may see a chimney pot on a house, while the other eye sees the bar of a window in its place; the final perceptual pattern involves the simultaneous awareness of both the bar and the chimney pot because the retinal images have meaning only if both are present in consciousness. So long as the individual retinal images can be regarded as the visual tokens of an actual arrangement of objects, it is possible to obtain a single percept, and there seems no reason to suppose that the final percept will be greatly influenced by the dominance of one or other eye. When a single percept is impossible, retinal rivalry enters; this is essentially an alternation of awareness of the two fields—the subject apparently makes attempts to find something intelligible in the combined presentation by suppressing first one field and then the other—and certainly it would be incorrect to speak of ocular dominance as an absolute and invariable imposition of a single field on awareness, since this does not occur. Dominance, however, has a well-defined physiological meaning in so far as certain cells of the cerebral cortex may be activated exclusively by one eye, either because the other eye makes no neural connections with it or because the influence of the other eye is dominant.

The choice of what is perceived

Ocular dominance

Binocular brightness sensation. When the two eyes are presented with differently illuminated objects or surfaces some interesting phenomena emerge. Thus fusion may give rise to a sensation of lustre. In other instances, rivalry takes place, the one or other picture being suppressed, while in still others the brightness sensation is intermediate between those of the two pictures. This gives rise to the paradox whereby a monocularly viewed white surface appears brighter than when it is viewed binocularly in such a way that one eye views it directly and the other through a dark glass. In this second case the eyes are receiving more light, but because the sensation is determined by both eyes, the result is one that would be obtained were one eye to look at a less luminous surface.

ELECTROPHYSIOLOGY OF THE VISUAL CENTRES

To elucidate the functions of the various stages in the visual pathway, one must examine the responses to a retinal light-stimulus of the individual neurons at the different stages.

Ganglion cells. The main features of the responses of ganglion cells have already been touched upon. These have receptive fields that indicate a dual type of connection with the rods and cones, as indicated by the centre-periphery organization. A spot of light falling on receptors in the centre of this field may provoke a discharge in the ganglion cell or its optic nerve fibre; it is called an on-response and consists usually in an increase in the background discharge occurring in darkness. If a spot of light falls on a ring of retina surrounding this central region, the effect is one of inhibition of the background while the light is on, and as soon as it is switched off there is a pronounced discharge, the off-response. Other ganglion cells have been shown to have a directional sensitivity, responding to a moving spot of light only if this moves in a preferred direction and showing inhibition of background discharge if movement is in the null direction.

Geniculate neurons. In general, the lateral geniculate neuron is characterized by an accentuation of the centre-periphery arrangement, so that the two parts of the receptive field tend to cancel each other out completely when stimulated together, by contrast with the ganglion cell in which one or another would predominate. Thus, when the retina is illuminated uniformly there is little response in the geniculate cells because of this cancellation. This represents a useful elaboration of the messages from the retina because, to the animal, uniformity is uninteresting; it is the nonuniformity created by a contour or a moving object that is of interest, and the brain is therefore spared from being bombarded by unnecessary information that would result if every receptor response were transmitted to the brain.

Cortical neurons. When investigators made records of responses from neurons in area 17 there was an interesting change in the nature of the receptive fields; there was still the organization into excitatory (on) and inhibitory (off) zones, but these were linearly arranged, so that the best stimulus for evoking a response was a line, either white on black or black on white. When this line fell on the retina in a definite direction, and on a definite part of the retina, there was, say, an on-response, while if it fell on adjacent areas there was an off-response. Changing the orientation of the line by as little as 15° could completely abolish the responses. The simplest interpretation of this type of receptive field is based on the connection of the cortical cell with a set of geniculate cells with their receptive fields arranged linearly.

Variations in response

Eye dominance. Most of these units (*i.e.,* cortical cells plus connections) could be excited by a light stimulus falling on either eye, although there was usually dominance of one eye, in the sense that its response was greater; when both eyes were stimulated together, the effects summated. In general, then, when a large number of units are studied, a certain proportion are fired by one eye alone, others by the opposite eye alone, others by both eyes with dominance of one or other eye, while still others respond only when both eyes are stimulated. It is interesting that when kittens are deprived of the use of one eye from birth for several months, this deprived eye is virtually blind and the distribution of dominance in the cortical neurons is changed dramatically; if the left eye is deprived, the right hemispherical cortical neurons show a marked fall in dominance by the left eye, and an increase by the right eye. Thus, the ability of the eye to make use of cortical neurons is not fully developed at birth.

Cortical architecture. When an electrode is directed downward into the cortex it picks up responses in individual units at successive depths; units having the same directional sensitivity are arranged in columns so that the receptive fields of successive neurons are similarly oriented. When units were classified on the basis of eye dominance, a similar vertical distribution of units was found, overlapping with those based on directional preference. The columns for eye preference were about one millimetre wide, but those for directional preference were considerably finer. This columnar organization of cortical cells is not peculiar to the visual area.

Complex neurons. The cortical units (cells) described above, with receptive fields organized on a linear basis, have been called simple units in contrast to complex and hypercomplex units. Four types of complex units have been described; as with the simple units, the orientation of a slit stimulus (that is, a line) is of the utmost importance for obtaining maximal response, but unlike the situation with the simple unit, the position on the retina is unimportant. This type of unit makes abstractions of a higher order, responding to direction of orientation but not to position. It is this type of neuron that would be concerned, for example, with determining the verticality or horizontality of lines in space. Space does not permit of a description of the receptive field of a hypercomplex cell, but in general its features could be explained on the basis of connections with complex cells.

Stereoscopic vision. Of special interest is the behaviour of binocularly driven (stimulated) cortical cells, since their responses provide a clue to the fusion of retinal images. The cortical nerve cell receiving impulses emanating from both retinas must select those parts of the two retinal images that are the images of the same point on an object; second, for stereopsis, the nerve cell must assess the small displacements from exact symmetry that give the binocular parallax. In experiments, maximal response was often obtained only when the stimuli fell on disparate parts of the two retinas; these cortical cells were obviously disparity detectors, in contrast to others that gave maximal response when the stimuli fell on strictly symmetrically related parts of the two retinas—*i.e.,* on corresponding points. When successive units, during penetration of the electrode, were recorded, it was found that those requiring the same degree of disparity for maximal response were arranged in columns, as with direction sensitivity, so that, in effect, all these nerve cells were responding to stimuli in a strip of space at a definite distance from the fixation point. (H.Da.)

Eye diseases and visual disorders

The human visual apparatus includes the eyeball or globe, and its socket, or orbit, and auxiliary structures such as the lids and the muscles that control eye movement. These organs and their normal functioning are covered in detail above. This section briefly describes the more common diseases of the eye and its associated structures, and the methods used in examination and diagnosis; it also indicates the treatment and prognosis. The first part deals with conditions affecting the orbit, lids, and external eye, and the second with diseases of structures within the globe. Later sections deal with injuries, ocular conditions associated with general disease, disorders of vision, methods of examining the eye, and devices for correcting visual defects.

THE OUTER EYE AND AUXILIARY STRUCTURES

The orbit and lacrimal apparatus. The orbit is the bony cavity in the skull that contains the globe of the eye, the muscles that move the eye, the lacrimal gland, and the blood vessels and nerves required to supply these structures. The remaining space within the orbit is filled with a fatty pad that acts as a cushion for the eye and allows free movement of the globe. In old age this pad of fat tends to atrophy so that the globe recedes, causing the sunken appearance often seen in old people.

Inflammatory conditions of the orbit. As the bone that separates the orbit from the nose and the nasal sinuses is rather thin, infection sometimes spreads from the nasal sinuses into the orbit, causing the orbital tissue to swell and the eye to protrude. The condition is serious because of the possibility that the infection may spread into the cranial cavity along the pathways of the cranial nerves that enter the orbit to reach the eye. Infections can also spread to the cranial cavity by way of the blood vessels that lie in the upper part of the orbit, from lesions such as a boil on the skin of the lids or face in the neighbourhood of the eye. Large doses of an antibiotic such as penicillin, given immediately, in most cases eliminate such infections. The lacrimal glands, the small glands that secrete tears and are behind the outer part of each upper lid, are rarely inflamed but may become so as a complication of mumps. Inflammations of the lacrimal sac are much more common. The lacrimal or tear sac lies in a hollow at the corner of the eye in the front part of the nasal wall of the orbit; under normal conditions tears run along the edges of the lids toward the nose and are drained through two tiny holes connected by small tubes to the upper part of the lacrimal sac. The lower part of the sac is connected to the nose by a duct, the nasolacrimal duct, and infection may ascend this passage from the nose and cause an acute painful swelling at the inner corner of the eye. Blockage of the nasolacrimal duct prevents the passage of tears into the nose and results in a watering eye. Such a blockage, which is nearly always accompanied by chronic inflammation in the lacrimal sac, is usually best treated by an operation in which a new opening from the lacrimal sac to the nasal cavity is made.

Infections of nasal origin

Tumours of the orbit. Tumours in this area are comparatively rare, the most common being a tumour of the lacrimal gland. If the tumour is behind the globe in the optic nerve, it will cause a slow and gradual protrusion of the eye; such an abnormal position of the eye may prevent ocular movements from being coordinated with those of the normal eye, and the images of the two eyes, which are normally fused, may separate and give rise to double vision (or diplopia).

The lids. *Inflammatory conditions.* The chronic inflammation of the lid margins known as blepharitis is a common and distressing condition. The inflammation may be mild and consist simply of redness of the lid margin with scaling of the skin, or more severe, affecting the follicles of the eyelashes and leading to their destruction and distortion. Both types tend to be associated with greasiness of the skin and dandruff. The skin of the lids is particularly sensitive to allergic processes, and itching and scaling of the lids is a common reaction to drugs or cosmetics applied to the eye of a sensitized person.

Another common inflammatory condition of the lid is a sty; *i.e.,* an infection of a lash follicle—the sheath of the eyelash root—corresponding to a boil on the skin elsewhere. It starts as a painful swelling of the whole lid, so that at first it may be difficult to find a localized lesion; but soon one area becomes more swollen and, as pus forms, a yellow point associated with an eyelash can be seen near the lid margin. A rather similar appearance can be produced by an inflammation of the meibomian glands; *i.e.,* tiny glands in the thickness of the lid, opening on the lid margin. As the glands are embedded in tough fibrous tissue, the pain and reaction may be more severe than in an ordinary sty. Examination of the internal surface of the lid will show a red velvety area with a central yellow spot through which pus will later discharge. Sometimes the meibomian glands suffer from a chronic infection, and a painless firm lump appears in the lid and slowly increases in size. The skin can be moved freely over the surface of the lump, showing that the latter is in the deeper tissue of the lid. The inner surface of the lid will show a grayish area surrounded by a little inflammation. The lesion is treated by making an incision on the inner surface of the lid, and scraping out the contents.

Herpes zoster (shingles) may affect the skin of the eyelids and is of particular importance because the cornea (the transparent covering of the front of the eyeball) and the inner eye may also be affected. The condition starts with pain and redness of one part of the forehead and the eyelids of the same side. Vesicles, or small blisters, form later in the affected area. The pain may be severe, and some constitutional disturbance is usual.

Ectropion and entropion

Displacements of the lid. Malpositioning of the lid is common in elderly people, and although not serious in itself gives rise to considerable discomfort and irritation. The commonest condition is called ectropion, in which the lower lid falls away from the globe so that the tears overflow the lid. This constant wetting of the skin of the lower lid excoriates the skin and causes it to retract, which in turn increases the tendency of the lid to turn out. In the early stages, massage of the lid and the use of a bland ointment on the skin may help, but usually some plastic procedure is necessary to bring the lid back into its normal position.

The opposite condition is entropion, in which the lid turns inward and the lashes cause much irritation by rubbing on the eye. Unlike ectropion, it may affect either the upper or the lower lid. It may be caused by scarring of the deeper tissues of the lid following infection, or may be due to senile changes in muscle tone in which a band of fibres of the circular muscle surrounding the lids contracts more strongly than the peripheral fibres, thus tending to turn the lid inward. Surgical treatment is required to restore the lid to its normal position.

Tumours. Benign overgrowths of the blood vessels, called hemangiomas, may occur in the lids and give rise to soft bluish swellings that can be reduced by pressure over them. They are present at birth and tend to grow rapidly in the first few years of life. Often they disappear spontaneously, but they can be treated by surgical removal or the insertion of radioactive material. Simple overgrowths of skin, called papillomas, are common along the lid margin but require no special treatment except excision for cosmetic reasons.

The lids and the skin of the nose near the inner margins of the lids are common sites for the development of skin cancer in older people; the most usual type, called a rodent ulcer, starts as a small nodule in the skin that gradually enlarges and breaks down to form an ulcer with a hard base and rolled edges. Bleeding may occur from the base of the ulcer. Although rodent ulcers are malignant in the sense that they destroy tissue locally, they do not spread to distant areas of the body by means of the lymph system or the blood vessels. A more serious malignant tumour of the lids is a carcinoma, which develops as a more irregular ulcer of the lid and may spread to involve the underlying bone. If left untreated the growth may spread to the lymph nodes in front of the ear or under the jaw.

Development of binocular vision; definition of squint

Squint. In the lower vertebrates such as the fish the eyes are situated on either side of the head, to give the maximum view of the surroundings and an early warning of the presence of predators. The field of vision of each eye is separate except for a narrow sector immediately in front of the animal, where the two visual fields overlap. From the evolutionary point of view the improved judgment of distance obtained by viewing an object with both eyes conferred considerable biological advantage in the struggle for survival. In the higher animals, particularly the predatory species of birds and mammals, binocular vision became more and more important and structural changes in the placement of the eyes in the head permitted a larger overlap of the two visual fields, until the situation was reached in the higher mammals in which the visual axes—that is, the line of direct sight—became parallel. This desirable visual function has been fully attained in man. The structural changes necessary to bring this about have, however, lagged behind the function, and the geometrical axes of most eyes are still slightly divergent (*i.e.,* the two eyes at rest are directed slightly away from the nose). The bony structure of the orbit has lagged even further behind, so that the axes of the two orbits make an angle of about 45°.

It is in fact the function of using two eyes together that keeps the optic axes straight in a normal person. If, for example, one eye becomes blind, it tends to revert to an anatomical position of rest in line with the axis of the orbit. A blind eye will therefore appear to be diverging. The visual axes can remain straight only if each eye has reasonably good vision, the ocular muscles can move the eyes in the required direction of gaze, and the complex neuromuscular reflexes required to coordinate the movements of the two eyes are intact. Failure to maintain the visual axes parallel may therefore result from a visual defect in one or both eyes, a muscular defect resulting in loss of normal movement of the eye, or a defect in the central nervous system involving the coordinating nervous pathways. A true squint is a condition in which the visual axes are no longer parallel. An apparent squint may be seen in children as the result of prominent skin folds in the inner sides of the eyes, which make the eyes appear to be converging (*i.e.,* appear to cross). These skin folds usually disappear when the bony structure of the nose develops more fully.

Clinically, squints are divided into concomitant, in which the abnormal angle between the visual axes remains constant in all positions of gaze, and paralytic, in which the angle of squint varies with the direction of gaze. The commonest type of squint is the convergent concomitant type seen in small children (*i.e.,* the children are consistently or intermittently cross-eyed). It is usually first noticed between the ages of one and two and may be precipitated by a systemic disease such as measles. There is often a family history of squint. Children of this age are particularly interested in objects close to them, and in order to view an object clearly at close range two things are necessary: first, the visual axes must converge, so that both eyes can view the same object; and second, the focus of the eye must be adjusted for near vision. The link between convergence of the eyes and focussing, or accommodation, is very strong, and normally the two actions work in harmony. Most small children, however, are long-sighted, which means that, in order to see clearly close to, they have to exert an extra amount of accommodative effort. As accommodation and convergence are closely linked, the extra effort of accommodation tends to produce an overconvergence; but, provided that the visual acuity of each eye is normal and the motor control of the eyes is normal, this tendency is controlled. If the vision of one eye is reduced, for example by disease or an error of refraction, binocular vision breaks down and overconvergence occurs.

Double vision; suppression of one image

Once parallelism of the visual axes is lost, the image of objects no longer lies on a familiar area of retina, and instead of the images from the two eyes being fused into one, two images are perceived. This condition of double vision, or diplopia, is intolerable to the child, who reacts by "suppressing" the image from the squinting eye. If the suppression is allowed to continue, the central vision of the affected eye drops rapidly to a low level, so that even if the original disturbance that started the squint is corrected, this loss of vision, or amblyopia, of the squinting eye will prevent the restoration of normal binocular vision and thus perpetuate the squint. The longer the suppression is allowed to continue, the less likely is the child to regain normal vision in the squinting eye. Covering the good eye will usually encourage the recovery of the suppressed vision but must be started as soon as the squint is noticed. Any refractive error present—any defect that prevents light rays from focussing properly on the retina—must be corrected by glasses, and retraining of the binocular reflexes can be aided by special exercises. Early treatment on these lines may be all that is necessary, but if the visual axes are still abnormal surgery of the extraocular muscles will be required to correct the deviation.

Paralysis of one of the muscles that control the movement of the eyes results in limitation of movement of the globe in the direction of action of the muscle, with the result that double vision with separation of the images occurs on attempts to move the eye in this direction.

As earlier stated, accommodation and convergence are normally perfectly linked together so that the movements of the two eyes bring the visual axes to the point of focus. In many people this balance is not quite perfect and the eyes tend to converge or diverge too much for a given distance—a condition known as heterophoria.

The conjunctiva. The marine origin of the human species is betrayed by the need for the anterior surface of the eye to be bathed in salt water. A thin membrane lines the lids and covers the anterior surface of the globe, forming a sac, the conjunctival sac, the contents of which are lubricated by the tear glands. This warm, moist habitat provides a suitable environment for the growth of bacteria and other organisms, leading to conjunctivitis, inflammation of the conjunctiva. Bacterial conjunctivitis starts with a feeling of grittiness and discomfort, the eye is red and there is a discharge from it. The discharge is particularly noticeable after sleep, when the lids may be stuck together by the exudate on the lashes. Vision is not affected except by the strands of mucus, which can be blinked away from the cornea. Antibiotic drops usually clear the condition in a few days. Vernal conjunctivitis or spring catarrh is, as its name suggests, an allergic condition occurring in the early summer; it is more common in young people and probably results from sensitivity to external irritants such as dust and pollen. It usually responds well to treatment with drops of corticosteroid hormone.

Chronic conjunctivitis, causing a gritty feeling, with redness of the eyes and a slight mucoid discharge, is a common condition, the cause of which may be difficult to find. Often there is an infective element such as a chronic inflammation of the lid margin, and sometimes the condition is allergic and may result from sensitivity to cosmetics or to drugs applied to the eye. An unsuspected foreign body or a deficiency of tear secretion may cause similar symptoms.

Viral conjunctivitis. With the enormous increase in the use of antibiotics since the 1940s, bacterial infections in general are becoming less common. This is also true of infections of the eye, and in most western countries bacterial conjuctivitis is now less common than viral infection. Viruses tend to attack the cornea as well as the conjunctiva; the infection is contagious and may be responsible for outbreaks of epidemic keratoconjunctivitis (inflammation of the cornea and the conjunctiva). The onset is acute, with redness and swelling of the eye and lids and a tender swelling of the lymph node in front of the ear.

Trachoma

Trachoma can truly be described as one of the scourges of mankind. Although rare in England and North America, it is the largest single cause of blindness in the world as a whole. Widespread in some Middle East countries, it has become more common in Asia, India, Central and South America, and Africa. It occurs sporadically in southern and eastern Europe. The agent responsible has now been isolated and shown to belong to the group of organisms known as chlamydiae. They occupy a taxonomic position between bacteria and true viruses and unlike the latter are susceptible to treatment with sulfonamides and some antibiotics. The disease is contagious and thrives where populations are crowded together in poor hygienic surroundings. Shortage of water for washing, and the myriads of flies attracted to human waste, aid the dissemination of the disease. In some ways trachoma is more of a social than a medical problem; if living standards can be improved, overcrowding reduced, flies discouraged, and adequate water supplies ensured, the incidence of trachoma decreases rapidly.

The early symptoms of infection are pain, watering of the eye, and sensitivity to light. At this stage the conjunctival lining of the lids is red and velvety in appearance, and the cornea shows gray areas. Later, the conjunctiva appears to have grains of sand embedded in its tissue. Blood vessels grow into the cornea, which becomes thickened and hazy. Secondary bacterial infections are common, but the real dangers of trachoma lie in the scarring and contracture of tissue that occur when healing takes place. These changes affect the upper lid in particular, causing it to buckle inward so that the lashes rub across the already diseased cornea, and it is the corneal scarring thus produced that can cause blindness.

Degenerative conditions. Exposure to wind and dust frequently causes degenerative changes in the exposed part of the conjunctiva, particularly in older people. A yellow nodule forms, first on the nasal side of the cornea and later on the other side. It is without blood vessels and is frequently unnoticed until an incidental conjunctivitis causes it to stand out clearly against the red background of dilated conjunctival vessels. It causes no symptoms and requires no treatment.

A more serious degeneration is that known as a pterygium, found particularly in people who live in hot, dusty climates. It appears as a fleshy growth at the edge of the cornea, with a tendency to progress across its front surface, where it may interfere with vision. Treatment consists of surgical removal, but recurrences are common.

The cornea and sclera. The cornea is the clear window of the eye and its most important refractive surface. Any surface irregularity, any scar in the substance of the cornea, is likely to have a profound effect on vision. Almost the whole nerve supply of the cornea consists of nerve fibres sensitive to pain, so that corneal diseases are always painful and elicit a flow of tears by a reflex action that is part of the protective system of the eye.

Infection of cornea with herpes virus; bacteria; fungi

Inflammation of the cornea. As with inflammations of the conjunctiva, bacterial infection of the cornea has become much less common and viral infections are increasingly important. Of these, the herpes viruses, which cause the common "cold sore" of the lips and skin and the venereal form of herpes, are the most frequent cause of corneal ulceration. Infection is spread by personal contact. The herpes virus causes a typical ulcer of the cornea called, from the pattern of the lesion, dendritic ("branching"). The disease starts with an acutely painful eye, with tearing, and sensitivity to light. The ulcer may heal spontaneously or after medical treatment, but the virus often lies dormant in the tissues; recurrences are common, and with each recurrence there is more danger that the virus will extend deeper into the cornea and cause an intractable inflammation.

Application of the drug deoxyuridine (5-iodo-2-deoxyuridine) to the cornea causes the ulcer to heal more rapidly and reduces the recurrence rate. The action of the drug depends upon its limiting the multiplication of the virus by interfering with the formation of virus deoxyribonucleic acid (DNA) in the host cell.

Bacterial infections of the cornea still occur, usually after injury to the corneal surface, as few bacteria have the power to penetrate the intact surface layers of the cornea. Such ulcers may be extremely severe, and there is always a danger of perforation of the eye, particularly in debilitated patients.

The spores of fungi are commonly present in the atmosphere. The normal cornea is resistant to infection by these organisms, but a fungal infection of the cornea can develop after a corneal injury or other lesion, particularly if corticosteroid drugs have been used in treatment. Intensive treatment with antifungal drugs is usually effective in killing the organisms, but a dense scar is usually left.

A corneal inflammation may start in the deeper layers, usually by spread of infection from the bloodstream. It is seen most commonly in adolescents who have congenital syphilis. Both eyes are usually attacked, although there may be an interval before the second eye is affected. The cornea rapidly becomes hazy, and blood vessels grow in from the surrounding tissues to form a red patch. With the decline in congenital syphilis in developed countries, the condition is now becoming a rarity.

Inflammation of the sclera. The sclera is the fibrous covering of the eye that shows up as a dense white layer beneath the transparent conjunctiva. A relatively mild nodular inflammation sometimes occurs in the superficial layers of the sclera; it is thought to be allergic in nature and usually responds well to anti-inflammatory treatment. Inflammation of the deeper sclera is more severe and often is painful. It occurs more frequently in older people and may be associated with tuberculosis or rheumatism; however, the cause of the condition is often not discovered.

Degenerative conditions. Keratoconus is the name of a curious condition in which the central part of the cornea, normally spherical in shape, begins to bulge and protrude forward as a cone. The only symptom is deterioration of vision due to the irregular astigmatism caused by the changing corneal curvature. Ordinary spectacles cannot correct the irregular refraction, but contact lenses are often of great value, and in more advanced cases corneal grafting is required.

There are numerous other rare types of corneal degeneration, some of which are familial; all produce a deterioration in vision that cannot be corrected with spectacles. Many of these conditions respond well to corneal grafting.

THE INNER EYE

The uveal tract. The uveal tract is a vascular layer of tissue—that is, a layer rich in blood vessels—lying next to the inner surface of the sclera. It is divided into three structures: the choroid, a highly vascular layer that supplies blood to the outer layers of the retina; the ciliary body, largely muscle tissue, which by its contraction and relaxation alters the focussing of the lens; and the iris, the coloured part of the eye, which forms the variable aperture of the eye, the pupil. The ciliary body, which lies at the base of the iris, also functions by forming the aqueous humour, the production and drainage of which regulate intraocular pressure; the aqueous humour also is the source of nutriment to the lens and cornea, which are avascular (without blood vessels).

Inflammation. Inflammations of the uveal tract are always potentially serious because of the secondary effects they may have on the retina and the lens. In most cases the disease affects either the anterior part of the uvea—that is, the iris and ciliary body—or the posterior part, the choroid. An attack of acute anterior uveitis starts with pain, redness, and mistiness of vision. The eye is sensitive to light, and, although there is no discharge as in conjunctivitis, the eye may water. The pupil tends to contract and the normally clear iris markings become less distinct. In chronic anterior uveitis the main symptom is blurring of vision. Acute choroiditis starts with sudden onset of blurring of vision with many black spots floating about in front of the sight.

Except for cases in which the uveitis follows a perforating injury or a corneal ulcer, it is believed that the inflammation is caused by an infective process within the body or by some other mechanism associated with systemic disease. Many infective conditions and parasitic diseases are known to cause uveitis. In a large proportion of cases, however, particularly when the inflammation is confined to the anterior segment, it proves impossible to be sure of the cause in any individual instance, and the investigation of a case of uveitis often poses one of the biggest problems in ophthalmology today. In men, a proportion of cases of anterior uveitis are associated with ankylosing spondylitis, a chronic disease of the joints of the spine, the cause of which is still obscure. Another association, again in men, is with Reiter's disease, a condition that starts as an infection of the urogenital tract with the later development of joint changes, particularly in the sacroiliac joints of the lower back, and recurrent attacks of anterior uveitis. The organism that is responsible for this venereally contracted infection is still unknown, but may be a virus. Infections in the teeth and tonsils have long been held to be a cause of uveitis, and eradication of dental decay does occasionally have a favourable effect on the course of the disease.

Toxoplasmosis and its effect on fetus

Inflammations of the choroid—the posterior portion of the uveal tract—and the retina are more likely to be infective in origin. One of the organisms most commonly involved is *Toxoplasma gondii,* a protozoon of worldwide distribution among domestic animals, small mammals, and man. Although antibodies to the organism can be found in a high proportion of most populations, showing that infection is widespread, overt signs of disease are rarely seen; most people can acquire the infection without being aware of any systemic disturbance at all, and only in special circumstances does the organism cause disease. One of these special circumstances is pregnancy. If a woman becomes infected during pregnancy there is a short period in which invasion of the tissues takes place before circulating antibodies are formed by the mother. During this period it is possible for the organism to pass through the placenta and infect the unborn child. Fetuses appear to be particularly susceptible to the organism, nearly half of those exposed showing some evidence of infection with toxoplasmosis. In severe cases the child may be stillborn or may be born with congenital toxoplasmosis, a serious disease affecting many organs of the body and particularly the brain and eye. In less serious cases small foci of infection are left in the nervous system and the retina of the eye; these may not be apparent at birth and may remain quiescent, only to become active 15 or 20 years later in the form of an inflammation of the choroid and the retina. Children of subsequent pregnancies are unaffected.

The treatment of uveitis has been transformed by the advent of corticosteroid drugs. Even when a specific infection cannot be discovered and treated with the appropriate specific drug, therapy with corticosteroids is usually successful in controlling the worst ravages of the inflammation.

Tumours of the uveal tract. Pigmented tumours are the commonest tumours involving the uveal tract. They may be benign (the nevus or mole) or malignant. The choroid is the commonest site of these lesions, which push the retina forward and cause a retinal detachment. Disturbances of vision are the commonest symptom, but the tumour if neglected may enlarge and cause inflammation and raised pressure within the eye. Small portions of the tumor often enter the bloodstream and settle in distant organs, particularly the liver. The growth of these secondary deposits is often slow; they may not be apparent until many years after the diagnosis of the tumour in the eye.

The lens. The lens is a transparent, avascular organ surrounded by an elastic capsule. It lies behind the pupil and is suspended from the ciliary body by a series of fine ligaments. Its transparency is the result of the regular arrangement of the lens fibres; since these are being formed continuously, the lens continues to grow throughout life. Interference with this growth pattern will result in the formation of abnormal lens fibres that cannot transmit light as well as the normal lens fibres. A small opacity is thus seen in the lens. Minor irregularities are common in otherwise perfectly normal eyes. If the opacity is severe enough to affect vision it is called a cataract.

Cataract in newborn child; in adult

Congenital lens opacities of many varieties have been recognized and described since the early days of ophthalmology, but they remained curiosities until the work of an Australian ophthalmologist, Norman M. Gregg, threw new light on their cause, and, indeed, on that of many other congenital defects. In 1941 Gregg noticed that after an epidemic of German measles (rubella) many of the children whose mothers had contracted the disease in the first two months of pregnancy were born with cataract, sometimes associated with deafness and congenital heart disease. It is now known that the virus can be recovered from the lens for several months after birth.

Cataract in the adult may be the result of injury to the lens by a perforating wound, by exposure to radiation such as X-rays, or as the result of the ingestion of toxic substances or even of some drugs. The lens relies for its nutrition on the aqueous humour secreted by the ciliary body and, if the latter is severely damaged as the result of long-continued uveitis or a tumour, the metabolism of the lens suffers and a cataract develops. The commonest form of cataract is senile cataract, so called because it becomes progressively more common with advancing age. In spite of a large amount of work on the biological and biochemical changes that take place in the lens, the underlying cause of senile cataract is still unknown. Whatever the underlying biochemical changes may be, they result in an increasing clouding of the lens until the whole lens loses its normal transparency and becomes white and opaque. The only symptom is progressive diminution of vision. In the early stages of the condition some visual improvement can usually be obtained with spectacles, but, as the cataract progresses, the visual deterioration becomes sufficiently severe to warrant surgical treatment.

With modern techniques cataract extraction can be done as soon as the visual deterioration interferes with normal activities, and it is no longer necessary for patients to wait for many years in semiblindness to allow the cataract to become mature. Cataract extraction is one of the most successful and satisfying operations in ophthalmic surgery; if the eye is otherwise normal the visual results are excellent, although the refractive power of the lens has to be replaced by a rather thick spectacle lens or special contact lens.

The retina. Developmentally, the retina is part of the brain and as such has only a limited capacity for repair of its damaged tissue. In particular, the highly specialized rods and cones (the photoreceptors), which are the structures sensitive to light, and the nerve cells of the retina, like those of the brain, cannot be replaced if they are damaged. Death of these cells inevitably has a permanent effect on vision.

The retina is a thin transparent membrane that lines the inner eye. Its outermost layer, the pigment epithelium, is formed of pigmented cells that are closely adherent to the underlying blood vessels of the choroid. The layer of rods and cones is more loosely attached to the pigment epithelium and has complicated nervous networks that culminate in the innermost layer of nerve fibres. These fibres run back through the optic nerve to the brain. The inner two-thirds of the retina derives its blood supply from a special complex of vessels that enters the eye through the optic nerve.

Detachment of retina

Retinal detachment. A retinal detachment is a condition in which the main part of the retina becomes separated from the pigment epithelium. This may follow an injury to the eye or a tumour; or inflammation of the underlying choroid. The commonest type of detachment, however, has no such predisposing factors: the distinctive feature is the formation of a small hole or tear in the retina, usually at the periphery where the retina is thinner. In most cases the hole is caused by an adhesion forming between the retina and the jelly-like substance called the vitreous humour that fills the interior of the eye. Sudden movement of the eye, or an injury, causes the vitreous to pull on the retina, thus creating a tear or hole. When this has happened, fluid can pass through the hole and strip the retina off the pigment epithelium. Myopic (nearsighted) eyes are particularly prone to retinal detachment because they are larger than normal, and the coats of the eye are thinned and stretched. The periphery of the retina, in particular, often shows weak areas, and the vitreous is usually unduly thin and fluid.

The history is often quite typical. The pull of the vitreous on part of the retina creates a sensation of light noticed by the person affected as flashes that occur on movement of the eye. When an actual tear has developed, the retina starts to become detached and the person has the sensation of a shadow coming down over the vision.

The essential factor in treatment is to seal off the hole in the retina. The part of the retina containing the hole must be brought into close contact with the choroid and then by means of a gentle inflammatory reaction caused by using heat, cold, or intense light, the retina is made to stick to the underlying choroid and seal off the leak. The remaining fluid can then be drained away, allowing the retina to fall back into place. Provided that the detachment has not been of long standing, the retinal function recovers quite well once the retina is reattached. The small central area of retina, however, that subserves the most acute vision has only one source of blood supply, the underlying choroid; once it is separated from this some permanent damage ensues, even if the retina is subsequently replaced in its correct position. Thus, it is most important therefore that retinal detachments be treated early, before the central area of the retina becomes detached.

Retinal degeneration. Cases of retinal degeneration can be grouped in two broad classes: hereditary and genetic, and senile. A large number of genetically determined degenerations of the retina have been described. Although they are quite rare, the bizarre appearances of the retina and the inexorable advance of the disease have excited considerable interest among ophthalmologists. These conditions are typified by the disease known as retinitis pigmentosa, a hereditary condition. The earliest symptom is night blindness, which may first be noticed in childhood and is due to alteration in the function of the rods, which are the visual receptors used in dim light. The more peripheral parts of the retina are affected first, and while central vision may be good the field of vision progressively decreases until only a small "tubular field" remains. Cause of the disease is unknown. It is easily recognizable by the narrowing of retinal vessels and the scattering of clumps of pigment throughout the retina.

In senile degeneration, unlike the hereditary type, it is the central part of the vision that is first affected. The central part of the retina, known as the macula, derives its blood supply only from the choroid, and it is probably for this reason that it is likely to suffer first from the slowing of the metabolic changes and from the deficiency of circulation that occur in old age. While degeneration of the macula does not cause blindness, in the sense that the person affected is unable to see anything, it is extremely disturbing because it affects central visual acuity and makes reading or fine work difficult or impossible. There is as yet no satisfactory medical or surgical treatment, but considerable improvement can be obtained by the use of special magnifying spectacles.

The retinal changes that may occur in diabetes, arteriosclerosis, and vascular hypertension are described in a later section.

The optic nerve. The optic nerve, which carries about one million nerve fibres, leaves the globe from the back of the eye and passes through the apex of the orbit into the cranial cavity. It is surrounded by an extension of the membranes that surround the brain and this connection with the intracranial cavity is of some importance, because in some intracranial diseases the pressure within the skull rises and is transmitted along the sheaths of the optic nerve to cause swelling of the optic nerve head, which is visible inside the eye. This swelling of the nerve head, or papilledema, is one of the most important signs of increased intracranial pressure. If the swelling persists, damage to the fibres of the optic nerve takes place, with subsequent loss of vision.

Causes and effects of swelling of optic nerve head

Swelling of the optic nerve may also be caused by inflammatory changes in the nerve itself or in the surrounding sheath; this condition is known as optic neuritis. The symptoms are loss of vision in the central part of the visual field and pain on moving the eye. The condition is most common in young adults and may be due to the spread of infection from the adjacent nasal sinuses. The majority of cases, however, are manifestations of multiple sclerosis, a condition in which the sheath of the nerves becomes altered and interferes with the transference of nervous impulses. This may occur in any part of the nervous system, but the optic nerve is a common site, and the lesion is often the first to be noticed by the patient because of the visual symptoms that result from it. The disease is characterized by long periods of remission from symptoms, and after optic neuritis it may be 10 years or more before other signs are apparent. Usually the function of the optic nerve recovers after an attack of optic neuritis, leaving little, if any, visual disturbance, but there is some atrophy of the fibres.

Optic atrophy may follow any serious disease of the retina involving a large amount of destruction of neural tissue. It may also follow damage to the optic nerve within the skull, or the optic chiasma—that is, the place where the optic nerves crisscross, close to the pituitary gland. Tumours of the pituitary gland nearly always compress the optic nerve fibres and cause some degree of atrophy with loss of vision in that part of the visual field subserved by the fibres concerned. Usually it is the fibres on the inner side of the optic nerve and those that cross at the chiasma that are most involved: these fibres supply the retina on the nasal half. This part of the retina receives visual images from the outer part of the visual field, and in pituitary lesions it is common to find that the outer parts of both visual fields are affected.

Certain chemicals and some drugs can also cause optic atrophy: among them are quinine and methyl alcohol. Optic atrophy is most unlikely to follow normal medical doses of quinine, and when it occurs it is usually from the large doses taken to cause abortion. Methyl alcohol (wood spirit or methylated spirits) is broken down in the body to acetyl aldehyde, which is toxic to neural tissue, and the risks of blindness from drinking methylated spirits are high.

Glaucoma. The thin coats of the eye are not sufficiently rigid in themselves to withstand distortion following the pull of the extraocular muscles when the eye is moved. The eyeball is kept rigid by the action of the ciliary body, which secretes sufficient amounts of the fluid called the aqueous humour to pump up the pressure of the eye to a level above the atmospheric pressure. This fluid is constantly being formed and drains away at the base of the iris through specialized drainage channels. Should these channels become blocked the pressure within the eye rises to abnormally high levels and impedes the entry of blood into the eye. The fibres of the optic nerve where it enters the eye are particularly susceptible to a reduction in blood supply, and if the intraocular pressure remains raised for long some of these nerve fibres will atrophy, causing loss of function of the retina from which they are derived. Glaucoma is the name given to a condition in which the intraocular pressure is raised to abnormal levels. In some persons this is due to other disease within the eye—such as inflammation or a tumour—but most have one of two distinct diseases, chronic simple glaucoma or closed-angle glaucoma.

Chronic simple glaucoma

Chronic simple glaucoma is a common disease that may affect one percent of people in the older age groups. Although the actual cause is not known it is almost certainly due to degenerative changes in the outflow channels for aqueous fluid. It is rare below the age of 40 but after this its incidence increases; in one recent survey it was found to affect 10 percent of those examined over the age of 80. Genetic influences are important and relatives of patients with glaucoma are five times more likely than others to develop the disease.

The symptoms are slight or absent in the early stages. The slow rise in pressure does not cause pain, and the early visual loss is in the peripheral parts of the visual field, affecting central vision only late in the disease. Both eyes are usually involved, although one may be more severely affected than the other. Since vision lost from glaucoma cannot be restored, successful treatment can only prevent further loss of vision. It is of great importance, therefore, that the disease be diagnosed as early as possible. Measurement of the intraocular pressure is of great value in the diagnosis of glaucoma: this is a simple test that can be applied as a screening method for surveys of the normal population.

From H.G. Scheie, *Medical Ophthalmology: Ophthalmalogic Manifestations of Systemic Diseases*

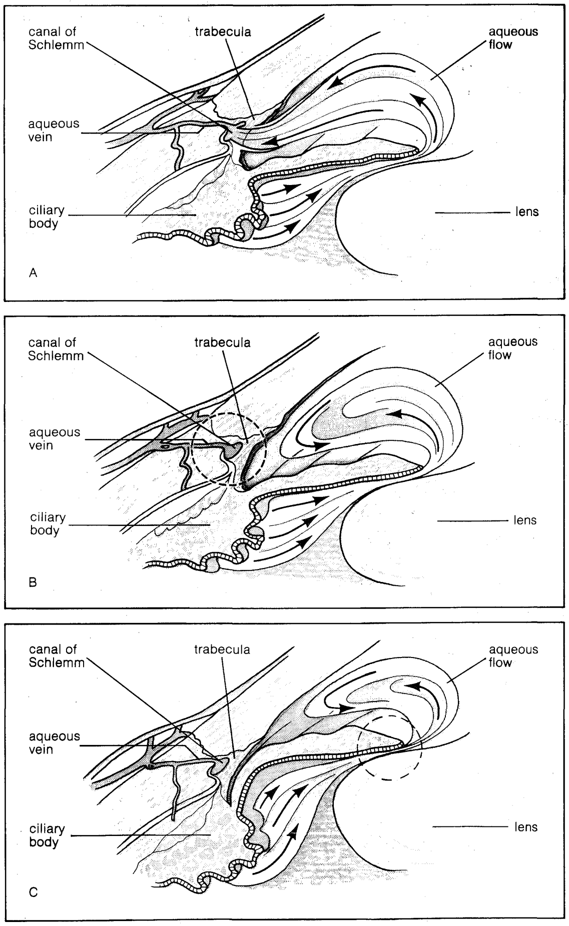


Figure 51: *Normal flow of aqueous humour contrasted with two types of obstruction.*
(A) Normal flow of aqueous humour. (B) Obstruction to flow of aqueous humour in chronic glaucoma. (C) Obstruction to flow of aqueous humour in closed-angle, or acute, glaucoma. Dotted circles indicate site of obstruction.

The medical treatment of chronic simple glaucoma consists of the use of drops that lower the intraocular pressure. Inhibitors of the enzyme carbonic anhydrase, when taken by mouth, reduce the formation of aqueous humour and are used as an additional measure when necessary. If the pressure remains raised in spite of all medical treatment, then surgical methods must be used to increase the drainage of fluid from the eye.

The other common type of glaucoma is called closed-angle glaucoma. This again has a familial incidence and occurs in people who have a rather small, long-sighted eye. Continued growth of the lens in these patients pushes the iris forward and narrows the gap at the root (the outer edge) of the iris where aqueous humour flows out of the eye. This fluid is formed in the ciliary body behind the iris and flows forward through the pupil to the angle of the anterior chamber. The lens pushing against the iris acts as a valve and impedes the flow of aqueous humour through the pupil. The root of the iris, which is rather thin, is then pushed forward and may eventually completely close the exit for aqueous humour, so that the intraocular pressure rises rapidly. The eye becomes painful and the vision is lost; the pain may be so severe as to cause vomiting and prostration. The eye becomes red and stony hard to the touch. Urgent treatment is required to lower the pressure and prevent strangulation of the blood vessels entering the eye.

In some cases an acute attack such as this heralds the onset of the disease, but more frequently minor, subacute, attacks, which are relieved by rest and sleep, occur for months or years. Modern methods of medical treatment are usually effective in lowering the pressure in the acute attack; an operation is usually necessary to prevent further recurrences.

OCULAR INJURIES

The bony orbit provides excellent protection for the eye from blunt injuries. A blow from in front with a rounded instrument such as a fist or tennis ball, however, can cause a shock wave to travel through the eye and damage the retina at the back of the eye. Central vision may be reduced after such injuries without any very obvious changes in the appearance of the eye. In severe cases the bones of the orbit may be fractured. Perforating wounds from glass, sharp metal fragments, and so on, are always serious. Injuries to the lens will result in the formation of a cataract, and often after penetrating injuries the eye remains inflamed for a considerable time.

Sympathetic ophthalmitis

One type of inflammation following injury, sympathetic ophthalmitis, is of particular importance; fortunately it is now rarely seen. The sequence of events is that an injured eye remains irritable and after some weeks, months, or even years, the fellow—previously normal—eye may take part in the inflammation. This is the "sympathizing eye." The cause of sympathetic ophthalmitis is not known, but it is known that if an injured eye is removed within 10 days sympathetic ophthalmitis never occurs in the other eye. In the past there was little effective treatment for the condition, but therapy with corticosteroid hormones has proved effective in controlling the inflammation in most cases, so that even if the disease becomes established the consequences are not as serious as they were previously.

Foreign bodies. Most foreign bodies that enter the eye remain near the surface. When they touch the cornea they cause intense pain and a flow of tears. The tears may be sufficient to wash the foreign body out of the eye, but if it becomes embedded in the cornea it may have to be removed surgically. Many small foreign bodies lodge in the under surface of the upper lid so that every time the eye blinks the foreign body rubs on the cornea, causing pain and irritation.

Small foreign bodies travelling at high speeds may penetrate the interior of the eye with remarkably few symptoms, and their presence may not be recognized until weeks or months later when inflammatory changes occur. The commonest foreign bodies to enter the eye in this way are fragments of metal from hammer-and-chisel accidents, or from moving parts of machinery. Whenever such injuries are suspected, it is important to locate the position of the fragment as carefully as possible, and to remove it by surgery. If the foreign body is magnetic, a large electromagnet is invaluable in attracting the foreign body to the site of the incision in the eye, making extraction comparatively simple.

Chemical and radiation injuries. Strong acids and alkalis always cause severe injury if they enter the eye. Speed is the vital factor in first-aid treatment, and copious irrigation with water the first essential. Delay of first-aid treatment in the hope of finding a neutralizing substance is a serious error, as strong acids and alkalis quickly become bound to the ocular tissues and cause severe necrosis (death of tissue).

Effects of ultraviolet and infrared rays

Except for extremely intense light such as that from a laser, the visible wavelengths of the electromagnetic spectrum—*i.e.,* visible light rays—rarely cause ocular injury. Ultraviolet light, however, is strongly absorbed by the cornea and is the cause of the not uncommon condition known as snow blindness. Symptoms, consisting of intense pain and copious flow of tears, may not occur until some time after exposure. Exposure to light is painful. The treatment consists of cold compresses to the eye and soothing lotions; usually the eyes recover without any permanent damage.

Long-continued exposure to infrared radiation, without adequate protection for the eyes, can cause cataract formation. The lens is also susceptible to X-rays, and the eyes must be shielded when therapeutic irradiation is used for growths around or near the eye. High-intensity microwaves such as those used in some military applications can also cause ocular damage. The widespread use of lasers in research departments and in industry has created a new ocular hazard, and a few cases of accidental exposure have already been reported.

MANIFESTATIONS OF SYSTEMIC DISEASE

The central nervous system. Since the optic nerve and retina are, embryonically, an extension of the brain, it is not surprising that central nervous system diseases frequently affect the eye, and visual defects may be the earliest evidence of the general disease. The nerve supply to the ocular muscles, particularly the extraocular muscles, may also be involved early in some diseases of the central nervous system: this will result in defective movement of the eyes, causing lack of coordination between the two eyes and diplopia, or double vision.

The nerve fibres that connect the retina with the site of visual sensation in the occipital cortex—*i.e.,* the outer brain substance at the back of the head—are arranged throughout the brain in a regular pattern, and many lesions of the brain, such as tumours, impinge on part of this pathway. From a detailed examination of the sensitivity of the different parts of the retina, using tests of the visual field, it is often possible to localize the exact site of an intracranial lesion. An optic neuritis causing sudden onset of loss of central vision in one eye is a frequent first symptom of multiple sclerosis. Detailed ophthalmic examination is therefore essential in the case of any patient suspected of having disease of the central nervous system.

Arteriosclerosis and vascular hypertension. The eye is the one structure in the body in which the blood vessels are easily visible to the examiner, and the changes that can be observed in the retinal vessels mirror those that are taking place in other parts of the body, particularly those in the brain. In arteriosclerosis degenerative changes occur in the walls of the arteries: this leads to thickening of the walls and narrowing of the bloodstream and may give rise to complete occlusion, or blockage of the vessel. If the central retinal artery is affected, loss of vision is complete and sudden and, unless the obstruction can be relieved within an hour or so, permanent. Occlusion of the retinal veins is more common than arterial occlusion and also has dramatic effects: the damming up of blood in the eye results in the bursting of small vessels, and multiple hemorrhages are scattered all over the fundus (that part of the inner eye which can be inspected through the pupil). Some degree of recovery of vision is usual but depends on whether a branch of the central vein or the vein itself is occluded.

Vascular hypertension, or raised blood pressure, usually occurs in association with arteriosclerosis. Typical changes can be recognized in the small vessels of the fundus, and in severe cases multiple hemorrhages and exudates, with swelling of the optic disk (the head of the optic nerve), may be present.

Diabetes. The satisfactory control of diabetes with insulin has increased the incidence of eye complications, for it has become apparent that it is the duration of the disease rather than its severity that determines the onset of ocular changes. A special type of cataract may occur in young diabetics with severe untreated disease, but the most serious complication involves the blood vessels of the retina. The actual cause of the changes in the retinal vessels is still unknown, but the natural history of the disease is well recognized. The retinal capillaries dilate at weak points in the vessel wall—*i.e.,* form small aneurysms; these weak portions of the vessel wall may give way and cause hemorrhages into the retina. In the later stages the hemorrhages become more extensive and spread into the vitreous. New vessels grow into hemorrhagic areas and are followed by fibrous changes that may pull on the retina and cause detachment. Extensive changes of this nature lead invariably to blindness.

Destruction of the pituitary gland, either by direct surgery or by the implantation of a radioactive material, has given some hope of alleviating these severe retinal changes. The procedure is, however, a drastic one. Destruction of affected areas of the retina by the use of an intense beam of light, a process called photocoagulation, promises to be a useful form of treatment in selected cases. Degenerative changes in the retina remain the most serious complication of diabetes.

Thyroid disease. The staring appearance of persons suffering from thyrotoxicosis, also called exophthalmic goitre or Graves' disease, is believed to be due to the stimulation of smooth muscle in the lids and orbit, causing the lid to retract a little from the globe and the globe itself to advance forward slightly. These changes normally regress if the thyrotoxicosis is treated. There is a more serious ocular complication of thyroid disease, which may follow excision of the thyroid for thyrotoxicosis or may, in some cases, arise in persons with normal or subnormal thyroid activity. It is characterized by swelling of the orbital tissues, including the extraocular muscles, so that the eyes cannot be moved properly and project forward between the lids to such an extent that the cornea becomes permanently exposed; the cornea may then ulcerate and even perforate and cause loss of the eye. Sewing the lids together may be sufficient to protect the cornea, but in many cases surgery to relieve the pressure in the orbit is necessary.

Rheumatism. The ocular complications of rheumatoid arthritis mainly involve the sclera, patches of inflammation occurring under the conjunctiva in the scleral and episcleral tissues (the latter are connective tissues between the conjunctiva and the sclera). Although the condition may respond to treatment with corticosteroids, recurrences are common.

VISUAL DISORDERS

"Floaters"; blind areas; flashing lights

Subjective symptoms. One of the commonest visual symptoms is the sensation of small, black objects floating in front of the eye. These move with the eye but lag slightly at the beginning of an eye movement and overshoot when the movement stops. They are due to cells and fragments of debris in the vitreous cavity of the eye. In certain conditions, as when looking at an empty sky, almost everybody can perceive them, and they are normal phenomena. A sudden increase in their number may indicate degenerative changes in the vitreous, which are particularly likely to occur in shortsighted eyes and in older people. These changes, although annoying, are of no serious import. The appearance of many "floaters," however, may be associated with inflammation or bleeding in the eye.

Blind areas in the field of vision occasionally force people to seek medical advice. Any condition that causes failure of function of part of the retina, the optic nerve, or the optic pathway to the brain, can cause such a blind spot, and the symptom requires careful investigation. There is a naturally occurring "blind spot" in each visual field that corresponds with the lack of retinal elements where the optic nerve enters the eye. The brain is so skillful in filling in the visual pattern that the normal blind spot can be detected only by special methods.

Flashing lights in the field of vision are caused by stimulation of the retina by mechanical means. Most commonly this occurs when the vitreous becomes degenerate and fluid and pulls slightly on its peripheral attachment to the retina. Similar symptoms also arise when the retina becomes detached, causing flashing lights to be seen.

Night blindness and defects of colour perception. Defective vision under reduced illumination may be a rare congenital condition or may be acquired as a result of severe deficiency of vitamin A.

Defective colour vision affects about four percent of men and 0.4 percent of women. Total colour blindness is extremely rare and is nearly always associated with poor vision in ordinary light. The colour-defective person is rarely aware of his disability until special matching tests are used, when it is discovered that he is unable to distinguish between hues in one or another part of the visual spectrum. Other visual functions are perfectly normal, and the only disadvantage is the restriction of certain types of occupation.

Eyestrain. Eyestrain, or asthenopia, is the term used to describe symptoms of fatigue and discomfort following the use of the eyes. Although such symptoms may result from intensive close work, particularly if this is unaccustomed, in people with perfectly normal eyes, they may indicate abnormalities of muscle balance or refractive errors. Eyestrain is more likely to be manifest during periods of fatigue or stress and is common among students working for examinations. Refractive errors require correction and muscle imbalance treatment. Psychological factors are often more important than physical factors.

Presbyopia; short sight; long sight

Refractive errors. In a normal eye rays of light from distant objects come to a focus on the retina. In near vision, the refractive power of the eye is increased by altering the shape of the lens to focus the image on the retina. A twelve-year-old can focus on an object four inches away from the eye but, with age, the ability of the lens to alter its shape decreases so that at the age of 40 the shortest distance at which an object can be kept in focus is about 10 inches. The near point continues to recede with age until fine print, for example, cannot be read at a normal reading distance. This condition is known as presbyopia; it is corrected by the use of convex lenses for reading.

In some eyes rays of light from distant objects are not brought to a focus on the retina but are focused on a plane in front of the retina, as in myopia (short sight), or behind the retina, as in hypermetropia (long sight). In myopia, near objects are brought into focus on the retina but distant objects can only be seen clearly with the aid of concave lenses. In hypermetropia, distant objects can usually be brought into focus by using the accommodative power of the lens, and in young people there is usually sufficient accommodation to enable them to see reasonably near to them. The constant accommodative effort required, however, may produce symptoms, and the necessity for accommodating for distance can be overcome by wearing convex glasses.

Another type of refractive error is astigmatism. In this condition the refractive power of the eye varies in different axes because of variation in curvature so that vision at all distances is distorted and can only be corrected by the use of cylindrical lenses or contact lenses.

Minor degrees of refractive error are extremely common. The refractive state is genetically determined and there are marked racial differences. Myopia, for example, is common in the Far East and rare in the African Negro. Although most refractive errors are easily correctible by spectacles and such errors are rarely accompanied by any serious disease of the eyes, hypermetropia is a factor in the development of some kinds of squint and high degrees of myopia are often associated with serious degenerative changes within the eye.

OPHTHALMOLOGICAL EXAMINATION AND CORRECTIVE DEVICES

Ophthalmological examination. An ophthalmological examination comprises a history of a patient's symptoms and signs, subjective tests to determine the visual function, and physical examination of the eyes by means of special devices. The most important subjective test is for visual acuity, and this is usually performed by presenting to the patient a series of letters of graded sizes at a set distance. He is required to read the lowest line legible to him; visual acuity can then be expressed in terms of the size of the letter and the distance at which it is read.

The visual field is assessed by moving an illuminated target inward from the periphery toward a central point viewed by the eye: the area in which the target is seen can then be drawn as a map of the visual field for that eye.

Other subjective examinations include colour-vision testing and tests of visual perception under reduced illumination. Examination of the external eye and part of the anterior segment is facilitated by the use of a binocular microscope mounted horizontally, to which is attached a slit-lamp, a variable source of light that projects the image of a slit onto the eye. The ophthalmoscope has an illuminating system that lights up the interior of the eye and a viewing system through which the fundus can be observed. Photography of the anterior part of the eye and of the fundus is both possible and widely used.

Other specialized methods of examination include examination of the angle of the anterior chamber by means of a specially designed contact lens with the slit-lamp microscope. The electrical responses of the retina and brain to light entering the eye can also be recorded and are of great value in certain conditions.

Estimation of the intraocular pressure is an important part of an ophthalmological examination and is accomplished by an instrument called a tonometer. This instrument is designed specifically to measure the tension or pressure that exists within the eyeball.

The refractive state of the eye can be measured objectively, or subjectively, or by a combination of both methods. The simplest method is subjective, using lenses of different powers to give a trial-and-error estimate of the best correcting lenses. More accurate results can be obtained by using an instrument known as a retinoscope, which gives an objective assessment of the refraction that can subsequently be modified by subjective methods to suit the individual requirements of the patient.

Types of spectacles

Optical aids. The most widely used optical aids are spectacles, and the technical design of spectacle lenses has advanced considerably in the last 50 years. A simple biconcave or biconvex lens causes considerable distortion of appearances if objects are viewed through the periphery of the lens, but if the back surface of the lens is made concave and the required power attained by altering the curvature of the front surface, improvement in peripheral definition results. All modern spectacle lenses are of this form.

Most older people require an additional lens for reading, and this can be incorporated with the distance correction in the form of a bifocal lens. In some occupations an intermediate distance is also required, and a third segment can be added, forming a trifocal lens. The complete range of correction from distance to near can only be achieved by means of a multifocal lens, and these are now available: the upper segment provides the correction for distance; as the eye moves lower down the lens its power increases, the lowest segment of the lens representing the reading correction. By slightly tilting the head it is possible to find the optimum correction for any intermediate distance.

The distortion of peripheral view when using conventional spectacles occurs because the correcting lens does not move when the eye moves. This problem can be completely overcome by the use of contact lenses, which are thin shells of plastic made to fit the anterior surface of the cornea and thus move with the eye. The earliest types of contact lens were larger than the cornea and were uncomfortable to wear, but the modern small "corneal" lenses have greatly increased the scope and usefulness of contact lenses and offer a practical alternative to the wearing of spectacles. Even so, there is a limit to the length of time for which they can be worn, and they should not be kept in at night. Their small size makes them easy to lose and difficult to manipulate, particularly for elderly persons. Further advances in design and in the use of new materials—for example, the flexible soft lenses—will doubtless extend the use of contact lenses even further.

For those persons who cannot obtain useful vision with ordinary spectacles or contact lenses, much can still be done by the use of compound lens systems known as low vision aids. These devices provide a magnified image but inevitably reduce the visual field. Their main value is to enable a person to read normal print who would otherwise be unable to read. They can be of use for distance, particularly when viewing conditions are relatively static, as with the cinema, theatre, or television.

Finally, for those who are completely blind from ocular causes there is new hope in the development of implants into the visual cortex that can be connected to a small television camera in such a way that electrical signals can be applied to the visual cortex, completely bypassing the normal optic pathways. The miniaturization of electrical circuitry resulting from space research has made the design of such devices a practical possibility. Their application to human subjects is, however, still in the experimental stage.

BLINDNESS

It is difficult to obtain reliable statistical information on the incidence of blindness on a worldwide basis. Even in countries in which the registration of blind people is attempted, the definitions of "blindness" vary from one country to another; in large parts of the world there is no registration, and the only estimate that can be made depends on random surveys of small parts of the population. An incidence of about 200 per 100,000 is fairly representative of countries in which the standard of medical care is high; it is probable that the incidence is 10 times higher in countries in which medical care is rudimentary.

There is wide variation in the causes of blindness in different parts of the world. This is partly due to geographic and climatic conditions but, more important, it is also due to differences in standards of hygiene and the availability of medical care. Infections, particularly trachoma, spread most easily in warm countries where the population is often crowded into small villages with lack of adequate hygienic facilities. Cataract is still high on the list of causes of blindness in many countries in the world, and this is all the more tragic in that it is so easily curable by surgical means. As the standards of general medical care increase and the expectation of life increases, so the pattern of blindness changes and degenerative conditions, diabetic disorders of the retina, and genetically determined diseases become predominant. Advances in the prevention and the medical and surgical treatment of blindness can only be of benefit to a population that has access to medical care. Until the nutritional and hygienic standards of a large part of the world population can be improved, preventable blindness will remain at its present high level. (E.S.P.)

HUMAN HEARING: STRUCTURE AND FUNCTION OF THE EAR

Anatomy of the auditory apparatus

Parts of the ear

The human ear, like that of other mammals, has three distinguishable parts: the external, the middle, and the inner ear. The external ear consists of the portion projecting from the side of the head, called the auricle or pinna, and the external auditory canal, which ends blindly at the eardrum. The middle ear is a narrow, air-filled space within the temporal bone, separated from the outside by the tympanic (eardrum) membrane and crossed by a chain of three tiny bones, the auditory ossicles. The inner ear is a complicated system of fluid-filled passages and cavities, deep in the rock-hard petrous portion of the temporal bone. It contains the sensory organs of hearing and equilibrium, the specialized endings of the auditory, or eighth cranial, nerve (Figure 52).

By courtesy of Abbott Laboratories, North Chicago, Illinois

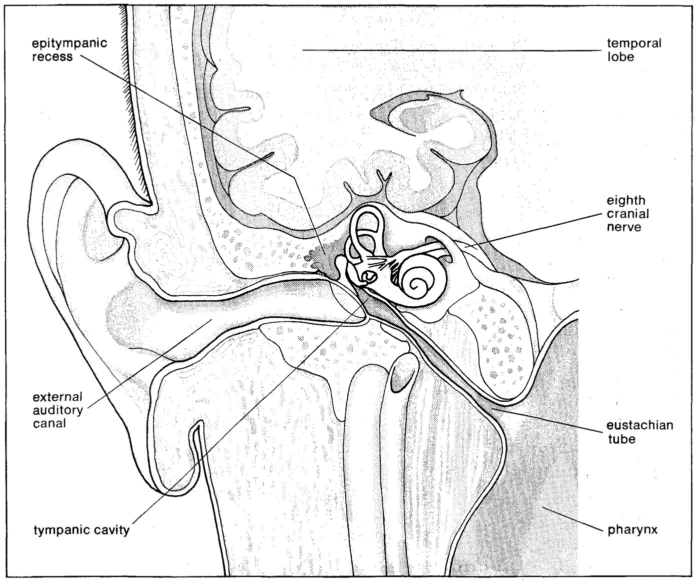


Figure 52: Section through the right side of the head showing the parts of the ear and the auditory area in the temporal lobe of the cerebral cortex.

EXTERNAL EAR

The most striking differences between the ear of man and that of many other mammals are seen in the structure of the auricle itself. In man the auricle is an almost rudimentary, usually immobile shell, more or less closely applied to the side of the head. It consists of a thin plate of yellow fibrocartilage covered by closely adherent skin. The cartilage is molded into the characteristic shape with clearly defined hollows, ridges, and furrows, forming an irregular, shallow funnel (Figure 53). The deepest depression, leading directly to the external auditory canal, or meatus, is called the concha. It is partly covered by two small projections, the tongue-like tragus in front and the antitragus behind. Above the tragus, a prominent ridge, the helix, arises from the floor of the concha and continues as the incurved rim of the upper portion of the auricle. An inner, concentric ridge, the antihelix, surrounds the concha and is separated from the helix by a furrow, the scapha (also called the fossa of the helix). In some ears a little prominence, known as Darwin's tubercle, is seen along the upper, posterior portion of the helix, the vestige of the folded-over point of the ear of a remote prehuman ancestor. The lobule, the fleshy, lower portion of the auricle, is the only part of the external ear that contains no cartilage. The auricle also has several small rudimentary muscles, which connect it to both the skull and the scalp. These muscles are usually without function but in some persons are capable of limited movements.

External auditory canal

The external auditory canal is a slightly curved tube, extending inward from the floor of the concha and ending blindly at the tympanic membrane. In the outer third the wall consists of cartilage; in the inner two-thirds, of bone. The entire length of the passage (24 millimetres, or almost one inch) is lined with skin, which also covers the surface

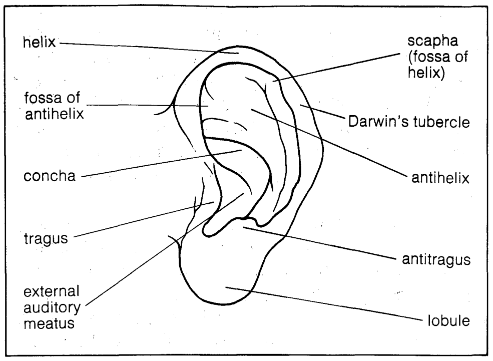


Figure 53: Structure of the outer ear.

of the eardrum membrane. Fine hairs directed outward and modified sweat glands producing earwax, or cerumen, discourage insects from entering the canal.

TYMPANIC MEMBRANE AND MIDDLE EAR

Tympanic membrane. The tympanic membrane lies obliquely across the end of the external canal in the form of a flattened cone with its apex directed inward. Its edges are attached to a ring of bone, the tympanic annulus. The diameter of the membrane is about nine millimetres (0.35 inch). The appearance and mobility of the membrane are important for the diagnosis of middle-ear disease. The healthy membrane is pearl gray; sometimes it has a pinkish or yellowish tinge.

The drum membrane is made up of three layers. The outer layer is continuous with the skin of the external canal; the inner with the mucous membrane lining the middle ear. Between them is a layer of circular and radial fibres that give the drum its stiffness and tension. It is well supplied with blood vessels and with sensory nerve fibres that make it acutely sensitive to pain.

The cavity, including the eustachian tube. The cavity of the middle ear is a narrow, air-filled space. A slight constriction divides it into two chambers, the tympanum (eardrum) proper, or atrium below, and the epitympanum, or attic above. Its outer wall is formed largely by the tympanic membrane, its inner wall by the bony capsule of the inner ear. Its roof and floor are thin plates of bone separating it from the cranial cavity and the brain above and from the carotid artery and the jugular vein below. Its posterior bony wall has an opening leading to a second air space, the tympanic antrum, and to the small air cells of the mastoid process, the portion of the temporal bone directly in back of the external auditory canal. In the narrow anterior wall is the opening of the eustachian or auditory tube.

The eustachian tube and its function

The eustachian tube, about 45 millimetres (1.75 inches) in length, leads downward and inward from the tympanic cavity to the nasopharynx, the space that is behind and continuous with the nasal passages and above the soft palate. At its upper end the tube is rather narrow and surrounded by bone. Toward the pharynx it widens and becomes cartilaginous. Its mucous lining is continuous with that of the middle ear. The cilia (small motile hairlike projections) that cover it help to speed the drainage of mucous secretions from the middle ear to the pharynx.

The major function of the tube is to ventilate the middle ear and thus maintain equal air pressure on both sides of the drum membrane. The tube is closed at rest and opens during swallowing, so that minor pressure differences are adjusted without conscious effort. During a dive or a rapid descent in an airplane the tube may remain tightly closed. The discomfort that is felt as the atmospheric pressure increases can usually be overcome by attempting a forced expiration with the mouth and nostrils tightly shut. This manoeuvre, which raises the air pressure in the pharynx and causes the tube to open, is named for the Italian anatomist Antonio Maria Valsalva (1666–1723), who recommended it for clearing pus from an infected middle ear.

Auditory ossicles. Three auditory ossicles with somewhat fanciful names form a short chain crossing the middle ear. From the outside inward, they are the malleus (hammer), the incus (anvil), and the stapes (stirrup). In appearance the malleus more nearly resembles a club, and the incus a premolar tooth with widely spreading roots. The stapes, on the other hand, is unmistakably a stirrup.

The head of the malleus and the body of the incus lie in the attic above the upper margin of the drum. The two small bones have a tightly fitting joint between them. The handle of the malleus is firmly attached to the upper half of the drum membrane. Its head is anchored to the walls and roof of the attic by three small ligaments. Another ligament fixes the short process (projection) of the incus in a slight depression in the rear wall of the cavity. Its long process is bent near its lower end and carries a small bony knob that forms a loose joint with the head of the stapes. This last, the smallest and lightest of the ossicles, is about three millimetres (about 0.1 inch) in height and weighs scarcely three milligrams (0.0001 ounce). It lies almost horizontally, at right angles to the long process of the incus. Its footplate fits nicely in the oval window, one of the two openings in the wall of the bony labyrinth, where it is held in place by a ringlike ligament called the annular ligament.

Two minute muscles are found in the middle ear. One, the slender muscle called the tensor tympani, emerges from a bony canal just above the opening of the eustachian tube, runs backward, changes direction as it passes over a pulley-like projection, and attaches to the upper part of the handle of the malleus. Its contractions tend to pull the malleus inward and thus increase the tension of the drum membrane. The other, called the stapedius, rises in the posterior wall and sends its minute tendon forward to attach to the neck of the stapes. Its contractions tend to pull the footplate out of the oval window by tipping the stapes backward.

When the healthy tympanic membrane is examined with the otoscope (an instrument designed for visual inspection of the interior of the ear), through the membrane the handle of the malleus is clearly seen projecting from above downward and backward and dividing the upper portion of the membrane into two almost equal parts. Behind and parallel to it, the long process of the incus can sometimes be made out. Above the malleus is a small triangular area in which the membrane is thin and slack. Behind this area, called the pars flaccida (as opposed to the pars tensa, which makes up the much larger portion of the membrane), lies the bare chorda tympani, a slender branch of the facial nerve that passes through the middle ear on its way to join the lingual nerve. It carries important secretory fibres to the parotid (salivary) gland, and sensory fibres to the taste buds of the tip of the tongue.

INNER EAR

Structure as a whole. The inner ear is enclosed in a bony case called the otic capsule, a part of which forms the inner wall of the middle ear. There, two openings between the middle and inner ear are found: the oval window above, which is filled by the footplate of the stapes, and the round window, which is covered by a thin membrane, sometimes referred to as the secondary tympanic membrane. Between them is a bulge called the promontory.

Because of its complicated galleries and chambers, the inner ear as a whole is referred to as the labyrinth. There are, in fact, two labyrinths, one inside the other. The passages hollowed out in the otic capsule constitute the bony labyrinth. In it is suspended a delicate system of ducts and sacs that constitutes the membranous labyrinth (Figure 54).

The bony labyrinth

The bony labyrinth consists of a central chamber called the vestibule, the three semicircular canals, and the spirally coiled cochlea. The last strongly resembles the shell of a snail, and its name is derived from the Greek word for a snail. The canals are designated, according to their position, superior, lateral, and posterior. The superior and posterior canals are in diagonal vertical planes that intersect at right angles. The lateral canal is often referred to as the horizontal because it lies approximately in that plane. Each canal has an expanded end, the ampulla, which opens into the vestibule. The ampullae of the lateral and superior canals lie close together, just above the

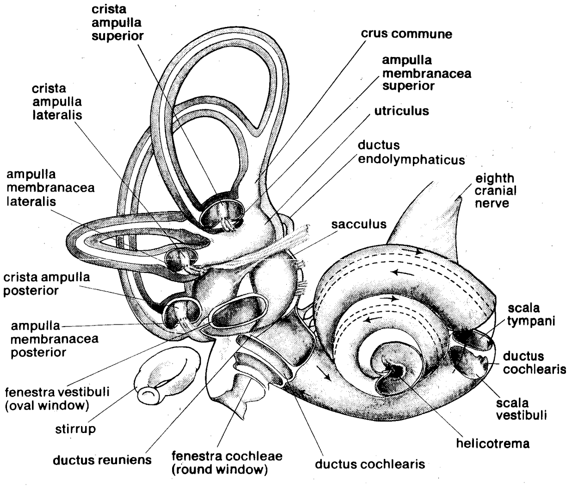


Figure 54: The membranous labyrinth, within the bony labyrinth. The stapes has been removed from the oval window.

By courtesy of Abbott Laboratories, North Chicago, Illinois

oval window, but the ampulla of the posterior canal opens on the opposite side of the vestibule. The other ends of the superior and posterior canals join to form a common stem, or crus, which also opens into the vestibule. Nearby is the mouth of a canal called the vestibular aqueduct, which opens into the cranial cavity. The lateral canal has an independent opening at its opposite end. Thus the vestibule, in effect, completes the circle for each of the semicircular canals.

The cochlea, a tube consisting of two and one-half spiral turns around a hollow central pillar, the modiolus, forms a cone approximately nine millimetres (0.35 inch) in diameter at its base and five millimetres (0.2 inch) in height. The tube is approximately 30 millimetres (1.2 inches) in total length, tapering from a width of two millimetres (0.08 inch) where the basal coil joins the vestibule and ending blindly in the cupula at the apex. The hollow centre of the modiolus is a spiral canal, containing the fibres and ganglion cells of the cochlear nerve; this enters the base of the modiolus through an opening in the petrous portion of the temporal bone called the internal meatus. A thin bony shelf, the osseous lamina, winds around the modiolus like the thread of a screw. It projects about half way across the cochlear canal, partly dividing it into two ramps, or galleries, the scala vestibuli above and the scala tympani below. The round window is in the wall of the scala tympani at its basal end. Nearby is the opening of the narrow cochlear aqueduct, through which passes the perilymphatic duct, connecting the interior of the cochlea with the posterior cranial fossa (the rear portion of the floor of the cranial cavity).

Suspended in the bony labyrinth, the membranous labyrinth occupies only a fraction of the available space. Each of the semicircular canals contains a narrow membranous semicircular duct, which by no means fills the channels or lumen of the canal, with a rounded expansion, the ampulla, at one end. In the same way that the semicircular canals communicate with the vestibule, the semicircular ducts open into an elongated tubular sac, the utricle, located in the upper part of the vestibule. The utricle is also connected, through a narrow duct, with a similar structure, the saccule, which occupies the lower part of the vestibule. The cochlear duct is a coiled, tapered tube, suspended between the scala vestibuli and scala tympani of the cochlea, ending blindly at both its basal and its apical ends. It communicates with (opens into) the saccule through a narrow connecting duct (the ductus reuniens) near the basal end. The duct that unites the utricle and the saccule is also connected to another narrow channel, the endolymphatic duct, which passes through the vestibular aqueduct and ends in a pouch on the cranial surface of the petrous bone, the endolymphatic sac.

Perilymph and endolymph

The space within the bony labyrinth that is not occupied by the various parts of the membranous labyrinth is filled with a watery fluid, the perilymph, which in composition closely resembles the cerebrospinal fluid, the aqueous humour of the eye, and other extracellular fluids of the body tissues. Like them, it is apparently formed locally from the blood plasma by ultrafiltration through the walls of the minute blood vessels called capillaries. Since it is possible for cerebrospinal fluid to enter the cochlea by way of the perilymphatic duct, a portion of the perilymph may come from that source.

The membranous labyrinth is filled with a second watery fluid of different composition, the endolymph. Chemical analysis has shown that perilymph is like most other extracellular fluids, being high in its concentration of sodium ions (about 150 milliequivalents per litre) and low in its concentration of potassium ions (about five milliequivalents per litre), but that endolymph is unique among the extracellular fluids in that it has potassium as the dominant ion (about 140 milliequivalents per litre), with a much reduced concentration of sodium (about 15 milliequivalents per litre). Contrary to the statement sometimes found in textbooks, mammalian endolymph is not a viscous fluid. Recent studies indicate that its viscosity is approximately that of water.

The process of formation of the endolymph and the maintenance of the difference in composition between it and perilymph is not yet completely understood. The tissue known as the stria vascularis, in the wall of the cochlear duct, is thought to play an important role in its secretion, but other tissues of the cochlea and the vestibular organs are probably also involved. Since the membranous labyrinth is a closed system, the question of removal of the endolymph is also of interest. Reabsorption is thought to occur from the endolymphatic sac, but this is probably only part of the story. In all probability other tissues also have important roles in regulating the inner-ear fluids.

Utricle and saccule

Parts of the inner ear. *Vestibule.* The vestibule includes the utricle and saccule, each of which contains a single sensory patch called a macula. In the utricle the macula projects from the anterior wall as an oval spot on a small rounded shelf. In the saccule, the macula lies against the medial (inner) wall of the vestibule directly overlying the bone. It is more elongated than the utricular macula and somewhat resembles the letter J in shape. Each macula is covered by neuroepithelium. This covering consists of sensory cells—called hair cells because of their hairlike projections—and supporting cells. The cells are separated by a basement membrane from the underlying connective tissue. Fibres from the vestibular branch of the vestibulocochlear (eighth cranial, or auditory) nerve enter the macula and pierce the basement membrane to end either at the base of the hair cells or as cuplike formations called calyces surrounding their cell bodies. The hair cells are thus the essential sensory receptor elements of the macula. Each hair cell is topped by a bundle consisting of about 100 fine, nonmotile "hairs" (stereocilia) of graded lengths and a single motile hairlike projection called a kinocilium. Covering the entire macula is a delicate acellular structure, the otolithic (or statolithic) membrane, which is often described as gelatinous although it has a distinct fibrillar (minute-fibre-like) pattern. The surface of the membrane is covered by a blanket of rhombohedral crystals, the otoconia (or statoconia), consisting of calcium carbonate in the form of calcite. These crystalline particles, which range in length from one to 20 microns (there are about 25,000 microns in an inch), have a specific gravity almost three times that of the membrane itself and thus add consider-

From H.H. Lindeman, "Studies on the Morphology of the Sensory Regions of the Vestibular Apparatus," *Ergebnisse der Anatomie und Entwicklungsgeschichte*, vol. 42.1 (1969); Springer-Verlag, New York

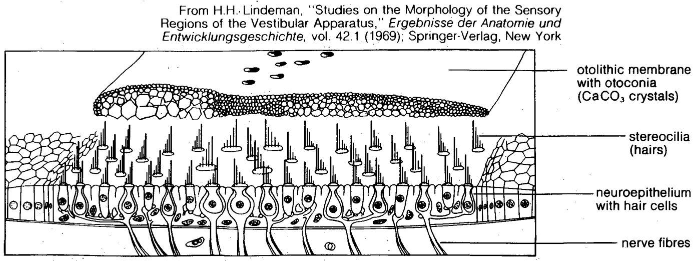

otolithic membrane with otoconia ($CaCO_3$ crystals)
stereocilia (hairs)
neuroepithelium with hair cells
nerve fibres

Figure 55: Structure of the macula of the utricle.

able mass to it. The hair cells of the maculae are of two types. One type has a rounded body enclosed by a nerve calyx, the other a cylindrical body with nerve endings at its base. The surfaces of the utricular and saccular maculae show characteristic organization and orientation of the two types of cells in definite mosaic patterns. The larger, rounded cells tend to have a curvilinear arrangement near the centre of the macula, with the cylindrical cells around the periphery. The significance of these patterns is poorly understood, but they are presumed to favour increased directional sensitivity to movements of the head (Figure 55).

The utricle and saccule with their maculae are often referred to as otolith organs. Because they respond to gravitational forces, they are also called gravity receptors.

Sensory end organ of semicircular duct

Semicircular canals. The sensory end organ of each semicircular duct is located in the expanded end, or ampulla. It consists of a saddle-shaped ridge of tissue covered with a sensory epithelium containing the same types of cells as do the maculae. The ridge, called the crista, extends across the ampulla from side to side, at right angles to the direction of the duct (Figure 56). It is covered by a gelatinous structure, the cupula, which extends to the roof of the ampulla immediately above it, dividing the interior of the ampulla into two approximately equal parts. On each of the hair cells of the sensory epithelium of the crista there are about 100 stereocilia, or "hairs," anchored in the dense cuticular plate. The single kinocilium springs from a small area at one side of the surface of the cell, where it is covered only by the thin plasma membrane. The stereocilia stand straight, in orderly rows. The longest are those nearest the kinocilium, which is motile. The height of the others decreases in stepwise fashion away from the kinocilium. Together with the longest stereocilia, the kinocilia extend well up into the substance of the cupula, occupying fine parallel channels. The cupula is thus attached at its base to the crista, but is free to move in either direction in response to pressure in the endolymph toward or away from the utricle. Its motion is transmitted to the hairs, causing them to bend just above the cuticular plate. This bending is the stimulus to the hair cells and elicits impulses in the vestibular fibres of the auditory nerve. Each macula and each crista is supplied by fibres from the vestibular nerve, one of the two parts of the vestibulocochlear (auditory, or eighth cranial) nerve.

From H.H. Lindeman, "Studies on the Morphology of the Sensory Regions of the Vestibular Apparatus," *Ergebnisse der Anatomie und Entwicklungsgeschichte*, vol. 42.1 (1969); Springer-Verlag, New York

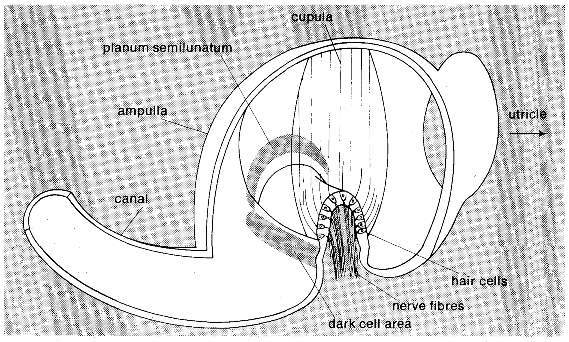


Figure 56: Ampulla of semicircular duct, showing hair cells and cupula.

Cochlea. The interior of the cochlea is divided longitudinally into three spiral ramps or scalae: the scala vestibuli, which communicates with the vestibule; the scala tympani, which ends blindly at the round window; and the scala media, or cochlear duct, which lies between the other two scalae and ends blindly both at the round window and at the apex. The scala vestibuli and scala tympani communicate at the apex, through an opening called the helicotrema (Figure 54).

The scala media

In cross section the scala media resembles a right triangle. Its base is formed by a thin bony shelf, the spiral osseous lamina, and by the basilar, or supporting, membrane, which separate it from the scala tympani, and its side by two important tissues lining the bony wall of the cochlea, the delicate stria vascularis and the tough, fibrous spiral ligament. The hypotenuse is formed by the thin membrane of Reissner, consisting of only two layers of flattened cells. A low ridge, the spiral limbus, rests on the margin of the osseous lamina. Reissner's membrane stretches from the inner margin of the limbus to the upper border of the stria, behind which it is inserted (Figure 57).

Upon the surface of the basilar membrane the sensory cells are arranged in more or less orderly rows. Together with their supporting cells they form a complex neuroepithelium called the basilar papilla or organ of Corti, named after the Italian microscopist, the Marchese Alfonso Corti, who first described it, in 1851. Viewed in cross section its most striking feature is the arch, or tunnel, of Corti, formed by two rows of pillar cells, or rods. The pillar cells furnish the major support of Corti's organ. They separate a single row of larger, more or less pear-shaped inner hair cells on the inner side of the tunnel from three or more rows of smaller, cylindrical, outer hair cells on the outer side. The inner hair cells are supported and enclosed by the inner phalangeal cells, which rest on the thin outer portion, called the tympanic lip, of the spiral limbus. On the inner side of the inner hair cells and the cells that support them is a curved furrow called the inner sulcus. This is lined with more or less cuboidal cells.

Each outer hair cell is supported by a cell known as a phalangeal cell of Deiters, which holds the base of the hair cell in a cup-shaped depression. From each Deiters cell a projection extends upward to the stiff membrane, the reticular lamina, that covers the organ of Corti. The top of the hair cell is firmly held by the lamina, but the body is suspended in fluid that fills the so-called space of Nuel and the tunnel. Although this fluid is sometimes referred to as the cortilymph, it appears to have the same composition as the perilymph. Beyond the hair cells and the cells of Deiters are epithelial cells of three other types, usually called the cells of Hensen, Claudius, and Böttcher, after the 19th-century microscopists who first described them. Their function has not been established, but they are assumed to help in maintaining the composition of the endolymph by secretory and absorptive activity.

The basilar membrane consists of two zones, the pars arcuata under the arch, and the pars pectinata extending from the feet of the outer pillars to the spiral ligament. The fibres of the membrane are fine and inconspicuous in the pars arcuata. Under the feet of the outer pillars the

From C.H. Best and N.B. Taylor, *The Physiological Basis of Medical Practice*, 8th ed. (© 1966); Williams and Wilkins Co., Baltimore

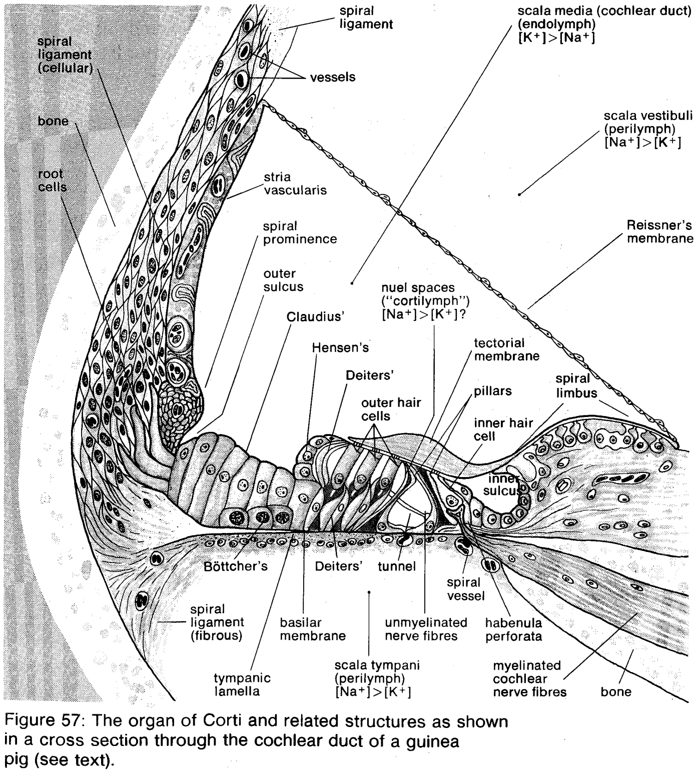


Figure 57: The organ of Corti and related structures as shown in a cross section through the cochlear duct of a guinea pig (see text).

membrane splits into two distinct layers, the upper one containing stout, parallel, radial fibrils that give the membrane its stiffness. These fibrils decrease in calibre and increase in length from the basal end of the cochlea to the apex, so that the basilar membrane as a whole decreases in stiffness as it increases in width from base to apex.

Beneath the fibrillar layer is the acellular ground substance of the membrane, which is covered in turn by a single layer of mesothelial cells forming the tympanic lamella on the tympanic surface.

The spiral vessels

Capillary blood vessels are found beneath the tympanic lip of the limbus and, in many species, beneath the tunnel. These vessels, called spiral vessels, do not enter the organ of Corti but are thought to furnish the major part of the supply of oxygen and nutrients to its various cells, including the hair cells. Although the outer spiral vessel is seldom found in adult animals of certain species such as the dog and cat and is somewhat irregularly distributed in man, it is always present during fetal development. To judge by its size in the fetus, it plays a major part in bringing blood to the rapidly differentiating organ of Corti.

Each hair cell is capped by a plate, which bears the stereocilia or hairs. On the inner hair cells 40 to 60 stereocilia are arranged in two or more irregularly parallel rows. On the outer hair cells approximately 100 stereocilia form a W pattern. Below the notch of the W the plate is incomplete. At this point is the base of the kinocilium, although the motile "hair" is absent.

The stereocilia are about three to five microns in length and extend upward from the cuticular plate to the underside of the tectorial membrane. This is an acellular, gelatinous structure that covers the top of the limbus spiralis as a thin layer and extends outward over the inner sulcus and the reticular lamina. Its fibrils extend radially and somewhat obliquely to end in finger-like projections at its free margin; these make contact with the stereocilia of the outermost hair cells.

The myelinated (sheathed) fibres of the vestibulocochlear nerve fan out in spiral fashion from the modiolus to pass into a channel near the root of the osseous lamina called the canal of Rosenthal. There, the bipolar cell bodies of these neurons, or nerve cells, are located, forming the spiral ganglion. Beyond the ganglion their dendrites extend radially outward in the lamina beneath the limbus to pass through the small pores of a structure called the habenula perforata, directly under the inner hair cells. At this point the fibres abruptly lose their multilayered coats of myelin and continue as thin, naked, unmyelinated fibres into the organ of Corti. They form a longitudinally directed bundle running beneath the inner hair cells and another at the foot of the inner pillars just inside the tunnel. The majority of the fibres end beneath the inner hair cells, but some of them cross the tunnel to form longitudinal bundles beneath the outer hair cells, on which they eventually terminate.

The endings of the nerve fibres beneath the hair cells are of two distinct types. The larger and more numerous endings contain many minute vesicles (liquid-filled sacs), which are related to impulse transmission at neural junctions. As described below these endings belong to a special bundle of nerve fibres arising in the brainstem and constituting an efferent system or feedback loop. The smaller and less numerous endings contain few vesicles or other cell structures (organelles). They are the terminations of the afferent fibres of the cochlear nerve, transmitting impulses from the hair cells to the brainstem.

The outer membranous wall of the cochlear duct is lined, as has been mentioned above, by the stria vascularis, a dense cellular layer containing a network of capillary blood vessels. The surface cells of the stria, called dark cells, are of secretory type. Interspersed among the dark cells is an intermediate layer of cells called light cells. Several layers of flat basal cells bound the stria and separate it from the spiral ligament.

Reissner's membrane is inserted at the upper margin of the stria. At the lower margin of the stria is the spiral prominence, a low ridge containing its own set of longitudinally directed capillary vessels. Below the prominence is a depression called the outer sulcus, the floor of which is lined by cells of epithelial origin that send long projections into the substance of the spiral ligament. Between these so-called root cells, capillary vessels descend from the spiral ligament. This region appears to have an absorptive rather than a secretory function, and it may be concerned with the removal of waste materials from the endolymph.

The spiral ligament

The spiral ligament lies between the stria vascularis and the bony wall of the cochlea. Extending above the attachment of Reissner's membrane, it is there in contact with the perilymph in the scala vestibuli; and extending below the insertion of the basilar membrane, it is there in contact with the perilymph in the scala tympani. It contains many stout fibres that anchor the basilar membrane, and numerous cells, mainly connective tissue cells (fibrocytes). Behind the stria the structure of the spiral ligament is denser than near the upper and lower margins.

Like the stria, the spiral ligament is well supplied with blood vessels. It receives the radiating arterioles that pass outward from the modiolus in bony channels of the roof of the scala vestibuli. Branches from these vessels form a network of capillaries above Reissner's membrane that is thought to be largely responsible for the formation of the perilymph as an ultrafiltrate of the blood plasma. Other branches enter the stria, and still others pass behind it to the spiral prominence and the floor of the outer sulcus. From these separate capillary networks, which are not interconnected, venules (small veins) descending below the attachment of the basilar membrane collect the blood and deliver it to the spiral vein in the floor of the scala tympani.

Viewed from above, the organ of Corti with its covering, the reticular lamina, forms a well-defined mosaic pattern. In man the arrangement of the outer hair cells in the basal turn of the cochlea is quite regular, with three distinct and orderly rows; but in the higher turns of the cochlea, it becomes increasingly irregular, as scattered cells appear representing incomplete fourth and fifth rows. The spaces between the outer hair cells are filled by the oddly shaped extensions (phalangeal plates) of the supporting cells. The double row of head plates of the inner and outer pillar cells cover the tunnel and separate the inner from the outer hair cells. The reticular lamina extends from the inner border cells near the inner sulcus to the Hensen cells but does not include either of these cell groups. When a hair cell degenerates and disappears as a result of injury, its place is quickly covered by the adjacent phalangeal plates, which form an easily recognized "scar."

The basilar membrane

The length of the basilar membrane (and of the organ of Corti that covers it) is about 35 millimetres (1.4 inches) in man. Its width varies from less than 0.001 millimetre near its basal end to 0.005 millimetre near the apex. The membrane is not under tension but decreases remarkably in stiffness from the base to the apex of the cochlea. Furthermore, at the basal end the osseous lamina is broader, the stria vascularis wider, and the spiral ligament stouter than at the apex. The mass of the organ of Corti, on the other hand, is least at the base and greatest at the apex. These considerations indicate that there is a certain degree of "tuning" provided in the structure of the cochlear duct and its contents. With greater stiffness and less mass, the basal end is more attuned to the higher frequencies of vibrations. Decreased stiffness and increased mass render the apical end more responsive to the lower frequencies.

The total number of outer hair cells in the cochlea has been estimated at 12,000, and the number of inner hair cells at 3,500. Although there are almost 30,000 fibres in the cochlear nerve, there is considerable overlap in the innervation of the outer hair cells. A single fibre may supply endings to many hair cells, which thus share a "party line." Furthermore, a single hair cell may receive nerve endings from many fibres. The actual distribution of the nerve fibres in the organ of Corti has not been worked out in any detail, but the inner hair cells appear to receive the lion's share of the afferent fibre endings (fibres bearing impulses to the brain), with less of the overlapping and sharing of fibres that are characteristic of the outer hair cells.

The auditory process

TRANSMISSION OF SOUND WAVES

Transmission to the inner ear. *Air conduction.* The auricle, or visible portion of the outer ear, because of its small size and virtual immobility in man, has lost most of the importance that it has in many animals as an aid in sound gathering and direction finding. For airborne sounds of relatively short wavelength—*i.e.,* those above 3,000 hertz (cycles per second)—the concha serves as a funnel, directing them into the canal. The canal itself contributes little of acoustic importance apart from a broad resonance centred at 3,800 hertz, which helps to determine the frequencies to which the ear is most sensitive.

Acoustic impedance

Sounds reaching the drum membrane are in part reflected and in part absorbed. Only that portion of the sound that is absorbed is effective in setting the drum membrane and the ossicles in motion and thus eventually reaching the inner ear. This tendency of the ear to oppose the passage of sound is called its acoustic impedance. The magnitude of the impedance depends upon the mass and stiffness of the drum membrane and the ossicular chain and on the frictional resistance that they offer. Direct or indirect measurement of the impedance of the ear in hard-of-hearing patients can give important information about the condition of the middle-ear mechanism.

The central portion of the drum membrane vibrates as a stiff cone in response to sound, at least at frequencies below 2,400 hertz. Its motion is transmitted to the handle of the malleus, the tip of which is at the umbo, or centre, of the membrane. At higher frequencies the motions of the drum membrane are no longer simple, and transmission to the malleus may be somewhat less effective. The malleus and incus are suspended by small elastic ligaments and are finely balanced, so that their masses are evenly distributed above and below their common axis of rotation. The head of the malleus and the body of the incus are tightly bound together, with the result that they move in and out as a unit with the movements of the drum membrane. At moderate sound pressures the stapes follows them, and the whole ossicular chain vibrates as a single mass. There may, however, be considerable freedom of motion and some loss of energy at the joint between the incus and the stapes because of the relatively loose coupling. The stapes itself does not move in and out, however, but rocks about the lower pole of its footplate as it transmits the vibrations to the perilymph that fills the vestibule. Its motion thus resembles that of a bell-crank lever rather than that of a piston (Figure 58).

Matching of impedances

The acoustic problem solved by the middle-ear mechanism of drum membrane and ossicular chain is that of transmitting sound from the air to the watery perilymph of the inner ear. Because of the great difference in density between the gas and the liquid, there is a serious mismatch of impedance between them. Ordinarily, this mismatch causes 99.9 percent of airborne sound striking a water surface to be reflected, so that only 0.1 percent passes into the water. In the ear this would represent a transmission loss of 30 decibels, enough to seriously limit its performance, were it not for the transformer action of the middle ear. The matching of impedances is accomplished in two ways, primarily by the reduction in area between the drum membrane and the stapes footplate and secondarily by the mechanical advantage of the lever formed by the malleus and incus. Although the total area of the drum membrane is about 69 square millimetres (0.1 square inch), the area of its central portion that is free to move has been estimated at about 43 square millimetres (0.07 square inch). All of the sound that causes this area of the membrane to vibrate is transmitted and concentrated in the 3.2-square-millimetre (0.005-square-inch) area of the stapes footplate. Thus the pressure (*i.e.,* force per unit area) is increased at least 13 times over. The mechanical advantage of the ossicular lever, because the handle of the malleus is longer than the long projection of the incus, amounts to about 1.3. The total increase in pressure at the footplate is, therefore, not less than 17-fold, depending upon the area of the drum membrane actually vibrating. At frequencies in the range of 3,000 to 5,000 hertz, to which the ear is most sensitive, the increase may be even greater because of the resonant properties of the ear canal as a closed tube.

Adapted from S.S. Stevens and H. Davis, *Hearing: Its Psychology and Physiology* (1938); John Wiley and Sons, Inc.

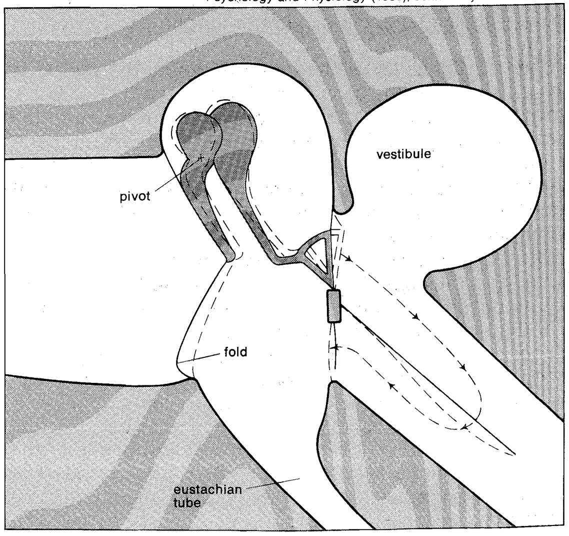


Figure 58: Mechanics (diagrammatic) of the drum membrane and ossicular chain (see text).

The ossicular chain not only concentrates sound in a small area but also applies sound preferentially to one window of the cochlea—*i.e.,* to the oval window, in which the footplate fits so neatly. If the oval and the round windows were equally exposed to airborne sound crossing the middle ear, the vibrations in the perilymph of the scala vestibuli would be opposed by those in the perilymph of the scala tympani, and little effective movement of the basilar membrane would result. As it is, sound is delivered selectively to the oval window, and the round window moves in reciprocal fashion, bulging outward in response to an inward movement of the stapes footplate and vice versa. The passage of vibrations through the air across the middle ear from the drum membrane to the round window is of negligible importance.

Thanks to these mechanical features of the middle ear, the hair cells of the normal cochlea are able to respond at the threshold of hearing to vibrations of the tympanic membrane no greater than one angstrom unit (0.0000001 millimetre) in amplitude. On the other hand, when the ossicular chain is immobilized by disease, as in otosclerosis, which causes the stapes footplate to become fixed in the oval window, increase in the threshold of hearing of as much as 60 decibels is common. Bypassing the ossicular chain through the surgical creation of a new window, as in the so-called fenestration operation, can restore hearing to within 25 or 30 decibels of the normal. Only if the stapes is removed and replaced by a tiny artificial stapes can normal hearing be approached. Fortunately, defects of the middle ear causing so-called conductive impairment can often be corrected by surgical means so that useful hearing is restored.

Action of middle-ear muscles

The muscles of the middle ear, the tensor tympani and stapedius, can influence the transmission of sound by the ossicular chain. Contraction of the tensor tympani pulls the handle of the malleus inward and, as the name of the muscle suggests, tenses the drum membrane. Contraction of the stapedius tends to pull the stapes footplate outward from the oval window and thereby reduces the intensity of sound reaching the cochlea. The stapedius responds reflexly with quick contraction to sounds of high intensity applied either to the same ear or to the opposite ear. The reflex has been likened to the eyeblink or constriction of the pupil in response to light and is thought to have protective value. Unfortunately, the contractions of the middle-ear muscles are not instantaneous, so that they do not protect the cochlea against damage by sudden intense noise, as of gunfire. They also fatigue rather quickly and thus offer little protection against injury by sustained high level noise, as in industry. Because dangerously intense noises are rare in nature, it has been argued that the chief biological significance of the aural reflex is not protection of the inner ear but selective high pass filtering of sound. In other words, since the contractions stiffen the ossicular chain, they favour the transmission of the higher frequencies. These carry more information than the lower frequencies, whether in the consonant sounds of human speech or in the rustle of leaves and the snap of a twig that may signal the approach of an enemy in the wild. At the same time, they reduce the transmission of the lower frequencies, which may otherwise mask those of greater significance for survival. The intra-aural muscle reflexes may also have a role in listening and attention, but this function is by no means fully understood.

Bone conduction. There is another route by which sound can reach the inner ear—that is, by conduction through the bones of the skull. When the handle of a vibrating tuning fork is placed on a bony prominence such as the forehead or the mastoid process behind the ear, its note is clearly audible. Similarly, the ticking of a watch held between the teeth can be distinctly heard. When the external canals are closed with the fingers, the sound becomes louder, indicating that it is not entering the ear by the usual channel. Instead, it is producing vibrations of the skull that are passed on to the inner ear, either directly or indirectly.

The higher audible frequencies cause the skull to vibrate in segments, and these vibrations are transmitted to the cochlear fluids by direct compression of the otic capsule, the bony case inclosing the inner ear. Since the round window membrane is more freely mobile than the stapes footplate, the vibrations set up in the perilymph of the scala vestibuli are not cancelled out by those in the scala tympani, and the resultant movements of the basilar membrane can stimulate Corti's organ. This type of transmission is known as compressional bone conduction.

At lower frequencies—*i.e.,* 1,500 hertz and below—the skull moves as a rigid body. The ossicles, because of their inertia and because they are suspended in the middle-ear cavity and only loosely coupled to the skull, are less affected and move less freely than the cochlea and the margins of the oval window. The result is that the oval window moves with respect to the footplate of the stapes, which gives the same effect as if the stapes itself were vibrating. This form of transmission is known as inertial bone conduction. The fixed stapes in otosclerosis interferes with this type of bone conduction, but not with compressional bone conduction.

Hearing aids for conduction deafness

In the presence of middle-ear disease, hearing aids with special vibrators are sometimes used to deliver sound to the mastoid process (the part of the temporal bone that is behind the ear) and then by bone conduction to the inner ear. Bone conduction is also the basis of some of the oldest and most useful tests in the repertory of the otologist, those employing tuning forks to distinguish between a conductive impairment affecting the middle ear and amenable to surgery and a sensorineural impairment—that is, an impairment affecting the inner ear and the cochlear nerve—for which surgery is usually not indicated. In the Rinne test the sounding tuning fork is placed on the mastoid process, and the person being tested is asked to report when he can no longer hear it. The examiner then removes the fork quickly and holds the prongs close to the open ear canal. The normal ear continues to hear it for about 45 seconds, and this "positive" result occurs also with incomplete sensorineural impairment of hearing. When the result is "negative" and the fork is heard longer by bone conduction than by air conduction, a conductive type of deafness is present. In the Schwabach test the presence of a sensorineural impairment is indicated when the patient is unable to hear the bone-conducted sound as long as the examiner with normal hearing can. The patient with a conductive hearing loss, on the other hand, can hear the fork longer than can the examiner because the conductive lesion excludes the extraneous masking noise of the surroundings. A bone-conduction audiometer would give the same result.

For the Weber test, the fork is simply placed on the patient's forehead, and the examiner asks in which ear he hears it. If a sensorineural lesion is present in one ear, the patient will localize the sound in the opposite, or "better," ear. If a conductive lesion is present, he will localize it in the "worse" ear—*i.e.,* in the one that is protected from interference by extraneous sounds. This simple test is a valuable aid in the diagnosis of otosclerosis.

Transmission through liquid of inner ear to organ of Corti. The vibrations of the stapes footplate at the oval window are transmitted through the perilymph of the vestibule and that of the scala vestibuli to the cochlear duct. Normally, they do not affect the semicircular canals, the utricle, or the saccule. Only the cochlear fluids and the membranes of the cochlear duct are free to vibrate in response to alternating pressures at the oval window, because only the cochlea has the round window as a "relief valve."

What takes place in the cochlea is both an analysis or sorting out of the frequencies of a complex sound and a transduction of the mechanical vibrations into nerve impulses for transmission to the brainstem by the fibres of the cochlear nerve. The analysis occurs because the cochlea is a tuned structure, different parts of which vibrate in response to different frequencies; the transduction because the hair cells of Corti's organ are capable of changing minute amounts of mechanical energy into an electrochemical form that stimulates the endings of the nerve fibres.

The ear's analysis of complex sounds

The idea of the inner ear as a resonant structure was proposed by several authors in the 17th and 18th centuries but received its most explicit statement just over a century ago by the German physicist and physiologist Hermann von Helmholtz. Inspired by the anatomical studies of Corti, Helmholtz postulated in the cochlea a series of resonators capable of analyzing complex sounds into their component frequencies. After considering various other structures, he concluded that the transverse fibres of the basilar membrane, which increase in length and decrease in stiffness from the basal end to the apex, were the resonators he sought. Although Helmholtz' resonance theory in its original form is no longer accepted, much experimental and clinical evidence supports the closely related "place theory," which holds that sounds of different frequency activate different regions of the basilar membrane and organ of Corti, there being an orderly progression from the basal end, where the highest tones are effective, to the apical end, where the lowest tones are received.

From the experiments of Georg von Békésy (for which he was awarded the Nobel Prize for Physiology or Medicine in 1961) it is clear that the frequency analysis performed by the cochlea occurs not because of a series of separate tuned resonators as postulated by Helmholtz but because the basilar membrane and organ of Corti vary continuously in stiffness and mass from base to apex, stiffness decreasing and mass increasing as the width of the basilar membrane increases. Vibrations reaching the basal end through the perilymph can be shown to proceed as travelling waves, attaining a maximum amplitude at a certain point along the membrane and then rapidly subsiding. The higher the frequency of the sound imposed, the shorter is the distance that the waves travel. Since the hair cells are arrayed in orderly ranks on the basilar membrane and the nerve fibres fan out in orderly fashion to innervate them, a tone of a given frequency causes a certain "place" on the basilar membrane to vibrate, the hair cells overlying it are stimulated, and the corresponding nerve fibres convey impulses to the brain.

The brain can recognize the place on the basilar membrane and thus the pitch of the stimulating tone, depending upon the particular group of nerve fibres activated. For the lower frequencies the rate of stimulation is also an important indicator of pitch because the frequency of the nerve impulses tends to follow the frequency of the tone. For the higher frequencies, place alone seems to be decisive. The intimate nature of the events that occur at the hair cells when they are stimulated by movements of the basilar membrane is by no means fully understood. The up-and-down movements of the basilar membrane are thought to be converted into shearing movements between the reticular lamina and the tectorial membrane, which overlies it. The stereocilia, or hairs of the hair cells, in contact with the tectorial membrane are displaced or bent by the shearing forces, and it is this bending of the stereocilia that triggers the electrochemical events in the hair cells that excite the endings of the cochlear nerve fibres.

Present understanding of the events occurring in the cochlea is based in part on von Bêkésy's experiments with cochlear models, in part on his direct microscopic observations by stroboscopic illumination of vibratory patterns in the human cochlea removed from the body after death and subjected to intense sound. Important insights have also been obtained through the study of the small electrical potentials that are produced by the living cochlea in response to sound. These alternating current potentials, first reported in 1929, reproduce the frequency and wave form of the stimulating sound and can be picked up by means of two electrodes, one of which is placed in contact with the round window membrane or with the bony wall of the cochlea. The potentials were shown to consist of two separate components, the microphonic potentials, which appear to be related to the transduction process at the hair cells, and the action potentials, which represent the nerve impulses in the terminal fibres of the cochlear nerve. There is also a direct current potential difference, the endocochlear potential of some 80 millivolts between the endolymph in the cochlear duct and the perilymph in the scala tympani. The alternating current potentials depend upon an intact organ of Corti, the direct current potential upon the stria vascularis. All require an adequate supply of oxygen to the cochlea and soon disappear when oxygen is lacking. When the organ of Corti is damaged by drugs or by intense sound, the potentials are diminished or abolished. In young children suspected of congenital deafness, the cochlear potentials can give important information about the state of the inner ear.

Electrical potentials of the cochlea

THE COCHLEAR NERVE AND THE CENTRAL AUDITORY PATHWAYS

The vestibulocochlear (acoustic, or eighth cranial) nerve consists of two anatomically and functionally distinct parts, the cochlear nerve, distributed to the organ of hearing, and the vestibular nerve, distributed to the organs of equilibrium. The fibres of the cochlear nerve have their peripheral terminals around the bases of the inner and outer hair cells and their central terminals in the groups of nerve cells called the dorsal and ventral cochlear nuclei, in the medulla oblongata, at the base of the brain. The cell bodies of these neurons (nerve cells), numbering some 30,000 in all, fill the spiral canal of Rosenthal at the root of the osseous lamina of the cochlea. Their peripheral, dendritic portions extend radially in the lamina to the habenula perforata, beneath the inner hair cells. At this point they lose their myelin sheaths and enter the organ of Corti as thin, unmyelinated fibres. The greater number are distributed as radial fibres directly to the inner hair cells, whereas others cross the tunnel either as radial fibres supplying a few outer hair cells or as spiral fibres that turn to run for relatively long distances and make contact with many outer hair cells. The central or axonal portions of the bipolar cochlear neurons unite to form the cochlear nerve trunk, the fibres of which are twisted like the strands of a rope. Leaving the modiolus through the internal meatus or passageway, they pass directly to the medulla. There each fibre divides into two, sending one branch to the dorsal and the other to the ventral cochlear nucleus.

Central auditory pathways

The central auditory pathways extend from the medulla to the cerebral cortex and consist of a series of nuclei (groups of nerve cell bodies) connected by fibre tracts made up of their axons. They form a more or less direct route for relaying acoustic information, encoded in the form of nerve impulses, directly to the highest cerebral levels in the cortex. At lower levels information as to pitch, loudness, and localization of sounds is processed, and appropriate responses, such as contractions of the intra-aural muscles, turning of the eyes and head, or movements of the body as a whole, are initiated.

Some fibres from the ventral cochlear nucleus pass across the midline to the cells of the superior olivary complex, whereas others make connection with the olivary complex of the same side. Together, these fibres form the trapezoid body. Fibres from the dorsal cochlear nucleus cross the midline to end on the cells of the nuclei of the lateral lemniscus. There they are joined by fibres from the ventral cochlear nuclei of both sides, and from the olivary complex. The lemniscus is a major tract, most of the fibres of which end in the inferior colliculus, the auditory centre of the midbrain, although some fibres may by-pass the colliculus and end at the next higher level, the medial geniculate body. From the medial geniculate there is an orderly projection of fibres to a portion of the cortex of the temporal lobe.

In man and other primates the primary acoustic area in the cerebral cortex is in the superior transverse temporal gyrus, a ridge in the temporal lobe, on the lower edge, or lip, of a cleft known as the sylvian fissure.

Because about half of the fibres of the auditory pathways cross the midline while others ascend on the same side, each ear is represented in both the right and the left cortex. For this reason both ears can continue to function normally, even if the auditory cortical area of one side is destroyed. Impaired hearing due to bilateral cortical injury involving both auditory areas has been reported, but it is extremely rare.

Parallel with the pathway ascending from the cochlear nuclei to the cortex is a pathway descending from the cortex to the cochlear nuclei. In both pathways some of the fibres remain on the same side, while others cross the midline to the opposite side of the brain. There is also evidence of a "spur" line ascending from the dorsal cochlear nucleus to the cerebellum and another descending from the inferior colliculus, the auditory centre of the midbrain, to the cerebellum. The significance of these cerebellar connections is not clear, but they may antedate the development of the cerebral cortex. In general, the descending fibres may be regarded as exercising an inhibitory function by means of a sort of "negative feedback." They may also determine which ascending impulses shall be blocked and which shall be allowed to pass on to the higher centres of the brain.

From the olivary complex, a region in the medulla oblongata, there arises also a fibre tract called the olivocochlear bundle. It constitutes an efferent system or feedback loop, by which nerve impulses, thought to be inhibitory in nature, reach the hair cells. This system, which apparently utilizes acetylcholine as a chemical synaptic transmitter, is presumably involved in sharpening the analysis that is made in the cochlea.

Evidence of an orderly spatial representation of the organ of Corti at the lower levels of the auditory pathway has been reported by many investigators. Such a pattern would seem to be required by the place theory of cochlear analysis of sound.

Physiological evidence of "tuning" of the auditory system has also been obtained by recording the electrical potentials from individual neurons at various levels by means of microelectrodes. Most neurons of the auditory pathway show a "best frequency"—*i.e.,* a frequency to which the individual neuron responds at minimal intensity. With each increase in the intensity of the sound stimulus, the neuron is able to respond to a wider band of frequencies, thus reflecting the broad tuning of the basilar membrane. With sounds of lower frequency, the rate of response of the neuron tends to reflect the stimulus frequency. Increased intensity of stimulation causes a more rapid rate of responding. In general, pitch tends to be coded in terms of the neurons that are responding, loudness in terms of the rate of response and the total number of active neurons.

Auditory function of the cortex

Although extensive studies have been made of the responses of single cortical neurons, the data do not yet fit any comprehensive theory of auditory analysis. Experiments in animals have indicated that the cortex is not even necessary for frequency recognition, which can be carried out at lower levels, but that it is essential for the recognition of temporal patterns of sound. It appears likely, therefore, that in man both pitch and loudness are distinguished at lower levels of the auditory pathways and that the cortex is reserved for the analysis of more complex acoustic stimuli, such as speech and music, for which the temporal sequence of sounds is equally important.

Presumably it is also at cortical levels that the "meaning" of sounds is recognized and behaviour is adjusted in accordance with their significance. Such functions were formerly attributed to an "auditory association area" immediately surrounding the primary area, but such a term might equally well be interpreted as embracing the entire cerebral cortex, thanks to the multiple interconnections between the various areas.

The localization of sounds is known to depend upon the recognition of minute differences in intensity and in the time of arrival of the sound at the two ears. A sound that arrives at the right ear a few microseconds sooner than it does at the left or that sounds a few decibels louder in that ear is recognized as coming from the right. In a real-life situation the head may also be turned to pinpoint the source of sound by maximizing these differences. For low-frequency tones a difference in phase at the two ears is the criterion for localization, but for higher frequencies the difference in loudness caused by the sound shadow of the head becomes all-important.

The superior olivary complex

Each cochlear nucleus receives impulses only from the ear of the same side. A comparison between the responses of the two ears first becomes possible at the superior olivary complex, which receives fibres from both ears. Electrophysiological experiments in animals have shown that some units of the accessory nucleus of the olivary complex respond to impulses from both ears. Others respond to impulses from one side exclusively, but their response is modified by the simultaneous arrival of impulses from the other side.

This system appears to be capable of making the extraordinarily fine discriminations of time and intensity that are necessary for sound localization. By virtue of such complex neural interconnections in the brain, the two ears together can be much more effective than one ear alone in picking out a particular sound in the presence of a background of noise. They also permit attention to be directed to a single source of sound, such as one instrument in an orchestra or one voice in a crowd. This is the basis of the "cocktail-party effect," whereby a listener with normal hearing can attend to different conversations in turn or concentrate on one speaker despite the surrounding babble. Whether the muscles within the ear play a part in filtering out unwanted sounds during such selective listening has not been established.

HEARING TESTS: AUDIOMETRY

Before the development of electroacoustic equipment for generating and measuring sound, the tests of hearing at the disposal of the otologist gave approximate answers at best. A patient's ability to hear could be specified in terms of whether he could distinguish the ticking of a watch or the clicking of coins or at what distance he could understand a whispered voice. Alternatively, the examiner might note the length of time a patient could hear the gradually diminishing note of the tuning fork, comparing the performance with his own. Other specialized tuning fork tests have been described above.

The electronic audiometer, introduced in the 1930s, makes it possible to measure the patient's threshold of hearing for a series of pure tones ranging from a lower frequency of 125 hertz to an upper frequency of 8,000 or 10,000 hertz. This span includes the three octaves between 500 and 4,000 hertz that are most important for speech.

Components of an audiometer

The audiometer consists of an oscillator or signal generator, an amplifier, a device called an attenuator, which controls and specifies the intensity of the tones produced, and an earphone. The intensity range is usually 100 decibels in steps of five decibels. The "zero dB" level represents "normal hearing" for young adults under favourable, noise-free laboratory conditions and was established in 1964 as an international standard.

For pure-tone audiometry, the patient wears the earphone and is asked simply to indicate when he hears a tone. The examiner proceeds to determine the lowest intensity for each frequency at which the patient reports that he is just able to hear the tone 50 percent of the time. If, for example, he hears 4,000 hertz only at 40 decibels, he is said to have a 40-decibel hearing level for that frequency—*i.e.,* a threshold 40 decibels above the normal threshold. A graph showing the hearing level for each frequency is called an audiogram. The shape of the audiogram for a hard-of-hearing patient can give the otologist or audiologist important information for diagnosing the nature and cause of the defective hearing. (The otologist is concerned with diseases of the ear; the audiologist focusses his attention on the measurement of hearing impairment.)

With the Békésy automatic recording audiometer, the patient himself controls the level of the tone presented. He is required to press a button so long as he is able to hear the tone, which is automatically reduced in intensity as the button is pressed. When it is released the intensity increases until the patient again signals that he hears it by pressing the button. In this way the patient's threshold is "tracked." At the same time the frequency is slowly increased and a graphic recording is made. Thus a complete, continuous audiogram can be obtained in less than 10 minutes.

A calibrated bone conduction vibrator is usually furnished with the audiometer so that hearing by bone conduction can also be measured. In the presence of otosclerosis or other conductive defect of the middle ear, there may be a sizable difference between the air-conduction and bone-conduction audiograms, the so-called "air-bone gap." This difference is a measure of the loss in transmission across the middle ear and indicates the maximum improvement that may be obtained through successful corrective surgery. When the defect is confined to the organ of Corti, the bone-conduction audiogram shows a loss similar to that for air conduction. In such cases of sensorineural impairment, surgery is seldom if ever capable of improving the hearing, but a hearing aid may prove useful.

Although faint sounds may not be heard at all by the ear with a sensorineural impairment, more intense sounds may be as loud as to a normal ear. This rapid increase in loudness above the threshold level is called recruitment. When the opposite ear has normal hearing, recruitment can be measured by the alternate binaural loudness balance (ABLB) test. The subject is asked to set the controls so that the loudness of the tone heard in his defective ear matches that of the tone heard in his normal ear. By repetition of the comparison at several intensity levels, the presence or absence of recruitment can be demonstrated. When recruitment is excessive, the range of useful hearing between the threshold and the level at which loudness becomes uncomfortable or intolerable may be extremely narrow, so that the amplification provided by a hearing aid is of limited value to the patient.

Speech reception threshold and discrimination

Although hearing thresholds for pure tone give some indication of the patient's hearing for speech, direct measurement of this ability is of interest to the otologist. Two types of tests are most often used. In one, the speech reception threshold (SRT) is measured by presenting words of spondee pattern—words containing two syllables of equal emphasis, as in "baseball" or "cowboy"—at various intensity levels until the level is found at which the patient can just hear and repeat half of the words correctly. This level usually corresponds closely to the average of the patient's thresholds for frequencies of 500, 1,000, and 2,000 hertz. A more important measure of socially useful hearing is the discrimination score. For this test a list of selected monosyllabic words is presented at a comfortable intensity level, and the subject is scored in terms of the percentage of the words heard correctly. This test is helpful in evaluating certain forms of hearing impairment in which the sounds may be audible but words remain unintelligible. Such tests are usually carried out in a quiet, sound-treated room that excludes extraneous noise. They may give a mistaken impression of the ability of the patient with sensorineural impairment to understand speech in ordinary noisy surroundings. Because such persons often have increased difficulty in understanding speech in the presence of noise, speech tests are best carried out against a standardized noise background as well as in the quiet. A person with a conductive defect may be less disturbed by noise than may the normal subject.

Tests suitable for young persons

For very young persons or others who are unable to cooperate in the usual audiometric tests, thresholds for pure tones may be established by electrophysiological means. A brief tone causes a small variation in the electrical potentials that can be recorded from the scalp and constitute the electroencephalogram, EEG (a recording of the brain waves). By repetition of the stimulus up to 100 times and by an averaging of the responses in a small computer, the responses can be selectively enhanced while the more or less random background of elctrical activity is cancelled out. In this way auditory thresholds can be established that closely approximate those obtained in conventional audiometry. If the responses indicate impaired hearing, they give little or no indication of the site of the defect.

Another form of electrophysiological hearing test is the electrocochleogram. Electrical potentials representing impulses in the cochlear nerve are recorded from the cochlea by means of a fine, insulated needle electrode inserted through the eardrum membrane to make contact with the promontory of the basal turn. This test gives a direct measure of cochlear function.

More elaborate tests, often involving speech or sound localization, are available for testing hearing impaired by defects of the central nervous system. The interpretation of the results is often difficult, and the diagnostic information furnished by the tests is seldom clear-cut.

A simple and objective means of testing hearing at the level of the cochlea and brainstem is supplied by impedance audiometry. Two small tubes are sealed into the external canal. Through one tube sound from a small loud-speaker is injected into the canal. The portion that is reflected from the drum membrane is picked up by the other tube and led to a microphone, amplifier, and recorder. When a sudden, moderately intense sound is applied to the opposite ear, the stapedius muscle contracts, the impedance is increased, and the recorder shows a slight excursion as more sound is reflected. This test can give valuable information not only about the condition of the cochlea and the auditory pathways of the medulla but also about the facial nerve that supplies fibres to the stapedius muscle. On the other hand, it does not give an actual measurement of the acoustic impedance of the ear, representing the state of the ossicular chain and the mobility of the eardrum. This information can be obtained by means of the acoustic bridge—a device that enables the observer to listen to a sound as reflected from the subject's eardrum and at the same time to a similar sound of equal intensity as reflected in an artificial cavity, the volume of which is adjusted to equal that of the external canal of the ear being tested. When the two sounds are matched by varying the acoustic impedance of the cavity, the impedance of the ear is equal to that of the cavity, which can be read directly from the scale of the instrument. Conductive defects of the middle ear, including discontinuity (disarticulation) of the ossicular chain and immobility of the malleus or stapes, can be recognized by the characteristic changes they cause in the impedance of the ear. Profound sensorineural deafness can occur as a result of viral and other infections, including mumps, measles, and meningitis. Rubella (German measles) in the mother during pregnancy can cause the child to be born with a severely damaged organ of Corti and profound hearing loss. Cochlear abnormalities may be present also as a result of genetic defects. In all such cases of deafness in young children it is essential that the condition be recognized as early as possible so that special educational measures may be instituted to minimize the handicap. The newer electrophysiological hearing tests described above—*i.e.,* electroencephalographic audiometry and the electrocochleogram—offer the best possibilities for detecting such losses in infants. (For further discussion of deafness and hearing impairments, see below *Ear diseases and hearing disorders.*)

VESTIBULAR FUNCTION

Equilibrium sense

The vestibular system represents the equilibrium sense and is concerned primarily with controlling the position of the head and the posture of the body. The adequate stimulus for its end organs is acceleration: angular acceleration for the semicircular canals; linear acceleration for the utricle and saccule. Functionally these organs are closely related to the cerebellum and to the reflex centres that govern the movements of the eyes, the neck, and the limbs. The information they deliver is proprioceptive in character, dealing with events within the body itself, rather than exteroceptive, dealing with events outside the body, as in the case of the responses of the cochlea to sound. Although the vestibular organs and the cochlea are derived embryologically from the same formation, the otic vesicle, their association seems to be a matter more of convenience than of necessity. Both from the developmental and from the structural point of view, the kinship of the vestibular organs with the lateral line system of the fishes is easily recognized. This system is made up of organs located in the skin at the side of the body and contains cristae with innervated hair cells and a cupula, like those of the semicircular canals. They are sensitive to waterborne vibrations and to pressure changes.

The anatomists of the 17th and 18th centuries were impressed by the orientation of the semicircular canals in the head, in three planes more or less perpendicular to one another. Since it was generally assumed that the entire labyrinth is devoted to hearing, the suggestion was put forward that the canals must perceive the direction of sound.

The first investigator to present evidence that the vestibular labyrinth is the organ of equilibrium was a French experimental neurologist, Marie-Jean-Pierre Flourens, who, in 1824, reported a series of experiments in which he had produced abnormal head movements in the pigeon by cutting each of the semicircular canals in turn. The plane of the movements was always the same as that of the injured canal. Hearing was not affected by destroying the nerve fibres to the vestibular organs, but it was abolished by cutting the cochlear nerve.

Endolymph and the "hydrodynamic concept"

It was almost half a century later that the significance of these experiments was appreciated and the semicircular canals were recognized as specific sense organs concerned with the position of the head. The German physiologist Friedrich Goltz in 1870 suggested in his "hydrostatic concept" that the canals are stimulated by the weight of the fluid they contain, the pressure it exerts varying with the head position. Three years later the Austrian scientists Ernst Mach and Josef Breuer, and the Scottish chemist Crum Brown, working independently, proposed their "hydrodynamic concept." They held that head movements cause a flow of endolymph in the canals, which are stimulated by the fluid movement or pressure changes. A German physiologist, J.R. Ewald, showed that the compression of the lateral canal in the pigeon by a small pneumatic hammer caused endolymph movement toward the crista and turning of the head and eyes toward the opposite side. Decompression reversed both the direction of endolymph movement and the turning of the head and eyes.

Later investigators have proved that the hydrodynamic concept is correct and that the cupula is deflected by endolymph movement in response to rotation. They were able to keep track of the deflection in live fish by means of a droplet of oil injected into the canal. At the start of rotation in the plane of the canal, the cupula was deflected in the opposite direction and then returned slowly to its position of rest. At the end of rotation it was deflected again, this time in the same direction as the rotation, and then returned once more to its upright resting position. These deflections were due to the inertia of the endolymph, which tended to lag behind at the start of rotation and to continue in motion after rotation had stopped. The slow return was a function of the elasticity of the cupula itself.

Polarization of hair cells

Other researchers found that they could keep the labyrinth of a cartilaginous fish, the thornback ray (*Raja clavata*), active for some time after it was removed from the animal and could record from the vestibular fibres of the vestibulocochlear nerve impulses arising in one of the ampullar cristae. At rest there was a slow continuous discharge, which was increased by rotation in one direction and decreased by rotation in the other. In other words, the level of excitation rose or fell depending upon the direction of rotation. Later studies with the electron microscope showed a remarkable polarization of the hair cells of the ampullar cristae. Each hair cell of the horizontal canals has its kinocilium facing toward the utricle, whereas each hair cell of the vertical canals has its kinocilium facing away from the utricle. In the horizontal canals deflection of the cupula toward the utricle—*i.e.,* bending of the stereocilia toward the kinocilium—depolarizes the hair cells and increases the rate of discharge. Deflection away from the utricle causes hyperpolarization and decreased discharge. In the vertical canals these effects are reversed.

The relation between the two labyrinths is reciprocal. When the head is turned to the left, the discharge from the left horizontal canal is increased while that from the right horizontal canal is decreased, and vice versa. Normal posture is the result of their acting both in cooperation and in opposition. When one labyrinth is injured, the unrestrained activity of the other causes a continuous false sense of turning (vertigo) and rhythmical, jerky movements of the eyes (nystagmus), both toward the normal side. When both labyrinths are injured or destroyed, as by the action of the antibiotic streptomycin, there may be a serious disturbance of posture and gait (ataxia) as well as severe vertigo and general disorientation. In younger persons the disturbance tends to subside as reliance is placed on vision and on proprioceptive impulses from the muscles and joints to compensate for the loss of information from the labyrinth. In elderly persons, the loss of labyrinthine function may be disabling.

The two otolith organs or gravity receptors, the maculae of the utricle and saccule, are more or less perpendicular to each other. The left and right utricular maculae lie in the same approximately horizontal plane; the saccular maculae in parallel, vertical planes. Both types of receptors are stimulated by shearing forces between the otolithic membrane and the cilia of the hair cells beneath it. The otolithic membrane is covered with a layer of crystals of calcite (otoconia), which add to its weight and increase the shearing forces set up in response to a slight displacement when the head is tilted. These receptors, particularly the utricle, have an important role in the righting reflexes and in reflex maintenance of tonic contraction of the muscles that keep the body in the upright position. The role of the saccule is less completely understood, and some investigators have suggested that it is responsive to vibration as well as to rectilinear acceleration of the head in the sagittal (fore-and-aft) plane. Of the two receptors, the utricle appears to be the dominant partner.

Tests of vestibular function depend mainly upon stimulation of the semicircular canals. Rotation, which can cause vertigo and nystagmus, as well as temporary disorientation and a tendency to fall, stimulates both labyrinths simultaneously. Since the otoneurologist is usually more interested in examining the right and left labyrinths separately, he usually employs warmth as a stimulant. Syringing the ear canal with warm water at 44° C (111° F) or cool water at 30° C (86° F) elicits nystagmus by setting up convection currents in the horizontal canal. The duration of the nystagmus may be timed with a stopwatch, or the movements of the eyes can be accurately recorded by picking up the direct current potentials of the eyeballs by means of electrodes pasted to the skin of the temples (a process called electronystagmography, or ENG). An abnormal labyrinth usually gives a reduced response or none. (J.E.H.)

Caloric stimulation

Ear diseases and hearing disorders

The ear has two important but quite different functions: the maintenance of equilibrium and hearing. For biped locomotion, the sense of equilibrium is especially necessary; and it is even more pressing in the dark, in which the eyes cannot help to maintain balance, or on uneven terrain, where sensations from the feet do not assist in maintaining the sense of equilibrium. Hearing, which developed later than equilibrium in the process of evolution, has become, with sight, one of the senses most needed by animals for flight and survival. For humankind, hearing is also the chief means of receiving communication by language and is so taken for granted that its importance is rarely appreciated until it is impaired or lost.

The human auditory system and its natural function are treated in detail above. This section deals with the more important diseases and disorders of the outer, the middle, and the inner ear. It concludes with brief discussions of the causes of deafness and impaired hearing; rehabilitation of hearing impairment and deafness; and the social and economic implications of deafness.

THE OUTER EAR

Each of the three main parts of the ear is afflicted by a particular set of diseases, according to the structure, tissues, and function of that part. Diseases of the outer ear are those that afflict skin, cartilage, and the glands and hair follicles in the outer ear canal. The sound-transmitting function of the outer ear is impaired when the ear canal becomes filled with tumour, infected material, or earwax, so that sound cannot reach the eardrum membrane. The most common diseases of the outer ear are briefly described in the following paragraphs.

Infections and injuries. *Frostbite.* The exposed position of the outer ear makes it the part of the body most frequently involved by freezing or frostbite. Humidity, duration of exposure, and, most of all, wind, in addition to degrees of temperature below freezing, predispose to occurrence of frostbite. The frozen area begins along the upper and outer edge of the ear, which becomes yellow-white and waxy in appearance, cold and hard to the touch, and numb with loss of skin sensation.

In treatment of frostbite the victim is placed as soon as possible in a warm room, but the frozen ear is kept cool until the returning blood circulation gradually thaws the frozen part from within. Massage of the frozen ear is avoided, for it is likely to injure the skin. Heat applied to the frozen area before circulation is established can result in clotting of the blood in the blood vessels. This in turn can result in death of that part of the ear, which turns black and eventually falls off, a process called dry gangrene.

Hematoma. Injury to the outer ear can cause bleeding between the cartilage and the skin, producing a smooth, rounded, nontender purplish swelling called hematoma. The accumulation of clotted blood is removed by a surgeon because, if it is left, it will become transformed into scar tissue and cause a permanent, irregular thickening of the outer ear commonly called cauliflower ear and seen in boxers and wrestlers whose ears receive much abuse.

Infections of the outer ear

Perichondritis. Infection of the cartilage of the outer ear, called perichondritis, is unusual but may occur from injury or from swimming in polluted water. It is due to a particular microorganism, *Pseudomonas aeruginosa.* There is a greenish or brownish, musty or foul-smelling discharge from the outer ear canal, while the affected outer ear becomes tender, dusky red, and two to three times its normal thickness. Prompt medical treatment is necessary to prevent permanent deformity of the outer ear.

External otitis. Infection of the outer ear canal by molds or various microorganisms occurs especially in warm, humid climates and among swimmers. The ear canal itches and becomes tender; a small amount of thin, often foul-smelling material drains from it. If the canal becomes clogged by the swelling and drainage, hearing will be impaired. Careful and thorough cleaning of the outer ear canal by a physician, application of antiseptic or antibiotic eardrops, and avoidance of swimming are indicated to clear up the infection.

Boil in the ear (furuncle). Infection of a hair follicle anywhere on the body is known as a boil, or furuncle. This can occur in a hair follicle in the outer ear canal, especially when there is infection of the skin of the canal. It always occurs because of a particular type of germ known as staphylococcus. Because the skin of the ear canal is closely attached to the underlying cartilage, a boil in the ear canal is especially painful, with swelling, redness, and tenderness but generally without fever. Heat applied to the outer ear by a hot-water bottle or electric pad helps the infection to come to a head and begin to drain. Antiseptic eardrops and careful cleaning of the outer ear canal are needed to prevent other hair follicles from becoming infected with a series of painful boils in the ear.

Erysipelas of the outer ear. Erysipelas is an infection in the skin caused by a particular type of streptococcus that causes a slowly advancing red, slightly tender thickening of the skin. It begins at the ear and spreads to the face and neck. Centuries ago erysipelas epidemics caused severe and often fatal infections. In AD 1089 one of the most severe epidemics was known as St. Anthony's fire; those who prayed to St. Anthony were said to recover; others, who did not, died. Today erysipelas is a rather mild and comparatively rare infection that clears up rapidly when sulfanilamide is taken by mouth or penicillin by injection.

Leprosy. Leprosy, seen rarely outside of the tropics today, was another scourge of ancient times that sometimes affected the outer ear. It is caused by the leprosy bacillus, *Mycobacterium leprae,* which causes a painless, slowly progressing thickening and distortion of the affected tissues. The diagnosis is made by examining a bit of the infected tissue under a microscope and finding the leprosy bacilli, which in appearance are not unlike the bacilli that cause tuberculosis. Fortunately, the antibiotics effective against tuberculosis are effective today in arresting the ravages of leprosy.

Osteoma of the bony ear canal. Osteoma of the bony ear canal is a bony knob that grows close to the eardrum membrane, especially in those who swim a great deal in cold water. It is not dangerous and does not need to be removed unless the bony overgrowth becomes large enough to block the ear canal.

Cyst of the ear. A cyst is a sac filled with liquid or semisolid material. A cyst of the ear is most often caused by a gland that lubricates the skin behind the earlobe, less often at the entrance of the ear canal. If the duct of this gland becomes stopped, the lubricating fatty material accumulates as a soft rounded nodule in the skin. Infection of the cyst causes a tender abscess to form and drain. The cyst will re-form unless it is removed completely by surgery.

Another type of cyst occurs above the ear canal, just in front of the outer ear or, rarely, in the neck behind and below the ear. This is a remnant of the primitive gill of the early embryo, a reminder of our ancient fishy ancestors. It may appear as a tiny pitlike depression that discharges a little moisture from time to time, or a cystic swelling may develop when the opening of the pit is closed, requiring surgical removal.

Enlargement of earlobe

Keloid of the ear. In dark-skinned people, overgrowth of scar tissue from any skin incision or injury causes a thickened elevation on the scar called a keloid. Having the earlobes pierced for earrings sometimes results in a large, painless nodular keloid enlargement of the earlobe, harmless but unsightly. Keloids are removed surgically (see also INTEGUMENTARY SYSTEMS).

Deformities. *Absence of the outer ear.* Congenital deformity or absence of the outer ear, usually on one side, sometimes on both, is often accompanied by absence of the outer ear canal. This failure of the primitive gill structures to become properly transformed into the normal outer and middle ear is, in rare instances, hereditary. More often it occurs for no known reason. In some cases it can be traced to the damaging effects on the embryo of rubella (German measles) in the mother during the first three months of her pregnancy. Since the inner ear and nerves of equilibrium and hearing come from the otocyst, separate from the gill structure, in most cases of deformed or absent outer ear the hearing nerve is normal. Surgical construction of a new ear canal and eardrum membrane can then often improve the hearing, which has been impaired by the failure of sound conduction to reach the hearing nerve in the inner ear.

Lop ear. Lop ear, excessive protrusion of the ear from the side of the head, is a more frequent and less serious deformity of the outer ear. Surgery may be performed to bring the ears back to a more normal and less conspicuous position.

Other ailments. *Eczema.* Eczema of the skin of the outer ear, like eczema elsewhere, is an itching, scaling redness, sometimes with weeping of the affected skin. It is often the result of an allergy to a food or substance such as hair spray that comes in contact with the skin. The best treatment is discovery and avoidance of the causative agent.

Impacted earwax. The waxy substance produced by glands in the skin of the outer ear canal normally is carried outward by slow migration to the outer layers of skin. When wax is produced too rapidly it can accumulate as a hard plug, firmly filling the outer ear canal and blocking the passage of sound to the eardrum membrane, causing a painless impairment of hearing. Large plugs of earwax need to be removed by a physician. Smaller amounts may be softened by a a few drops of baby oil left in the ear overnight, then syringed out with warm water and a soft-rubber baby ear syringe.

Cancer of the outer ear. Cancer of the outer ear occurs chiefly in areas exposed for many years to the direct sun. A small and at first painless ulcer, with a dry scab covering it, that slowly enlarges and deepens may be a skin cancer. Removal of a small bit of tissue from the edge (biopsy) and examining it under a microscope comprise the method of diagnosis. Complete removal by surgery or properly applied irradiation is needed for a cure. Cancer that begins in the ear canal is more serious, for it may deepen into the bone before it is diagnosed. It is then more difficult to cure by removal. Cancers of the ear canal are rather rare, while cancers of the skin of the outer ear are more common, as well as more readily cured by removal.

THE MIDDLE EAR

Location of middle ear

The air-filled middle-ear cavity and the air cells in the mastoid bone that extend backward from it are located a third of the distance from the side of the head toward its centre. The brain cavity lies just above and behind the middle ear and mastoid air spaces, separated from them only by thin plates of bone. The nerve that supplies the muscles of expression in the face passes through the middle-ear cavity and mastoid bone; it, too, is separated from them by only a thin layer of bone. In some instances this bony covering is incomplete, so that the facial nerve lies directly against the mucous membrane that lines the middle ear and mastoid air cells. This mucous membrane, an extension of a similar mucus-producing membrane that lines the nose and upper part of the throat, extends all the way through the eustachian tube into the middle ear and mastoid. It is subject to the same allergic reactions and infections that afflict the nasal passages. Thus, an acute head cold or other infection of the nose and throat, such as measles or scarlet fever, may extend through the eustachian tube into the middle ear and mastoid air cells. The proximity of the brain cavity to the mastoid air cells is such that an infection, if severe and untreated, may lead to meningitis (inflammation of the covering of the brain) or brain abscess. The large vein that drains blood from the brain passes through the mastoid bone on its way to the jugular vein in the neck. Infection from the middle ear can extend to this vein, resulting in "blood poisoning" (infection of the bloodstream, also called septicemia). Paralysis of the facial nerve and infection extending from the middle ear to the labyrinth of the inner ear are other possible complications of middle-ear infection. All of these possibilities spring from the particular location of the small but important middle-ear cavity.

Acute middle-ear infection. Fortunately, acute middle-ear infections, called acute otitis media, are nearly always due to microorganisms that respond quickly to antibiotics. As a result, acute infection of the mastoid air cells resulting in a dangerous mastoid abscess with the possibility of meningitis, brain abscess, septicemia, infection of the labyrinth, or facial nerve paralysis, complicating an acute infection of the middle-ear cavity, have become rare. Abscess of the mastoid and the other complications of acute middle-ear infection are seen chiefly in remote regions and countries lacking adequate medical attention.

Chronic otitis media

While serious and life-threatening acute infections of the middle ear and mastoid air cells have become rare, chronic infections, mentioned below, continue to occur, and another type of middle-ear disease, secretory otitis media, is frequent.

Secretory otitis media. In secretory otitis media the middle-ear cavity becomes filled with a clear, pale-yellowish, noninfected fluid. The disorder is the result of inadequate ventilation of the middle ear through the eustachian tube. The air in the middle ear, when it is no longer replenished through this tube, is gradually absorbed by the mucous membrane, fluid taking its place. Eventually, the middle-ear cavity is completely filled with fluid instead of air. The vibratory movements of the eardrum membrane and the ossicular chain are impeded by the fluid, with a painless impairment of hearing.

The usual causes for secretory otitis media are an acute head cold with swelling of the membranes of the eustachian tube; an allergic reaction of the membranes in the eustachian tube; and an enlarged adenoid (nodule of lymphoid tissue) blocking the entrance to the eustachian tube. The condition is cured by finding and removing the cause and then removing the fluid from the middle-ear cavity, if it does not disappear by itself within a week or two. Removal of the fluid requires puncturing the eardrum membrane and forcing air through the eustachian tube to blow out the fluid. In some cases a tiny plastic tube is inserted through the eardrum membrane to aid in reestablishing normal ventilation of the middle-ear cavity. After a time, when the middle ear and hearing have returned to normal, this plastic tube is removed. The small hole left in the eardrum membrane quickly heals.

Aero-otitis media. The sudden change of altitude by a rapid descent in a nonpressurized or poorly pressurized plane during a head cold or allergic reaction may not permit the normal equalization of air pressure that occurs by the periodic opening of the eustachian tube on swallowing or yawning. The eardrum membrane becomes sharply retracted when the air pressure becomes less within than without, while the opening of the tube into the upper part of the throat becomes pressed tightly together by the increased air pressure in the throat, so that the tube cannot be opened by swallowing. A severe sense of pressure in the ear is accompanied by pain and a decrease in hearing. Sometimes the eardrum membrane ruptures because of the difference in pressure on its two sides. More often, the pain continues until the middle ear fills with fluid, or surgical puncture of the eardrum membrane is done. This condition occurring from airplane flights is called aero-otitis media. Usually, however, pain and hearing loss produced during a flight is of a temporary nature and disappears of its own accord.

Chronic middle-ear infection. Chronic infection of the middle ear occurs when there is a permanent perforation of the eardrum membrane that allows dust, water, and germs from the outer air to gain access to the middle-ear cavity. This results in a chronic drainage from the middle ear through the outer ear canal. There are two distinct types of chronic middle-ear infection, one relatively harmless, the other caused by a dangerous bone-invading process that leads, when neglected, to serious complications.

The harmless type of chronic middle-ear disease is recognized by a stringy, odourless, mucoid discharge that comes from the surface of the mucous membrane that lines the middle ear. Medical treatment with applications of antiseptic solutions and powders is all that is needed to dry up the chronic drainage. The perforation in the eardrum membrane may then be closed, restoring the normal structure and function of the ear with recovery of hearing.

The dangerous type of chronic middle-ear drainage is recognized by its foul-smelling discharge, often scanty in amount, coming from a bone-invading process beneath the mucous membrane. Such cases are usually caused by a condition known as cholesteatoma of the middle ear. This is an ingrowth of skin from the outer ear canal that forms a cyst within the middle ear and mastoid. If untreated, the cyst enlarges slowly but progressively, gradually eroding the bone until the cyst reaches the brain cavity or the nerve that supplies the muscles of the face or a semicircular canal of the inner ear. The infected material within the cyst then produces a complication: meningitis or brain abscess, paralysis of the facial nerve, or infection of the labyrinth of the inner ear with vertigo, often leading to total deafness.

Dangers of neglected cholesteatoma

Fortunately, cholesteatoma of the middle ear is now rarely so neglected as to permit development of a serious complication. By careful examination of the eardrum-membrane perforation and by X-ray studies, the bone-eroding cyst can be diagnosed; it can then be removed surgically before it has caused serious harm. This operation is known as a radical mastoid or a modified radical mastoid operation. If at the same procedure the perforation in the eardrum membrane is closed and the ossicular chain repaired, the operation is known as a tympanoplasty, or plastic reconstruction of the middle-ear cavity.

Ossicular interruption. The ossicular chain of three tiny bones needed to carry sound vibrations from the eardrum membrane to the fluid that fills the inner ear may be disrupted by infection or by a jarring blow on the head. Most often the separation occurs at its weakest point, where the anvil (incus) joins the stirrup (stapes). If the separation is partial there is a mild impairment of hearing; if it is complete there is a severe hearing loss. Testing of the hearing in such a case demonstrates that the nerve of hearing in the inner ear is functioning normally but that sound fails to be conducted from the eardrum membrane to the inner ear. The separated ossicles can be brought back together by repositioning, thus restoring the conduction of sound to the inner ear. This is one of the most successful of hearing-restoring operations.

Otosclerosis. The commonest cause for progressive hearing loss in early and middle adult life is a disease process of the hard shell of bone that surrounds the labyrinth of the inner ear. This disease of bone is known as otosclerosis, a name that is misleading, for in its early and actively expanding stage the nodule of diseased bone is softer than the ivory-hard bone that it replaces. The more appropriate name otospongiosis is sometimes used, but such is the tenacity of tradition that the older name, applied before the process was well understood, has persisted and is the term generally used.

The cause for the occurrence of the nodule of softened otosclerotic bone is unknown. There is a certain familial tendency, half the cases occurring in families in which one or several relatives have the same condition. It is one-tenth as common among blacks as among whites, and is twice as common in women as in men. The nodule of softened otosclerotic bone first appears in late childhood or in early adult life. Fortunately, in most cases it remains quite small and harmless, producing no symptoms, and is discoverable only if the ear bones are removed after death and examined under a microscope. Such evidence indicates that approximately one in 10 white adult men and one in five white adult women will be found to have such a nodule of otosclerotic bone by middle adult life.

In about 12 percent of cases of otosclerosis the nodule of softened bone becomes large enough to reach the oval window containing the footplate of the stirrup. Increasing pressure caused by the expanding nodule begins to impede its vibratory movements in response to sound striking the eardrum membrane. Gradually and insidiously, such an affected person begins to lose his sharpness of hearing. First he begins to lose the ability to hear faint sounds of low pitch; next he begins to have difficulty hearing the whispered voice; then he has difficulty in hearing conversation from a distance; and finally he can hear and understand the spoken voice only when it is quite loud or close to the ear. One of the characteristics of impaired hearing due to stirrup fixation by otosclerosis is retained ability to hear over the telephone by pressing the receiver against the head so that the sound is carried to the inner ear by bone conduction. Another characteristic of this type of impaired hearing is that hearing seems to be better while one is riding in an automobile, in a plane, or on a train. The reason is that the low-pitched roar of motors causes normally hearing persons to unconsciously raise their voices, while the individual with stirrup fixation fails to hear the low-pitched roar and thus hears better and enjoys the raised voices around him.

Diagnosis of stirrup fixation

The diagnosis of stirrup fixation by otosclerosis is made on the basis of a history of a gradually increasing impairment of hearing with absence of any chronic infection of the middle ear or of perforation of the eardrum membrane and with hearing tests showing that the nerve of hearing in the inner ear is functioning but that sound fails to be conducted properly to it. The hearing tests demonstrate that the hearing by bone conduction is better than by air conduction.

The final and conclusive diagnosis of otosclerosis is made by surgical exploration and finding that the stirrup bone (stapes) is fixed and unable to be moved because of a nodule of bone that has grown against it. A special type of X-ray of the ear called polytomography is sometimes used to demonstrate that the footplate of the stirrup bone has been invaded by otosclerosis.

Fixation of the stirrup bone can be corrected surgically. This was accomplished formerly by constructing a new window into the inner ear to admit sound to the hearing nerve. This operation originated in Europe in 1924 and in 1937 was brought to the United States, where it was improved and named the fenestration operation. In 1952 it was found possible to mobilize (loosen) the fixed stirrup bone in some cases, thus restoring hearing without the need of constructing a new opening. In 1956 it was found that the fixed stirrup bone could be removed and replaced by a plastic or wire substitute in cases in which it could not be mobilized. Today this operation, known as stapedectomy, is the one most often used to correct fixation of the stapes (stirrup bone) by otosclerosis.

The otosclerotic bone disease in some cases expands as far as the cochlea of the inner ear, causing a gradual deterioration of the hearing nerve. This progressive nerve deafness may precede, accompany, or follow fixation of the stapes. In some cases it may occur without fixation of the stapes.

While the exact cause for the softening of a nodule of bone known as otosclerosis is not known, it may be associated in some cases with lack of fluoride in drinking water. There is evidence, not yet conclusive, that increasing the intake of fluoride may promote hardening of the softened nodule of otosclerotic bone, thus arresting or retarding its expansion. In this way it is possible that the gradual impairment of hearing-nerve function that often occurs with fixation of the stapes may be retarded or arrested.

THE INNER EAR

The labyrinth of the inner ear contains the nerve endings of the vestibular nerve—the nerve of equilibrium—and the auditory nerve, or nerve of hearing. The vestibular-nerve ends supply the semicircular canals and the otolithic membranes in the vestibule. The auditory nerve supplies the cochlea (see above *Anatomy of the auditory apparatus: inner ear).* Diseases of the labyrinth of the inner ear may affect both the vestibular nerve and the auditory nerve; or they may affect only the auditory nerve, with loss of hearing, or the vestibular nerve, bringing on vertigo. The commoner inner-ear diseases are touched upon in the following paragraphs.

Nerve deafness. *Congenital nerve deafness.* Congenital nerve deafness, a defect of the hearing nerve in the cochlea, may be present at birth or acquired during or soon after birth. Usually both inner ears are affected to a similar degree, and as a rule there is a severe impairment of hearing, although, in some cases of congenital nerve loss the impairment is moderate in degree. Many cases of congenital nerve deafness have been caused by the rubella (German measles) virus in the mother during the first three months of her pregnancy, causing arrest of development of the otocyst. This can happen during a rubella epidemic, even when the mother has no symptoms of the infection. In most cases the vestibular nerve is not affected or is affected to a lesser degree, and in most (but not all) cases the outer- and middle-ear structures are not affected. A vaccine against the rubella virus that has recently been introduced promises to result in a marked reduction in the number of cases of congenital nerve deafness if it is given to prospective mothers who have not had rubella before becoming pregnant and thus have not had an opportunity to build up immunity against infection when they are carrying a child.

Anoxia and kernicterus

Congenital nerve deafness, acquired at or soon after birth, may result from insufficient oxygen (anoxia) during a difficult and prolonged delivery or from the condition known as kernicterus, in which the baby becomes jaundiced because of incompatibility between its blood and that of the mother. In a few cases congenital nerve deafness is an inherited failure of the cochlea to develop properly. There is no medical or surgical treatment that can improve or restore hearing in cases of congenital nerve deafness. When the hearing loss is severe, speech cannot be acquired without special training. Children so afflicted must attend special classes or schools for the severely deafened, where they can be taught lipreading and speech. Electrical hearing aids can be helpful, especially during classes, to utilize the remnants of hearing usually present in such cases.

Viral nerve deafness. Virus infections can cause severe degrees of hearing-nerve loss in one ear and sometimes in both, at any age. The mumps virus is one of the most common causes of severe hearing-nerve loss in one ear. The measles and influenza viruses are less common causes. There is no effective medical or surgical treatment to restore hearing impaired by a virus.

Effect of injury and trauma. *Ototoxic drugs.* Ototoxic (ear-poisoning) drugs can cause temporary and sometimes permanent impairment of hearing-nerve function. Salicylates such as aspirin in large enough doses may cause ringing in the ears and then a temporary decrease in hearing that recovers when the person stops taking the drug. Quinine can have a similar effect but with a permanent impairment of hearing-nerve function in some cases. Certain antibiotics, such as streptomycin, dihydrostreptomycin, neomycin, and kanamycin, may cause permanent damage to the hearing-nerve function. The susceptibility to damage to the hearing from ototoxic drugs varies greatly among individuals. In most cases, except when streptomycin is the drug taken, the more durable and less easily damaged vestibular-nerve function is not affected. Streptomycin affects the vestibular nerve more than the auditory nerve.

Skull fracture and concussion. Skull fracture and concussion from a severe blow on the head can impair the functioning of the hearing nerve and of the nerve of balance in varying degrees. The greatest hearing loss arises when a fracture of the skull passes through the labyrinth of the inner ear, totally destroying its function.

Exposure to noise. The effects of noise exposure on hearing depend on the intensity and duration of the noise. The effects may be temporary or permanent. A single exposure to an extremely intense sound, such as an explosion, may produce a severe and permanent loss of hearing. Repeated exposures to sounds in excess of 80 to 90 decibels may cause gradual deterioration of hearing by destroying the hair cells of the inner ear and possible subsequent degeneration of nerve fibres (Figure 59). The levels of noise produced by rock music bands frequently exceed 110 decibels. The noise generated by farm tractors, power mowers, and snowmobiles may reach 100 decibels. In the United States, legislation requires that workers exposed to sound levels greater than 90 decibels for an eight-hour day be provided some form of protection, such as earplugs or earmuffs.

Sources of high level noise

By courtesy of Lars-Goran Johnsson

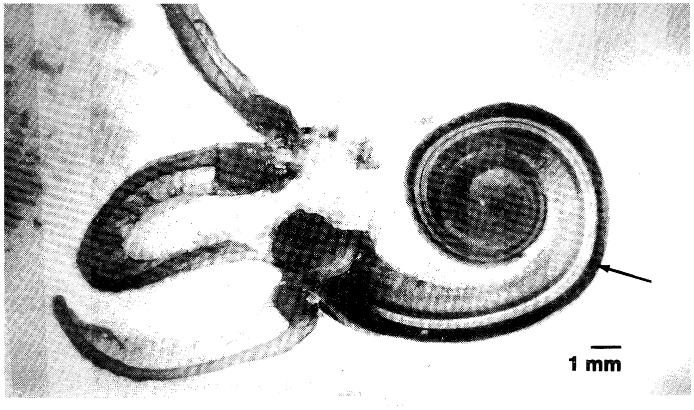


Figure 59: Dissection of the human cochlea and semicircular canals. The arrow indicates an area of degeneration of Corti's organ and the cochlear nerve fibres in the basal turn, corresponding to a loss of hearing for frequencies of 4,000 hertz and higher, presumably caused by noise exposure.

Individuals differ in their susceptibility to hearing loss from noise exposure. Because hearing loss typically begins at the higher frequencies of 4,000 to 6,000 cycles per second, the effects of noise exposure may go unnoticed until the hearing loss spreads to the lower frequencies of 1,000 to 2,000 cycles per second.

Inflammation and tumour. *Labyrinthitis.* Labyrinthitis, an inflammation of the labyrinth of the inner ear, occurs when microorganisms enter as a result of meningitis, syphilis, acute otitis media and mastoiditis, or chronic otitis media and cholesteatoma. Loss of both equilibrium and hearing occurs in the affected ear. Prompt treatment sometimes arrests the damage with the possibility of partial recovery of the function of the inner ear.

Acoustic neuroma. An acoustic neuroma is a benign tumour that grows on the acoustic nerve near the point where it enters the labyrinth of the inner ear. The tumour causes gradual and progressive loss of auditory- and vestibular-nerve function on one side. Eventually the tumour grows out into the brain cavity, causing headache and paralysis. If it is not removed, blindness and death may result. Fortunately, acoustic neuroma can be diagnosed early and removed before it has serious consequences.

Menière's disease. Menière's disease, also called endolymphatic hydrops, is a fairly common involvement of the labyrinth of the inner ear that affects both the vestibular nerve, with resultant attacks of vertigo, and the auditory nerve, with impairment of hearing. It was first described in 1861 by a French physician, Prosper Menière. It is now known that the symptoms are caused by an excess of endolymphatic fluid in the inner ear. The diagnosis is made from the recurring attacks of vertigo, often with nausea and vomiting, ringing or roaring in the ear, impairment of hearing with a distortion of sound in the affected ear that fluctuates in degree, and a sense of fullness or pressure in the ear. The cause of Menière's disease is not always known, although in many cases it results from defective functioning of the endolymphatic duct and sac, the structures that normally resorb endolymphatic fluid from the inner ear as fast as it is produced. The treatment of Menière's disease is directed toward controlling the excess of endolymphatic fluid. If medical treatment does not relieve the repeated attacks of vertigo, surgery may be necessary.

Presbycusis. Presbycusis, the gradual decline of hearing function of old age, is similar to other aging processes because it affects some people more rapidly and at an earlier age than others. Usually the slow diminishing of hearing does not begin until after age 60. The affected individual notices increasing difficulty in hearing sounds of high pitch and in understanding conversation. There is no medical or surgical treatment that can restore hearing in uncomplicated presbycusis. The physician must make certain that the patient does not have a correctible impairment, such as accumulated earwax, secretory otitis media, or stirrup fixation by otosclerosis, as part of his difficulty. An electrical hearing aid is of limited help to some, while others find that a hearing aid makes voices louder but less clear and therefore is of little help. Lipreading must then be used.

DEAFNESS AND IMPAIRED HEARING

Causes. Impaired hearing is, with rare exception, the result of disease or abnormality of the outer, middle, or inner ear, located in the temporal bone of the skull. One rare exception has already been mentioned, loss of hearing, usually on one side only, caused by a tumour, acoustic neuroma, on the acoustic nerve; an even rarer exception is impairment because of a brain lesion. Since nerve fibres from each ear go to both sides of the brain, a brain tumour or stroke causing paralysis of one side of the body rarely affects hearing.

Serious impairment of hearing (deafness) at birth is nearly always of the nerve type and cannot be improved by medical or surgical treatment. Nerve deafness arising from rubella infection of the mother during the first three months of her pregnancy, from insufficient oxygen during birth, and from kernicterus have already been mentioned.

Causes in childhood

In early and late childhood the most frequent cause for impaired hearing is poor functioning of the eustachian tubes with the accumulation of a clear, pale-yellowish fluid in the middle-ear cavity, known as serous or secretory otitis media. The vibratory movements of the tympanic (eardrum) membrane and ossicular chain in response to sound are impeded by the fluid, causing a moderate loss of hearing. Since this hearing loss is due to impairment of sound conduction to the inner ear, it is known as a conductive loss. In these children, evacuation of the fluid from the middle ear and its replacement with air restores normal sound conduction and normal hearing. The enlarged adenoid or allergic swelling of the membrane lining the eustachian tube needs to be corrected to prevent recurrences of secretory otitis media.

In early and middle adult life the usual cause for progressive impairment of hearing is otosclerosis, the disease process of bone surrounding the inner ear that has been described above. The conductive hearing loss caused by fixation of the stirrup bone (stapes) can be relieved by surgical mobilization of the stirrup or its replacement with a fine stainless-steel wire. In some cases the otosclerotic bone nodule grows inward to the inner ear, resulting in a gradual loss of hearing-nerve function. This may occur at the same time as the fixation of the stapes; it may precede stapes fixation; or it may occur some years after fixation of the stirrup bone, whether or not the stapes has been successfully operated upon. Hearing-nerve deterioration due to otosclerosis or any other cause cannot be improved surgically or by medical treatment. The progressive loss of hearing caused by the enlarging otosclerotic nodule may stop after a time, if the nodule becomes matured and inactive. Treatment of the patient with sodium fluoride tablets taken by mouth is a promising method for promoting inactivation of an active otosclerotic process but has not yet been proved conclusively to be of value.

The usual cause of impaired hearing after the age of 60 is presbycusis, the normal aging of the auditory system, for which there is no effective medical or surgical treatment.

Other varieties of ear disease causing impaired hearing have been mentioned above in the sections on diseases of the ear. In most cases when loss of hearing is of the conductive type surgical restoration of useful hearing by correcting the defect in the outer or middle ear is a possibility. When loss of hearing is of the sound-perceiving (nerve) type, surgical restoration cannot be expected. Medical treatment for nerve types of hearing loss is helpful only in rare cases when the loss is due to syphilis and in some early cases of Menière's disease.

More important than cure for nerve types of hearing loss is prevention. Preventable especially are cases of deafness in the newborn due to rubella in the mother. Excessive and prolonged noise-exposure nerve deafness is preventable by early detection (routine testing of the hearing of persons engaged in noisy occupations), by a change of occupation, or by the wearing of ear protectors, either specially designed earplugs or earmuffs.

The incidence of impaired hearing in the general population depends upon the degree of hearing loss defined as impaired. According to U.S. statistics, by age six, 0.2 percent of all children have impaired hearing in one or both ears sufficient to warrant consultation of an ear specialist (otologist). By age 18 the number of children with loss of hearing sufficient to require diagnostic examination reaches 2.5 to 3 percent. By age 65 the number of adults with a recognizable hearing impairment reaches 5 percent. Beyond 65 the incidence of impaired hearing rises rapidly as presbycusis, the normal aging of the auditory system takes its toll.

Comparable figures from Britain show that one in six persons is estimated to have some hearing difficulty, but only a quarter of these have any real handicap, with a third of this latter group needing hearing aids and one in 20 being deaf to all speech and beyond useful help with a hearing aid. With British children, one in 1,000 is severely deaf and as high as seven per thousand are estimated to have sufficient impairment to need some form of help.

Rehabilitation. The child born deaf or with a severe hearing impairment cannot acquire speech by the normal process. He must attend special classes or a school for the deaf to be taught speech and lipreading. Most of these children have remnants of the sense of hearing that can be utilized in their schooling by the use of aids to amplify sound. The child with a moderate or mild hearing impairment is able to acquire speech by himself but a little more slowly than the child with normal hearing, while speech-correction instruction is usually required to improve his diction.

Advances in hearing-aid technology have served to increase the proportion of hearing-impaired individuals who can benefit substantially from amplification. Selection of an appropriate hearing aid for individuals with sensorineural (or nerve-type) hearing loss may be difficult and time-consuming. Research has demonstrated repeatedly, however, that the ability of listeners with sensorineural hearing loss to understand speech at conversational levels often can be enhanced significantly by use of an appropriate hearing aid. For those individuals whose hearing loss causes severe distortion of speech, use of a hearing aid in combination with lipreading may increase the amount of speech the individual can understand through lipreading alone. One the other hand, selection of a hearing aid is often a simpler matter for listeners with hearing loss of the conductive type. Careful selection is necessary to ensure that maximum understanding of speech is obtainable in noisy environments. The hearing-impaired individual should consult with trained professionals such as audiologists, who are trained in evaluating the benefit derived from the use of a hearing aid.

Lipreading

Lipreading, which actually entails attentive observation of the entire facial expression rather than the movements of the lips alone, is utilized even by persons with normal hearing who, in the presence of background noise, need these visual clues to supplement hearing. As hearing begins to be impaired, lipreading, which might better be termed speechreading, becomes increasingly valuable and important.

The hearing-impaired individual who knows a spoken language can learn lipreading by careful observation of a speaker of that language. Formal instruction in lipreading by a teacher individually or in classes is necessary for those hearing-impaired persons who have not acquired knowledge of a spoken language. The greater the loss of hearing, the more essential becomes lipreading, for which good lighting is essential.

Speech-correction instruction, needed for the young with serious degrees of impaired hearing, also becomes necessary for the adult who suddenly loses all hearing in both ears. Without the monitoring effect of hearing his own voice, his speech begins to deteriorate and to acquire the flat, toneless quality of the profoundly deaf.

Social and economic handicaps. The social handicap of severe degrees of hearing impairment is particularly important to the individual and his family. An individual who has a hearing impairment can feel isolated and embarrassed by being unable to join group conversations. He therefore tends to withdraw within himself and often develops false ideas that others are talking about him and ridiculing him. No one has more poignantly written of this isolation than Beethoven, who is generally believed to have suffered from otosclerosis. Had he lived today, his hearing might have been restored or improved through surgery. As it was, it gradually became worse until, after conducting the first performance of his ninth and last symphony, he had to be told to turn around to face the audience to acknowledge the tremendous applause because he could not hear it.

Small devices, such as installing a buzzer instead of a high-pitched door and telephone bell for the nerve-deafened person who cannot hear tones of high pitch, using an amplifier on the telephone, and being patient when communicating with the hard-of-hearing person, make life easier both for him and for those around him.

The economic handicap of impaired hearing is in proportion to the degree of loss. Nevertheless, persons with a hearing disability, while unable to engage in jobs that require keen hearing, tend to have better records of dependability and fewer days away because of illness than persons with normal hearing. (G.E.S.)

BIBLIOGRAPHY

Sensory reception: E.D. ADRIAN, *The Basis of Sensation: The Action of the Sense Organs* (1928), a basic work in the field of sensory physiology; *Cold Spring Harbor Symposia on Quantitative Biology,* vol. 30, *Sensory Receptors* (1965), a report of current concepts and research in the field; JOHN FIELD (ed.), *Handbook of Physiology,* sect. 1, *Neurophysiology,* vol. 1 (1959), a comprehensive treatise on historical as well as current concepts of sensory reception. AINSLEY IGGO (ed.), *Somatosensory System* (1973), is another comprehensive handbook.

Mechanoreception: J. FIELD (gen. ed.), *Handbook of Physiology,* sec. 1, *Neurophysiology,* 3 vol. (1959–60), chapters on nonphotic receptors, posture and locomotion, vestibular mechanisms, initiation of impulses at receptors, touch and kinesis, and pain; J.D. CARTHY and G.E. NEWELL (eds.), *Invertebrate Receptors* (1968), chapters on mollusk statocysts, invertebrate proprioceptors, chordotonal organs, and mechanoreceptive transduction; P.H. CAHN (ed.), *Lateral Line Detectors* (1967), contributions of 35 investigators; M.J. COHEN and S. DIJKGRAAF, "Mechanoreception," in T.H. WATERMAN (ed.), *The Physiology of Crustacea,* vol. 2 (1961); S. DIJKGRAAF, "The Functioning and Significance of the Lateral-Line Organs," *Biol. Rev.,* 38:51–105 (1963); E. VON HOLST, "Die Arbeitsweise des Statolithenapparates bei Fischen," *Z. Vergl. Physiol.,* 32:60–120 (1950), a classical study on statoreception in the labyrinth; I.P. HOWARD and W.B. TEMPLETON, *Human Spatial Orientation* (1966), extensive coverage of the regulation of body posture; O. LOWENSTEIN, "Labyrinth and Equilibrium," pp. 60–82 in *Physiological Mechanisms in Animal Behaviour,* in *Symp. Soc. Exp. Biol.,* no. 4 (1950); D. MELLON, *The Physiology of Sense Organs* (1968); C.L. PROSSER and F.A. BROWN, *Comparative Animal Physiology,* 2nd ed. (1961), a textbook survey of mechanoreception and equilibrium; A.V.S. DE REUCK and J. KNIGHT (eds.), *Myotatic, Kinesthetic, and Vestibular Mechanisms* (1967), deals primarily with mammals; J SCHWARTZKOPFF, "Mechanoreception," in M. ROCKSTEIN (ed.), *The Physiology of Insecta,* vol. 1 (1964).

Thermoreception: J. BLIGH and H. HENSEL, "Modern Theories on Location and Function of the Thermoregulatory Centers in Mammals Including Man," in *Advances in Biometerology,* vol. 1 (1973), covers thermosensors in the central nervous system; H. HENSEL, "Physiologie der Thermoreception," *Ergebn. Physiol.,* 47:166–368 (1952), a comprehensive review with references, *Allgemeine Sinnesphysiologie: Hautsinne, Geschmack, Geruch* (1966), discusses skin receptors, with comprehensive references, and "Cutaneous Thermoreceptors," in *Handbook of Sensory Physiology,* vol. 2 (1973), describes thermoreceptors in the skin; R.W. MURRAY, "Temperature Receptors," *Advances Com. Physiol. Biochem.,* 1: 117–175 (1962), covers the comparative physiology of thermoreceptors; H. PRECHT, J. CHRISTOPHERSEN, and H. HENSEL, *Temperatur und Leben* (1955), a comprehensive review on temperature and life, covering microorganisms, plants, man, and other animals; Y. ZOTTERMAN, "Thermal Sensations," in J. FIELD (gen. ed.), *Handbook of Physiology,* sect. 1, *Neurophysiology,* 1:431–458 (1959), and "Specific Action Potentials in the Lingual Nerve of Cat," *Skand. Arch. Physiol.,* 75:105–120 (1936), a classic showing first electrical records from specific thermosensitive nerve fibres; M. BLIX, "Experimentela bidrag till lösning av frågan om hudnervernas specifika energi," *Uppsala LäkFör. Förh.,* 18:87–102 (1882–83), reports on the discovery of cutaneous hot and cold spots; E.A. BREARLEY and D.R. KENSHALO, "Electrophysiological Measurements of the Sensitivity of Cat's Upper Lip to Warm and Cool Stimuli," *J. Comp. Physiol. Psychol.,* 70:5–14 (1970); T.H. BULLOCK and F.P.J. DIECKE, "Properties of an Infrared Receptor," *J. Physiol.,* 134:47–87 (1956); C.B. DE WITT, "Precision of Thermoregulation and Its Relation to Environmental Factors in the Desert Iguana, Dipsosaurus Dorsalis," *Physiol. Zoöl.,* 40:49–66 (1967); H. HENSEL and K.K. BOMAN, "Afferent Impulses in Cutaneous Sensory Nerves in Human Subjects," *J. Neurophysiol.,* 23:564–578 (1960), contains first records of neural impulses from human thermoreceptors; H. HENSEL and D.R. KENSHALO, "Warm Receptors in the Nasal Region of Cats," *J. Physiol.,* 204:99–112 (1969); A. IGGO, "Cutaneous Thermoreceptors in Primates and Subprimates," *J. Physiol.,* 200:403–430 (1969); S. LANDGREN, "Convergence of Tactile, Thermal, and Gustatory Impulses on Single Cortical Cells," *Acta Physiol. Scand.,* 40:210–221 (1957); R. LOFTUS, "The Response of the Antennal Cold Receptor of Periplaneta Americana to Rapid Temperature Changes and to Steady Temperature," *Z. Vergl. Physiol.,* 59:413–455 (1968); T. NAKAYAMA *et al.,* "Thermal Stimulation of Electrical Activity of Single Units of the Preoptic Region," *Am. J. Physiol.,* 204:1122–1126 (1963), first records of impulses from thermosensors in the cat hypothalamus; D.A. POULOS and R.M. BENJAMIN, "Response of Thalamic Neurons to Thermal Stimulation of the Tongue," *J. Neurophysiol.,* 31: 28–43 (1968). See also GIORGIO MACCHI, ALDO RUSTIONI, and ROBERTO SPREAFICO (eds.), *Somatosensory Integration in the Thalamus* (1983).

Chemoreception: J.E. AMOORE, *Molecular Basis of Odor* (1970), a technical discussion of molecular shapes and odours; M. BEROZA (ed.), *Chemicals Controlling Insect Behavior* (1970), technical reports at a symposium on pheromones and defensive secretions of insects; T.H. BULLOCK and G.A. HORRIDGE, *Structure and Function in the Nervous Systems of Invertebrates,* 2 vol. (1965), a monumental review of invertebrate sensory physiology and neurophysiology, with extensive bibliographies; V.G. DETHIER, *The Physiology of Insect Senses* (1963), a technical review, with sections on chemoreception; H. and M. FRINGS, *Animal Communication* (1964), a semipopular survey, including sections on chemical signalling in the animal kingdom; R. HARPER, E.C. BATE SMITH, and D.G. LAND, *Odour Description and Odour Classification* (1968), a technical review of odour theory and practical schemes of classification; T. HAYASHI (ed.), *Olfaction and Taste II* (1967), technical reports at a symposium on vertebrate chemoreception, especially electrophysiological and electron microscope studies, and discussion of theories; J.W. JOHNSTON, D.G. MOULTON, and A. TURK (eds.), "Communication by Chemical Signals," *Advances in Chemoreception,* vol. 1 (1970), a technical discussion of the field; M.R. KARE and O. MALLER (eds.), *The Chemical Senses and Nutrition* (1967), technical reports at a symposium, mostly on human chemoreception, with an extensive bibliography on taste for the years 1566–1966; W.W. KILGORE and R.L. DOUTT (eds.), *Pest Control: Biological, Physical, and Selected Chemical Methods* (1967), technical reviews by specialists, including chapters on pheromones, repellents, and antifeedants; H. KLEEREKOPER, *Olfaction in Fishes* (1969), a semipopular review especially on orientation by odours; L. and M. MILNE, *The Senses of Animals and Men* (1962), a popular survey of senses and behaviour; R.W. MONCRIEFF, *The Chemical Senses,* 3rd ed. (1967), a standard technical reference on chemoreception in vertebrates, particularly man; G.H. PARKER, *Smell, Taste, and Allied Senses in the Vertebrates* (1922), a classic summary of earlier research and theories; H.W. SCHULTZ, E.A. DAY, and L.M. LIBBEY (eds.), *Symposium on Foods: The Chemistry and Physiology of Flavors* (1967), technical reports at a symposium, particularly on chemical analysis for odorants in foods; see the *Scientific American* for excellent semipopular articles on many aspects of chemoreception (February 1964, August 1964, June 1967, May 1968, and February 1969); T.A. SEBEOK (ed.), *Animal Communication* (1968), technical reviews by specialists, with chapters on chemical signalling; E. SONDHEIMER and J.B. SIMEONE (eds.), *Chemical Ecology* (1970), technical reviews by specialists on effects of environmental chemicals on animals, including chapters on plant feeding stimulants, communication signals, defense chemicals, and fish orientation; T.H. WATERMAN (ed.), *The Physiology of Crustacea,* vol. 2 (1961), technical reviews by specialists, including chapters on senses and behaviour; V.B. WIGGLESWORTH, *The Principles of Insect Physiology,* 6th ed. (1965), a standard textbook in the field, including a chapter on chemoreception; K.M. WILBUR and C.M. YONGE (eds.), *The Physiology of Mollusca,* vol. 2 (1966), technical reviews by specialists, including chapters on chemoreception and behaviour; G.E.W. WOLSTENHOLME and J. KNIGHT (eds.), *Taste and Smell in Vertebrates* (1970), technical reports at a symposium, particularly on morphology of receptors, electrophysiology, and theories; D.L. WOOD, R.M. SILVERSTEIN, and M. NAKAJIMA (eds.), *Control of Insect Behavior by Natural Products* (1970), technical reports at a symposium particularly concerned with methods of research on feeding stimulants, deterrents, and pheromones; R.H. WRIGHT, *The Science of Smell* (1964), a semitechnical discussion of odour theories, particularly the molecular vibration theory; Y. ZOTTERMAN (ed.), *Olfaction and Taste* (1963), technical reports at a symposium, particularly on morphology, electrophysiology, and theories. Later works include H. ACKER and R.G. O'REGAN (eds.), *Physiology of the Peripheral Arterial Chemoreceptors* (1983); DIETLAND MÜLLER-SCHWARZE and ROBERT M. SILVERSTEIN (eds.), *Chemical Signals in Vertebrates: Proceedings of the Third International Symposium* (1983); D. MICHAEL STODDART (ed.), *Olfaction in Mammals: Proceedings of a Symposium of the Zoological Society of London* (1980); A.D. HASLER, A.T. SCHOLZ, and R.W. GOY, *Olfactory Imprinting and Homing in Salmon* (1983); KLAUS REUTTER, *Taste Organ in the Bullhead (Teleostei)* (1978); R.H. WRIGHT, *The Sense of Smell* (1982).

Photoreception: M.H. PIRENNE, *Vision and the Eye,* 2nd ed. (1967), optics and physiology of vision (vertebrate and invertebrate eyes) for the beginner and non-specialist; *Handbook of Sensory Physiology:* vol. 7, pt. 1, H.J.A. DARTNALL (ed.), *The Photochemistry of Vision;* vol. 7, pt. 2, M.G.F. FOURTES (ed.), *Physiology of Photoreceptor Organs* (1971–72), authoritative treatises by leading scientists; C.G. BERNHARD (ed.), *The Functional Organization of the Compound Eye* (1966), on the optics, morphology, photochemistry, and physiology of the compound eye and other primitive eyes; GORDON L. WALLS, *The Vertebrate Eye and Its Adaptive Radiation* (1942, reprinted 1963), a classic and encyclopaedic store of knowledge of the vertebrate retina and eye; TSUNEO TOMITA, "Electrical Activity of Vertebrate Photoreceptors," *Q. Rev. Biophys.,* 3:179–222 (1970), a summary of vertebrate photoreceptor physiology; T.H. GOLDSMITH and G.D. BERNARD, "The Visual System of Insects," in MORRIS ROCKSTEIN (ed.), *The Physiology of Insecta* (1972), on the visual system of insects, optics, visual pigments, and physiology; M.F. MOODY, "Photoreceptor Organelles in Animals," *Biol. Rev.,* 39:43–86 (1964); RICHARD M. EAKIN, "Lines of Evolution of Photoreceptors," in DANIEL MAZIA and ALBERT TYLER (eds.), *General Physiology of Cell Specialization* (1963), and "Structure of Invertebrate Photoreceptors," in H.J.A. DARTNALL (ed.), *Photochemistry of Vision* (1972); three articles on photoreceptor structure and function; BRADLEY R. STRAATSMA *et al.* (eds.), *The Retina* (1969), on the morpholgy, function, and clinical characteristics of the vertebrate retina; HUGH DAVSON (ed.), *The Eye,* vol. 2, *The Visual Process* (1962), a textbook of photoreception and function of the visual system; the *Journal of the Optical Society of America,* vol. 53, no. 1 (1963); the Cold Spring Harbor Symposia on Quantitative Biology, vol. 30, *Sensory Receptors* (1965); and the Proceedings of the International School of Physics, "Enrico Fermi," course 43, ed. by W. REICHARDT, *Processing of Optical Data by Organisms and by Machines* (1969), all contain recent research on photoreception and retinal function; H.J.A. DARTNALL, *The Visual Pigments* (1957); and GEORGE WALD, "Molecular Basis of Visual Excitation," *Science,* 162:230–239 (1968), reviews of visual pigment chemistry; T.H. BULLOCK and G.A. HORRIDGE, *Structure and Function in the Nervous Systems of Invertebrates,* vol. 2 (1966), contains a broad survey of invertebrate sensory receptors; H.K. HARTLINE, "Visual Receptors and Retinal Interaction," *Science,* 164: 270–278 (1969); FLOYD RATLIFF, *Mach Bands: Quantitative Studies on Neural Networks in the Retina* (1965), physiological effects on vision of neural activity in the retina. Physiology of vision is also explored in JONATHAN STONE, *Parallel Processing in the Visual System* (1983); LEO M. HURVICH, *Color Vision* (1981); EBERHART ZRENNER, *Neurophysiological Aspects of Color Vision in Primates* (1983); GERALD H. JACOBS, *Comparative Color Vision* (1981).

Sound reception: There are no general texts on sound reception. Included below are some of the specialized references dealing with this subject in a technical manner.

J. SCHWARTZKOPFF, "Mechanoreception," in MORRIS ROCKSTEIN (ed.), *The Physiology of Insecta,* vol. 1, pp. 509–561 (1964); M.L. WOLBARSHT, "Electrical Characteristics of Insect Mechanoreceptors," *J. Gen. Physiol.,* 44:105–122 (1960); E.G. GRAY, "The Fine Structure of the Insect Ear," *Phil. Trans. R. Soc.,* Series B, 243:75–94 (1960); C. WALCOTT and W.G. VAN DER KLOOT, "The Physiology of the Spider Vibration Receptor," *J. Exp. Zool.,* 141:191–244 (1959); E.G. WEVER and J.A. VERNON, "The Auditory Sensitivity of Orthoptera," *Proc. Natn. Acad. Sci. U.S.A.,* 45:413–419 (1959); H. and M. FRINGS, "Uses of Sounds by Insects," *A. Rev. Ent.,* 3:87–106 (1958); R.J. PUMPHREY, "Hearing in Insects," *Biol. Rev.,* 15:107–132 (1940); "Sensory Organs: Hearing," in A.J. MARSHALL (ed.), *Biology and Comparative Physiology of Birds,* vol. 2. (1961). See also BRIAN LEWIS, *Bioacoustics: A Comparative Approach* (1983); WILLIAM C. STEBBINS, *The Acoustic Sense of Animals* (1983); JAMES F. WILLOTT (ed.), *The Auditory Psychobiology of the Mouse* (1983).

Human sensory reception: E.G. BORING, *Sensation and Perception in the History of Experimental Psychology* (1942), a classic historical account of the early work in sensation and perception; J. FIELD (ed.), *Handbook of Physiology, Section 1, Neurophysiology,* vol. 1 (1959), a technical and detailed review of modern sensory physiology; F.A. GELDARD, *The Human Senses* (1953), a scholarly overview of the senses, suitable as an introduction to the subject; D.R. KENSHALO (ed.), *The Skin Senses* (1968), a rather specialized report of a symposium held in 1968 that gives the reader with a special interest in this field a good idea of current research; P.M. MILNER, *Physiological Psychology* (1970), an advanced textbook in general physiological psychology with a good section on the senses; C. PFAFFMANN (ed.), *Olfaction and Taste, III* (1969), a somewhat specialized but good overview of current research in both olfaction and taste; J.S. WILENTZ, *The Senses of Man* (1968), an excellent popular account for the general reader that serves as a good introduction to the field. Later monographs on human sensory physiology include discussions of modern theories of pain, and on space perception, in addition to tradional studies of smell, taste, vision, and hearing. See CHRISTIAAN BARNARD and JOHN ILLMAN (eds.) *The Body Machine* (1981); HERBERT HENSEL, *Thermal Sensations and Thermoreceptors in Man* (1982); TRYGG ENGEN, *The Perception of Odors* (1982); RONALD MELZACK and PATRICK D. WALL, *The Challenge of Pain,* rev. ed. (1983); LAWRENCE KRUGER and JOHN C. LIEBESKIND (eds.), *Neural Mechanisms of Pain,* (1984); HERBERT L. PICK, Jr., and LINDA P. ACREDOLO (eds.), *Spatial Orientation: Theory, Research, and Application: Proceedings of a Conference on Spatial Orientation and Perception, July 14–*

16, 1980 (1983); R. ROBIN BAKER, *Human Navigation and the Sixth Sense* (1982); MICHAEL POTEGAL (ed.), *Spatial Abilities: Development and Physiological Foundations* (1982).

Structure and function of the human eye: H. DAVSON (ed.), *The Eye,* 4 vol. (1962; 2nd ed., vol. 1, 1969), covers the whole field of eye physiology, written by a group of experts in their particular fields; W.S. DUKE-ELDER *et al.* (eds.), *System of Ophthalmology,* 15 vol. (1958–), authoritative accounts of the anatomy and physiology of the eye; E. WOLFF, *Anatomy of the Eye and Orbit,* 5th ed. (1961), the classic work on this aspect; H. DAVSON, *Physiology of the Eye,* 3rd ed. (1971), an up-to-date account of eye physiology covering all aspects; M.H. PIRENNE, *Vision and the Eye,* 2nd ed. (1967), a simple account of certain features of eye physiology; R.A. WEALE, *The Eye and Its Function* (1960), a short and elementary account; H. VON HELMHOLTZ, *Handbuch der physiologischen Optik,* 3rd ed. (1886–96; Eng. trans., *Physiological Optics,* 3 vol., 1924–25; reprinted in 2 vol., 1962), a classic account of the psychological aspects of vision—not at all out of date, although written nearly 100 years ago; H.H. EMSLEY, *Visual Optics,* 5th ed., 2 vol. (1952–53), a technical account of the detailed optics of the eye. Physiological aspects of vision are discussed in HITOSHI SHICHI, *Biochemistry of Vision* (1983). Psychology of vision, including motion, depth, binocular vision, visual effects, color, etc., is discussed in MARK FINEMAN, *The Inquisitive Eye* (1981).

Eye diseases and visual disorders: SIR STEWART DUKE-ELDER (ed.), *Parsons' Diseases of the Eye,* 15th ed. (1970), a textbook for students concentrating on the more common eye conditions, and (ed.), *System of Ophthalmology,* vol. 1–5, 7–12 (1958–71; vol. 6, 13–15 in prep.) perhaps the most comprehensive textbook on any medical subject; F.W. NEWELL, *Ophthalmology,* 2nd ed. (1969), an up-to-date standard textbook; E.S. PERKINS and P. HANSELL, *Atlas of Diseases of the Eye,* 2nd ed. (1971), illustrations of common eye conditions with brief text; D.T. VAIL, *The Truth About Your Eyes,* 2nd ed. (1959), a description for the layman of the function of the eye and management of the more common eye diseases; F.B. WALSH and W.F. HOYT, *Clinical Neuro-Ophthalmology,* 3rd ed., 3 vol. (1969), a very detailed account of ophthalmic conditions associated with neurological diseases. A wide range of disorders and visual defects is covered in ROBERT SEKULER, DONALD KLINE, and KEY DISMUKES (eds.), *Aging and Human Visual Function* (1982); and JOHN H. DOBREE and ERIC BOULTER, *Blindness and Visual Handicap, the Facts* (1982).

Structure and function of the human ear: Reliable and readable introductory treatises are S.S. STEVENS and FRED WARSHOFSKY, *Sound and Hearing* (1965); and W.A. VAN BERGEIJK, J.R. PIERCE, and E.E. DAVID, *Waves and the Ear* (1960). HALLOWELL DAVIS and S.R. SILVERMAN, *Hearing and Deafness,* 3rd ed. (1970), although somewhat more technical, is also written for the non-specialist. A useful account of the anatomy of the ear may be found in WILLIAM BLOOM and D.W. FAWCETT, *A Textbook of Histology,* 9th ed. (1968). B.J. ANSON and J.A. DONALDSON, *The Surgical Anatomy of the Temporal Bone and Ear* (1967), presents the gross and microscopic structure of the ear from the point of view of the surgeon. Details of inner ear anatomy can be studied in SALVATORE IURATO, *Submicroscopic Structure of the Inner Ear* (1967); HANS ENGSTROM, H.W. ADES, and ANTON ANDERSSON, *Structural Pattern of the Organ of Corti* (1966); and HEINRICH SPOENDLIN, *The Organization of the Cochlear Receptor* (1966). For the comparative anatomy of the ear from fish to man, the illustrations in GUSTAF RETZIUS' two folio volumes *Das Gehörorgan der Wirbelthiere* (1881–84), are still unsurpassed, but available only in a few medical libraries. The physiology of hearing and of vestibular function is discussed in four chapters by JOSEPH E. HAWKINS in C.H. BEST and N.B. TAYLOR (eds.), *The Physiological Basis of Medical Practice,* 8th ed. (1966). Modern classics in the field of hearing are: S.S. STEVENS and HALLOWELL DAVIS, *Hearing* (1938, reprinted 1960), equally divided between psychological and physiological aspects of the subject; E.G. WEVER, *Theory of Hearing* (1949, reprinted 1970), including a good historical treatment of auditory theories developed through the centuries; and E.G. WEVER and MERLE LAWRENCE, *Physiological Acoustics* (1954), concerned mainly with middle-ear and inner-ear mechanics. G. VON BEKESY, *Experiments in Hearing* (1960), is the best source of information about the experimental work that won von Békésy the Nobel Prize, but it is not recommended for the novice. Biochemical aspects of cochlear physiology and pathology are treated in SIGURD RAUCH, *Biochemie des Hörorgans* (1964); and in *Biochemical Mechanisms in Hearing and Deafness,* ed. by MICHAEL M. PAPARELLA (1970). I.C. WHITFIELD, *The Auditory Pathway* (1967), gives a useful summary of our imperfect knowledge of auditory processes in the central nervous system. Later monographs on acoustic perception include ÅAGE R. MØLLER, *Auditory Physiology* (1983); SANFORD E. GERBER and GEORGE T. MENCHER, (eds.), *The Development of Auditory Behavior* (1983); JOHN M. PALMER, *Anatomy for Speech and Hearing,* 3rd ed. (1984); M.E. LUTMAN, and M.P. HAGGARD (eds.), *Hearing Science and Hearing Disorders* (1983).

The measurement of human hearing is described in numerous textbooks of audiology. *Audiometry: Principles and Practices,* ed. by ARAM GLORIG (1965), gives a good overview of the field. A thorough, sober, and factual account of studies concerning *The Effects of Noise on Man* has been written by K.D. KRYTER (1970). There are few comprehensive publications devoted to vestibular function, but the symposium volume, ed. by R.J. WOLFSON, *The Vestibular System and Its Diseases* (1966), is a good source of information, as is the NASA series ed. by ASHTON GRAYBIEL, *The Role of the Vestibular Organs in Space Exploration,* 4 vol. (1965–68).

Ear diseases and hearing disorders: G.E. SHAMBAUGH, "A Restudy of the Minute Anatomy of Structures in the Cochlea with Conclusions Bearing on the Solution of the Problem of Tone Perception," *Amer. J. Anat.,* 7:245–257 (1907), the first detailed description of the hearing nerve end-organ in the cochlea, where sound waves are converted into nerve impulses depending upon the pitch of the tone; GEORG VON BEKESY, "The Ear," *Scient. Am.,* 197:66–78 (1957), a description in lay terms of the mechanism of hearing by today's foremost research authority on the ear; G.E. SHAMBAUGH, JR., *Surgery of the Ear,* 2nd ed. (1967), a well-illustrated text on diseases of the ear and their surgical correction; and with A. PETROVIC, "Effects of Sodium Fluoride on Bone," *J.A.M.A.,* 204:969–973 (1968), a summary of recent research on the arrest of progressive deafness due to otosclerosis by means of sodium fluoride; PHILIP H. BEALES, *Noise, Hearing and Deafness* (1965), a useful review, written in lay language, of the problem of deafness and the adverse influence on hearing of excess noise exposure. Hearing disorders are also discussed in ERNEST J. MOORE (ed.), *Bases of Auditory Brain-stem Evoked Responses* (1983); JAMES JERGER (ed.), *Hearing Disorders in Adults* (1984); KATHRYN P. MEADOW, *Deafness and Child Development* (1980); SHIRLEY HANAWALT MCARTHUR, *Raising Your Hearing-Impaired Child* (1982).

Seoul

Seoul (Sŏul-t'ŭkpyŏlsi [Special City of Seoul]) was, except for a brief interregnum (1399–1405), the capital of Korea from 1394 until the formal division of the country in 1948, when it became the capital of the Republic of Korea (South Korea). The name itself has come to mean "capital" in the Korean language. The city was popularly called Seoul in Korean during both the Yi dynasty (1392–1910) and the period of Japanese rule (1910–45), although the official names in those periods were Hansŏng and Kyŏngsŏng, respectively. The city was also popularly and, during most of the 14th century, officially known as Hanyang. Seoul became the official name of the city only with the founding of the Republic of Korea.

This article is divided into the following sections:

PHYSICAL AND HUMAN GEOGRAPHY

The landscape. *The city site.* Modern Seoul was founded in 1394 by Gen. Yi Sŏng-gye, the founder of Korea's Yi dynasty, as the capital of a unified nation. The site was a militarily defensible natural redoubt that was also an especially suitable site for a capital city, lying at the centre of an undivided Korea and adjoining the navigable Han-gang (Han River), one of the peninsula's major rivers flowing into the Yellow Sea. The contact afforded by this riverain site both with inland waterways and with coastal sea routes was particularly important to Yi because these were the routes by which grain, taxes, and goods were transported. In addition to the practical advantages, the site was well situated according to *p'ungsuchirisol,* the traditional belief in geomancy. The district chosen by Yi remains, after 600 years, the centre of Seoul; it is located immediately north of the Han-gang in the lowland of a topographic basin surrounded by low hills of about 1,000 feet (300 metres). The natural defensive advantages of the basin were reinforced two years after the city's founding by the construction of an 11-mile (18-kilometre) wall along the ridges of the surrounding hills.

Today, the remains of the fortifications are a popular attraction. The old city centre is drained by a small tributary of the Han-gang, which has been covered over by streets and expressways. Main streets and major shopping areas occupy the lower part of the basin. The original city district served to contain most of the city's growth until the early 20th century; for, although the population had grown to approximately 100,000 by the census of 1429, it had risen to only about 250,000 by the time of the Japanese annexation in 1910, almost five centuries later. The modernizations brought by the Japanese began the first of several cycles of growth during the 20th century that extended the city limits by successive stages, so that they now contain both banks of the Han-gang river plain, as well as the banks of several tributary rivers. The city's boundaries now form a ragged oval about eight to 12 miles distant from the original site, except to the northwest, where they are approximately half that distance. The present boundary of Seoul is largely that established in 1963 and encompasses roughly 234 square miles (605 square kilometres), more than twice the city area of 1948. Seoul has grown rapidly since the Korean War (1950–53). Suburbs have sprung up in the rural areas surrounding the city, and such satellite cities as Sŏngnam, Suwŏn, and Inch'ŏn have undergone considerable expansion as the capital has grown.

Climate. Seoul's climate is characterized by a large annual range of temperature. The coldest month, January, has a mean temperature of 26° F (−3° C), and the warmest month, August, has a mean temperature of 78° F (25° C). Yearly precipitation in the city is about 54 inches (1,370 millimetres), with a heavy concentration during the summer months. Air pollution in the basin and in Yŏngdŭng-p'o, an industrial area, has become a serious problem, caused in large part by the increasing number of automobiles and factories. For years the Han-gang was highly polluted, but since the early 1980s pollution levels have been reduced significantly by measures to control the river's water level and by the construction of large-scale sewage treatment facilities.

The city plan. Street patterns in the city centre are basically rectangular. Streets and buildings stretch out in all directions from the old city wall's four major gates that still stand: toward Mia-dong and Suyu-dong to the north, Ch'ŏngnyang-ni to the east, Yongsan and Yŏngdŭng-p'o to the south, and Map'o and Hongje-dong to the west. Main streets, such as Ŭlchi-ro and Chong-no, are oriented east to west, but, toward the foot of the surrounding hills, topographic irregularities have some influence on the pattern. Outside the basin area of the central city, however, there are a number of radiating streets, which are interconnected by a series of circular roads. The Capitol Building and other government offices are concentrated along Sejong-no, although the National Assembly building is on Yŏido island; banks, department stores, and other business offices are located along Namdaemun-no and T'aep'yŏng-no. The area of Chong-no, Myŏng-dong, and Ŭlchi-ro constitutes the central business district. The district has been transformed from an area of wooden, tile-roofed houses to one of concrete high-rise office buildings. Much of the city's expansion has been to the south of the Han-gang, resulting in the creation of three new urban centres at Yŏido-Yŏngdŭngp'o, Yŏngdong, and Chamshil.

Housing. A shortage of housing has been a chronic problem. A number of large-scale apartment blocks were built, especially along the banks of the Han-gang. In addition, much residential housing has been developed along the suburban fringes of the city. Old-style houses—with the traditional heated floors (*ondol*) designed for the cold winters—are still found in areas of the old city and adjacent to the remains of the city wall.

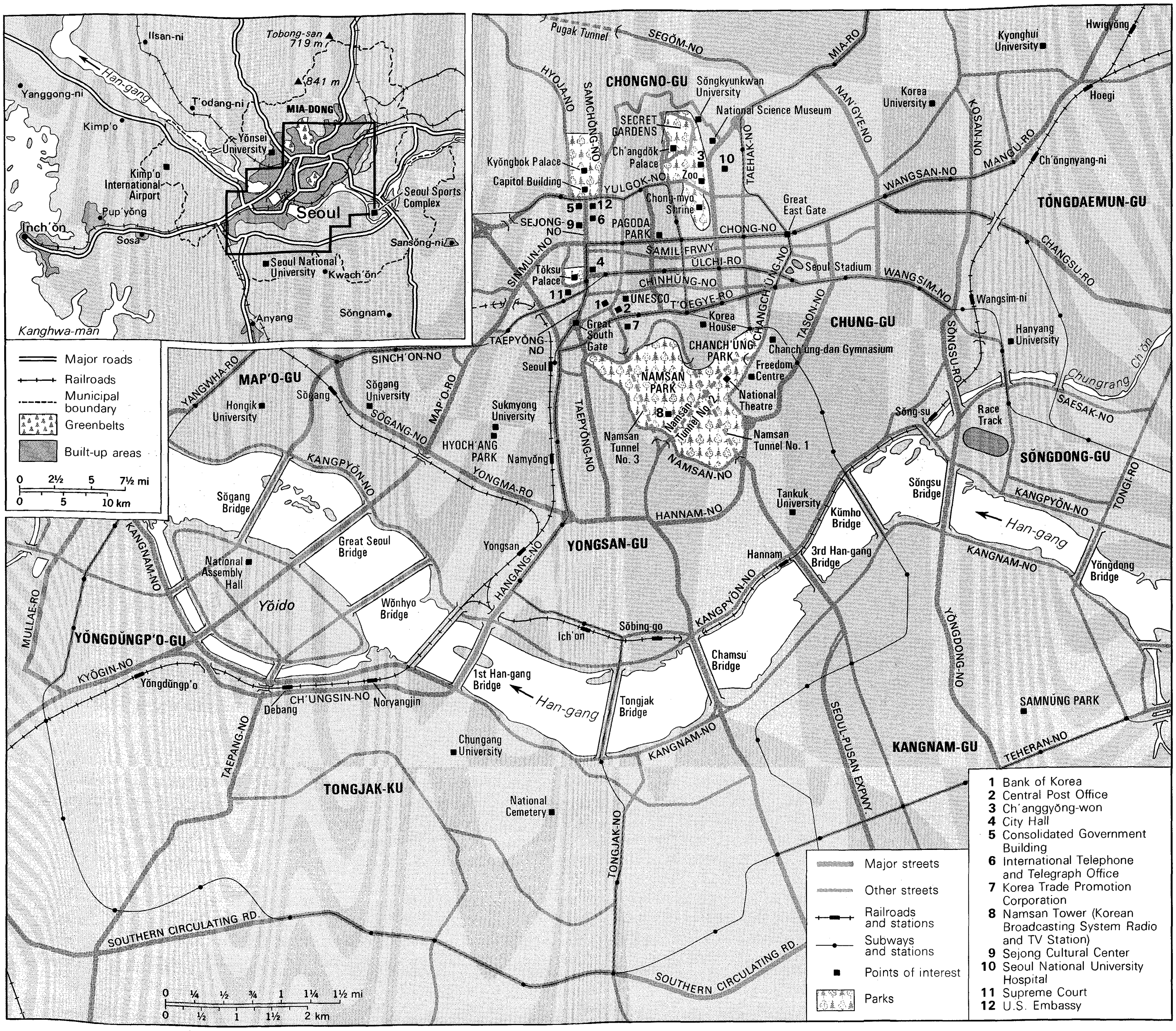
The people. The population of Seoul has grown extremely rapidly since 1950, and the city now has one of the highest population densities in the world. The most densely populated areas are distributed within and outside the old city and in the apartment belts along the Han-gang, while residential areas in the suburbs have a relatively low density. Koreans constitute nearly all of the population, the number of foreign residents being insignificant.

Principal employers

The economy. *Industry and commerce.* Manufacturing, commerce, and service industries are the principal employers. The textile, machinery, and chemical industries are the chief sectors in manufacturing, but food and beverage processing, printing, and publishing are also important.

The two most important traditional shopping areas are the extensive Great East Gate Market (Tongdaemun Sijang) and the smaller Great South Gate Market (Namdaemun Sijang), located near their respective gates. Comprising numerous individually owned shops, these markets serve not only Seoul but the entire country. There are also several large downtown department stores and modern shopping centres in the apartment blocks.

Seoul is the centre of finance for the country. The headquarters of the major stock exchanges and banks are



Central Seoul and (inset) its metropolitan area.

located there, and the city plays host to many annual trade shows.

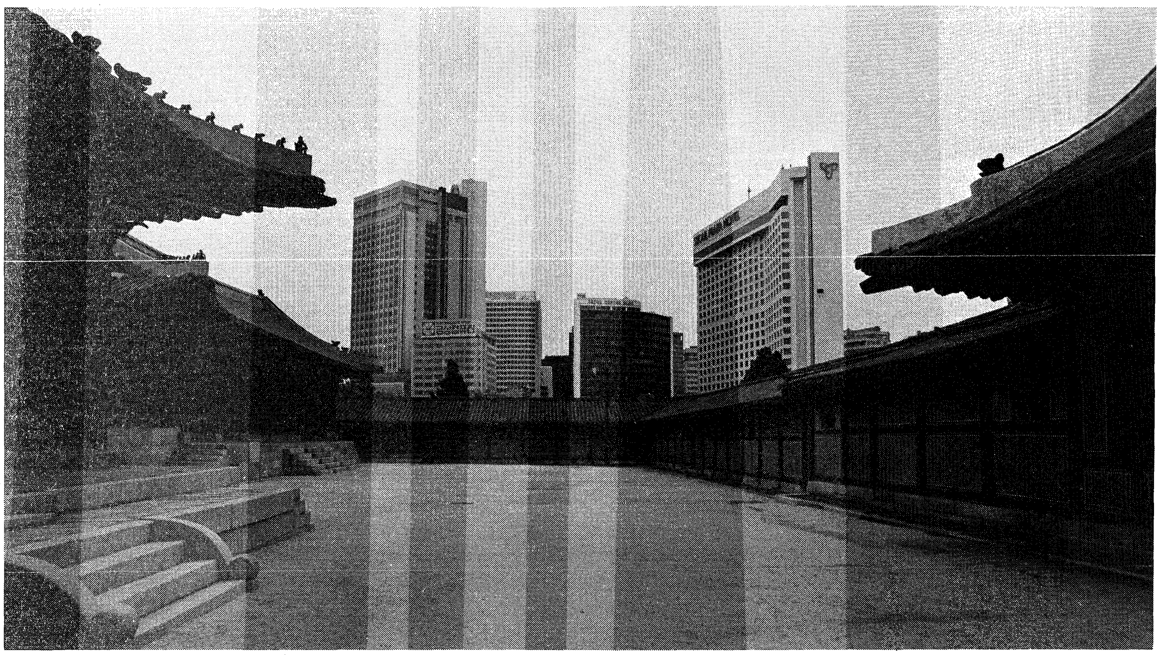
Transportation. Although Seoul is an old city, it has a good road system; vast improvements have been made in the system since the Korean War, notably in the construction of more than a dozen bridges across the Han-gang. Transportation facilities, however, have not been able to keep up with the demands of a large and expanding population, resulting in crowded streets. An extensive subway system has replaced the older streetcars; this has alleviated traffic congestion and has become, with buses and railways, one of the main forms of public transport. The capital is the hub of railway lines connecting it with most provincial cities and ports, including Inch'ŏn and Pusan. Before the Korean War, small vessels navigated up the river 37 miles to Seoul, but the demilitarized zone that now divides Korea into North and South runs partly through the mouth of the river and has deprived Seoul of its role as a river port. Hence, most goods are transported to and from the city on railways and highways. Kimp'o International Airport, located in the western part of Seoul, serves as the centre of the nation's airline network.

Administration and social conditions. Water and sewage-disposal facilities are inadequate, notwithstanding strenuous efforts to keep up with the demand stemming from rapid expansion of the city. Medical facilities are relatively good, and there are many small clinics as well as numerous doctors of herbal medicine. Fires occur frequently during winter and spring (the cold, dry season), placing a heavy demand on the city's fire services.

Compulsory education applies only to the six-year elementary school, but a large proportion of elementary school graduates receive a secondary education. There is a shortage of elementary school facilities because of the rapid increase in population. Most of South Korea's major universities, colleges, and research institutes are located in Seoul.

Cultural life. Seoul is the country's cultural centre. It is the home of the National Academy of Arts and the National Academy of Sciences and nearly all of the nation's learned societies and libraries. The National Classical Music Institute, engaged in the preservation of the traditional court music of Korea and in the training of musicians, is complemented by two Western-style symphony orchestras. In addition, there are a national theatre, an opera, and a number of public and private museums, including the National Museum on the grounds of the Kyŏngbok Palace. The Sejong Cultural Center, to the south of the palace, has facilities for concerts, plays, and exhibitions.

Surrounded by hills, Seoul has numerous small and large



Downtown Seoul seen from the Tŏksu Palace.
Woon Gu Kang

parks within easy reach. Places of historical interest—including Ch'anggyŏng, Ch'angdŏk, Kyŏngbok, and Tŏksu palaces and Chong-myo Shrine—annually attract large numbers of citizens and tourists. The city also has excellent sports and recreational facilities, notably the Seoul Sports Complex, which was built for the Asian Games in 1986 and the Olympic Games in 1988.

HISTORY

The earliest historical mention of Seoul and the surrounding area dates from the 1st century BC. During the Three Kingdoms period (57 BC–AD 668) of Silla, Koguryŏ, and Paekche, the area formed a borderland between the three countries, although during the early part of the period it was most closely associated with the Kingdom of Paekche. Historical accounts as well as archaeological records indicate that the original site of Paekche's capital, Wiryesŏng, was in the northeastern part of present-day Seoul. Shortly thereafter the capital was moved south across the Han-gang; a number of remains, including earthen walls, dwellings, and tombs have been uncovered at that site. It was not, however, until King Munjong of Koryŏ built a summer palace in 1068 that a fairly large settlement existed on the site of the modern city.

After the formal establishment of Seoul as the capital of the unified Yi state in 1394, construction and growth were very rapid. Construction on the Kyŏngbok Palace began in 1392; it was the residence of the Yi kings from 1395 until 1592. Before residence had even been established, the construction of the city's defensive walls had been completed, although so hastily that they had to be reconstructed in 1422. The Tŏksu Palace, the construction of which began in the late 15th century, was the residence of the Yi kings from 1593 until 1611. The Ch'angdŏk Palace, begun in 1405, was the residence from 1611 to 1872, when the king moved back into the reconstructed Kyŏngbok Palace (it had been burned by the Japanese in 1592 and was not rebuilt until 1867). Throughout this period Seoul remained the centre of the "Hermit Kingdom," with little contact permitted with the outside world. The opening of Korea to diplomatic contacts with the West in 1876, at a time when the weakening Yi dynasty was unable to control Western influence, led in 1905 to the establishment of a Japanese protectorate over the kingdom.

Seoul under Japanese rule

A year after the annexation of Korea to Japan in 1910, the name of the Seoul area was changed to Kyŏngsŏng, and minor changes were made in the boundary. Seoul served as the centre of Japanese rule, and modern technology was imported. Roads were paved, old gates and walls partly removed, new Western-style buildings built, and streetcars introduced.

After the end of Japanese control in 1945, Seoul came under the direct control of the central government, and in 1962 it was placed directly under the control of the prime minister. The city was left devastated by the Korean War. Out of the rubble has risen a modern city of skyscrapers and highways that has become one of the largest metropolises in the world.

BIBLIOGRAPHY. Upon its completion, the Korean-language *Sŏul yukpaengnyŏnsa* ("Six Hundred Year History of Seoul"; 1977–), issued by the SŎUL T'ŬKPYŎLSI SA P'YŎNCH'AN WIWŎNHOE (History Compilation Committee of the Special City of Seoul), will be the most authoritative source; 5 vol. had appeared to 1985. Also published by SŎUL T'ŬKPYŎLSI SA P'YŎNCH'AN WIWŎNHOE is *The Hyangto Seoul: The Special City of Seoul Historical Review* (annual). In Western languages, JONG-KI KIM, *Seoul: Host City of the '88 Olympics* (1983), is a comprehensive guidebook with numerous colour pictures and maps. ALLEN D. CLARK and DONALD N. CLARK, *Seoul: Past and Present* (1969), is a well-written and comprehensive guidebook concentrating on antiquities. EDWARD B. ADAMS, *Through Gates of Seoul*, 2nd ed., 2 vol. (1974–77), is a more extensive work, well prepared and containing numerous pictures and maps, some in colour.

(C.Le.)

Set Theory

Between the years 1874 and 1897, the German mathematician and logician Georg Cantor created a theory of abstract sets of entities and made it into a mathematical discipline. This theory grew out of his investigations of certain concrete problems regarding certain types of infinite sets of real numbers (see ANALYSIS: *Real analysis*). A set, wrote Cantor, is a collection of definite, distinguishable objects of perception or thought conceived as a whole. The objects are called elements or members of the set.

The theory had the revolutionary aspect of treating infinite sets as mathematical objects that are on an equal footing with those that can be constructed in a finite number of steps. Since antiquity, a majority of mathematicians had carefully avoided the introduction into their arguments of the actual infinite (*i.e.,* of sets containing an infinity of objects conceived as existing simultaneously, at least in thought). Since this attitude persisted until almost the end of the 19th century, Cantor's work was the subject of much criticism to the effect that it dealt with fictions; indeed, that it encroached on the domain of philosophers and violated the principles of religion. Once applications to analysis began to be found, however, attitudes began to change, and in the 1890s Cantor's ideas and results were gaining acceptance. By 1900, set theory was recognized as a distinct branch of mathematics.

At just that time, however, it received a severe setback through the derivation of several contradictions in its superstructure (see below *Cardinality and transfinite numbers*). The main thrust of this article is to present an account of one response to such contradictions as these. The purpose of the development here related has been to provide an axiomatic basis for the theory of sets analogous to that developed for elementary geometry. The degree of success that has been achieved in this development, as well as the present stature of set theory, has been well expressed in the Bourbaki *Éléments de mathématique:* "Nowadays it is known to be possible, logically speaking, to derive practically the whole of known mathematics from a single source, The Theory of Sets." (R.R.S./Ed.)

See the article MATHEMATICS, THE FOUNDATIONS OF for further discussion of the role and scope of set theory in the study of mathematics.

For coverage of related topics in the *Macropædia* and *Micropædia,* see the *Propædia,* sections 10/21 and 10/22.

This article is divided into the following sections:

Introduction to set theory

FUNDAMENTAL SET CONCEPTS

If the elements and sets to be considered are restricted to some fixed class of objects, such as the letters of the alphabet, the universal set (or the universe), which is commonly denoted by U, can then be defined as that which includes all of the elements—in this case, the set of all of the 26 letters. Thus, if A is one of the sets being considered, it will be understood that A is a subset of U. Another set may now be defined that includes all of the elements of U that are not elements of A. This set, which is called the complement of A, is denoted by A'. (Some writers, employing the convention of "difference sets," speak of "the complement of A with respect to U," which they denote by "$U-A$.")

Empty set

The empty (or void, or null) set, which is usually symbolized by Ø, contains no elements. One description of the empty set is that of all whole numbers that are neither even nor odd.

Operations on sets. The symbol $\cup$ is employed to denote the union of two sets. Thus, the set $A \cup B$—read "A union B" or "the union (or join) of A and B"—is defined as the set that consists of all elements belonging either to set A or set B. If sets A and B, however, have one or more members in common, their union will not duplicate those members. A committee, for example, consisting of Jones, Blanshard, Nelson, Smith, and Hixon (Committee A) may for some common purpose sit in joint session with Committee B, consisting of Blanshard, Morton, Hixon, Young, and Peters. Clearly, the union of Committees A and B must then consist of eight members rather than 10, namely, Jones, Blanshard, Nelson, Smith, Morton, Hixon, Young, and Peters.

The intersection operation is denoted by the symbol $\cap$. $A \cap B$—read "A intersect B" or "the intersection of A and B"—is defined as that set composed of all elements that belong to both A and B. Thus the intersection of the two committees in the foregoing example is the set consisting of Blanshard and Hixon.

If the set E denotes all positive even numbers and the set Q denotes all positive odd numbers, then the union of the two yields the entire sequence of positive natural numbers, and their intersection is the empty set. Any two sets the intersection of which is the empty set are said to be disjoint.

A product of two sets A and B, called a Cartesian product, is denoted by $A \times B$. This product is defined in terms of ordered pairs, analogous to the coordinates (or x and y values) of points on a Cartesian grid in analytic geometry. It is conventional to denote such pairs by enclosure in parentheses to differentiate them from unordered pairs, or sets, the members of which are enclosed in braces. In other words, the set $\{x, y\}$ is identical to the set $\{y, x\}$, but (x, y) is not the same as (y, x); two ordered pairs (a, b) and (c, d) are defined to be equal if and only if $a = c$ and $b = d$. The Cartesian product $A \times B$ may now be defined as the set consisting of all ordered pairs (x, y) for which x is an element of A and y is an element of B. An example is easily constructed: $A = \{x, y\}$, $B = \{3, 6, 9\}$, $A \times B = \{(x, 3), (x, 6), (x, 9), (y, 3), (y, 6), (y, 9)\}$.

Relations involved in set theory. The relations between sets can be of many sorts; *e.g.,* "is a subset of" ($\subseteq$), "is equivalent to" ($\sim$), "is a complement of" ($'$), "is in one-to-one correspondence with," and "has the same cardinal number as" (see below). In addition, pairing relations, defined in terms of some specific criterion, can exist between the individual elements of a set. Examples of pairing relations are: "is parallel to" ($\parallel$), "is equal to" ($=$), "is less than" ($<$), and "is the same colour as." More broadly

Pairing relations between elements

conceived, pairing can include the relations depicted on charts and graphs, on which, for example, calendar years may be paired with automobile production figures, weeks with Dow-Jones averages, and degrees of angular rotation with the lift accomplished by a cam.

The relation of one-to-one correspondence between two sets can be conceived as one in which each element of a set A is matched with an element of another set B. If $A = \{x, z, w\}$, for example, and $B = \{4, 3, 9\}$, then A is in one-to-one correspondence with B if and only if a matching such as 4 with x, 3 with z, and 9 with w obtains without any element in either set left unmatched as a remainder.

Many relations display identifiable similarities. The relations "is parallel to," "is the same colour as," and "is in one-to-one correspondence with," for example, all bear the stated relation to themselves as well as to other elements; thus, these relations are said to be reflexive. These same relations share, in addition, the property that, if an element bears the stated relation to a second element, then the second also bears that relation to the first—a property known as symmetry. Relations also have the property that, if two elements bear the stated relation to a third element, then they bear it to one another as well—a property known as transitivity.

Equivalence classes and numbers

Those relations that have all three properties—reflexivity, symmetry, and transitivity—are called equivalence relations. In an equivalence relation, all elements related to a particular element are related to each other, thus forming what is called an equivalence class. For the relation "is parallel to," for example, the equivalence class of a particular line ℓ is the set of all lines parallel to ℓ.

For each of the equivalence classes of sets, it is possible to construct an ordered set—for which not only the membership but also the sequence of its elements is significant—that can be used to name the class.

With appropriate qualifications, the cardinal of the empty set $\varnothing$ can be defined as 0; *i.e.,* $n(\varnothing) = 0$. The number 1, then, is assigned to be the cardinal of the set $\{0\}$ that contains only a single element; it is thus called the successor of 0. Similarly, the number 2, the cardinal of $\{0, 1\}$, is called the successor of 1; and 3, the cardinal of $\{0, 1, 2\}$, is the successor of 2. Continuing in this manner, the set $\mathfrak{N}$ of the natural numbers in the proper sequence $\{0, 1, 2, 3, 4, \cdot \cdot \cdot\}$ is obtained, for which the ordering is given by the successor relation.

It is this ordering that is used when one learns to count. The words one, two, three, four, · · · in proper sequence are associated with the elements in the set that are being counted. If this process stops, the set is said to be finite. Otherwise, it is said to be infinite.

There is a technical difference between cardinal and ordinal numbers. The distinction can be seen in the way that these numbers are used. A number used to designate the size of a set—*i.e.,* to answer the question, "How many?"—is used cardinally. Any use that depends on the position of the number in the prescribed sequence is the ordinal use of the number. The number found at the top or bottom of a page in a book is an example of the ordinal use of the number. (Jo.Ha./Ed.)

ESSENTIAL FEATURES OF CANTORIAN SET THEORY

At best, the foregoing description presents only an intuitive concept of a set. Essential features of the concept as Cantor understood it include: (a) that a set is a grouping into a single entity of objects of any kind; and (b) that, given an object x and a set A, exactly one of the statements "x is an element of A" (symbolized $x \in A$) and "x is not an element of A" (symbolized $(x \notin A)$) is true and the other is false. The definite relation that may or may not exist between an object and a set is called the membership relation.

Extension and abstraction

A further intent of this description is conveyed by what is called the principle of extension, viz., that a set is determined by its members: that sets A and B are equal (symbolized $A = B$) if and only if every element in A is also in B and every element in B is in A; or, in other terms, $x \in A$ implies $x \in B$ and vice versa. There exists, for example, exactly one set the members of which are 2, 5, and 7; and this set will be written by listing its elements in some order (which order is immaterial) between small braces, possibly $\{5, 2, 7\}$.

A set A is finite if for some natural number n there is a pairing of the elements of A with those of the initial segment $0, 1, \cdot \cdot \cdot, n-1$ of the natural numbers $0, 1, 2, \cdot \cdot \cdot$ in their usual order. This definition amounts to classifying a set as finite if it has a natural number n as its cardinal number (see below *Cardinality and transfinite numbers*). In principle, the brace notation is adequate for defining all finite sets.

To define infinite (*i.e.,* nonfinite) sets, Cantor used (sentential) formulas. The phrase "x is a professor" is an example of a formula; if the symbol "x" in this phrase is replaced by the name of a person, there results a declarative sentence that is true or false. The notation "$S(x)$" will be used to represent such a formula. The phrase "x is a professor at university y and x is a male" is a formula with two variables. If the occurrences of x and y are replaced by names of appropriate, specific objects, the result is a declarative sentence that is true or false. Given any formula $S(x)$ that contains the letter "x" (and possibly others) but not the letter "A," Cantor's principle of abstraction asserts the existence of a set A such that for each object x, $x \in A$ if and only if $S(x)$ holds. The unique (because of the principle of extension) set A corresponding to $S(x)$ is symbolized by $\{x \mid S(x)\}$ and read "The set of all objects x such that $S(x)$." For instance, $\{x \mid x \text{ is blue}\}$ is the set of all blue objects. This illustrates the fact that the principle implies the existence of sets the elements of which are all objects having a certain property. It is actually more comprehensive. For example, it asserts the existence of a set B corresponding to "Either x is an astronaut or x is a natural number." Astronauts have no property in common with numbers (other than both being members of B). The formula "$x \neq x$" defines the only set without elements. It is called the empty set and symbolized by $\varnothing$. The empty set is finite, for its members can be paired with those of the initial segment defined by $n = 0$.

Equivalent sets. Cantorian set theory is thus founded on the principles of extension and abstraction. To describe some results based upon these principles the notion of equivalence of sets will be defined. The set A is defined as equivalent to the set B (symbolized $A \sim B$) if and only if there exists a third set the members of which are ordered pairs such that: (a) the first member of each pair is an element of A and the second is an element of B, and (b) each member of A occurs as a first member and each member of B occurs as a second member of exactly one pair. Thus, if A and B are finite and $A \sim B$, then the third set that establishes this fact provides a pairing or matching of the elements of A with those of B. Conversely, if it is possible to match the elements of A with those of B, then $A \sim B$, because a set of pairs meeting requirements (a) and (b) can be formed (if $a \in A$ is matched with $b \in B$, then the ordered pair (a, b) is one member of the set). By thus defining equivalence of sets in terms of the notion of matching, it is formulated independently of finiteness. As an illustration involving infinite sets, $\mathfrak{N}$ may be taken to denote the set of natural numbers $0, 1, 2, \cdot \cdot \cdot$ (some authors exclude 0 from the natural numbers). Then $\{(n, n^2 \mid \in \mathfrak{N}\}$ establishes the equivalence of $\mathfrak{N}$ and the set of the squares of the natural numbers. (R.R.S.)

Equivalence and subsets

A set B is included in, or is a subset of, a set A (symbolized $B \subseteq A$) if every element of B is an element of A. So defined, a subset may possibly include all of the elements of A, so that A can be a subset of itself. Furthermore, the empty set, because it by definition has no elements that are not included in other sets, is a subset of every set.

If every element of set B is an element of set A, but the converse is false (hence $B \neq A$), then B is said to be properly included in, or is a proper subset of, A (symbolized $\mathrm{B} \subset \mathrm{A}$). Thus, if A is $\{3, 1, 0, 4, 2\}$, both $\{0, 1, 2\}$ and $\{0, 1, 2, 3, 4\}$ are subsets of A; but the latter is not a proper subset. A finite set is nonequivalent to each of its proper subsets. This is not so, however, for infinite sets, as is illustrated with the set $\mathfrak{N}$ in the earlier example. (The equivalence of $\mathfrak{N}$ and its proper subset of the even natural numbers was essentially the paradox noted by Galileo in 1638.) (R.R.S./Ed.)

Infinite sets and power sets

Cardinality and transfinite numbers. The application of the notion of equivalence to infinite sets was first systematically explored by Cantor. His initial significant finding was that the set of all rational numbers (see ARITHMETIC; ANALYSIS: *Real analysis*) is equivalent to $\mathfrak{N}$ but that the set of all real numbers is not equivalent to $\mathfrak{N}$. The existence of nonequivalent infinite sets justified Cantor's introduction of transfinite cardinal numbers as measures of size for such sets. Cantor defined the cardinal of an arbitrary set A as the concept that can be abstracted from A taken together with the totality of other equivalent sets. Gottlob Frege, in 1884, and Bertrand Russell, in 1902, both mathematical logicians, defined the cardinal number $\overline{\overline{A}}$ of a set A somewhat more explicitly, as the set of all sets that are equivalent to A; this definition thus provides a place for cardinal numbers as objects of a universe whose only members are sets.

These definitions are consistent with the usage of natural numbers as cardinal numbers. A natural number, 2 for example, is first assigned to the set $\{0, 1\}$ as a measure of its size; then 2 is assigned to every set equivalent to $\{0, 1\}$. Turning matters around, 2 is the concept that can be abstracted from the collection of sets equivalent to $\{0, 1\}$, or 2 may be defined as this collection of sets. Intuitively, a cardinal number, whether finite (*i.e.,* a natural number) or transfinite (*i.e.,* nonfinite), is a measure of the size of a set. Exactly how a cardinal number is defined is unimportant; what is important is that $\overline{\overline{A}} = \overline{\overline{B}}$ if and only if $A \sim B$.

To compare cardinal numbers, an ordering relation—denoted by $<$—may be introduced by means of the definition: $\overline{\overline{A}} < \overline{\overline{B}}$ if A is equivalent to a subset of B and B is equivalent to no subset of A. Clearly this relation is irreflexive $\overline{\overline{A}} \not< \overline{\overline{A}}$ and transitive $\overline{\overline{A}} < \overline{\overline{B}}$ and $\overline{\overline{B}} < \overline{\overline{C}}$ imply $\overline{\overline{A}} < \overline{\overline{C}}$.

When applied to natural numbers used as cardinals, $<$ coincides with the familiar ordering relation for $\mathfrak{N}$, so that $<$ is an extension of that relation.

The symbol $\aleph_0$ (aleph-null) is standard for the cardinal number of $\mathfrak{N}$ (sets of this cardinality are called denumerable) and $\aleph$ (aleph) is usually used for that of the set of real numbers. Then $n < \aleph_0$ for each $n \in \mathfrak{N}$ and $\aleph_0 < \aleph$.

This, however, is not the end of the matter. If the power set of a set A—symbolized $P(A)$—is defined as the set of all subsets of A, then, as Cantor proved, $\overline{\overline{A}} < \overline{\overline{P(A)}}$ for every set A—a relation that is known as Cantor's theorem. It implies an unending hierarchy of transfinite cardinals: $\overline{\overline{N}} = \aleph_0$, $\overline{\overline{P(N)}}$, $\overline{\overline{P(P(N))}}$, $\cdot\ \cdot\ \cdot$ Cantor proved that $\aleph = \overline{\overline{P(N)}}$ and was led to the question whether there is a cardinal number between $\aleph_0$ and $\aleph$, which is known as the continuum problem. A solution was completed in 1963 (see below).

There is an arithmetic for cardinal numbers based on natural definitions of addition, multiplication, and exponentiation (squaring, cubing, and so on), which deviates, however, from that of the natural numbers when transfinite cardinals are involved. For example, $\aleph_0 + \aleph_0 = \aleph_0$ (because the set of integers is equivalent to $\mathfrak{N}$), $\aleph_0 \cdot \aleph_0 = \aleph_0$ (because the set of ordered pairs of natural numbers is equivalent to $\mathfrak{N}$), and $c + \aleph_0 = c$ for every transfinite cardinal c (because every infinite set includes a subset equivalent to $\mathfrak{N}$).

The extension of the natural numbers as cardinal numbers to transfinite numbers described earlier is a typical facet of Cantorian set theory.

The Cantor paradox

The so-called Cantor paradox, discovered by Cantor himself in 1899, is the following: By the principle of abstraction, the formula "x is a set" defines a set U. It is the set of all sets. Now $P(U)$ is a set of sets and so $P(U)$ is a subset of U. By the definition of $<$ for cardinals, however, if $A \subseteq B$, then it is not the case that $\overline{\overline{B}} < \overline{\overline{A}}$. Hence, by substitution, $\overline{\overline{U}} \not< \overline{\overline{P(U)}}$. But by Cantor's theorem, $\overline{\overline{U}} < \overline{\overline{P(U)}}$. This is a contradiction. In 1902, Bertrand Russell devised another paradox of a less technical nature. The formula "x is a set and $(x \notin x)$" defines a set R of all sets not members of themselves. Using proof by contradiction, however, it is easily shown that (A) $R \in R$. But then by the definition of R it follows that (B) $(R \notin R)$. Together, (A) and (B) form a contradiction.

Axiomatic set theory

The attitude adopted in an axiomatic development of set theory is that it is not necessary to know what the "things" are that are called "sets" nor what the relation of membership means. Of sole concern are the properties assumed about sets and the membership relation. Thus, in an axiomatic theory of sets, the terms set and the membership relation $\in$ are undefined. The assumptions adopted about these notions are called the axioms of the theory. Its theorems are the axioms together with the statements that can be deduced from the axioms using the rules of inference provided by a system of logic. Criteria for the choice of axioms include: (A) their consistency (*i.e.,* that it should be impossible to derive as theorems both a statement and its negation), (B) their plausibility (*i.e.,* that they should be in accord with intuitive beliefs about sets), and (C) their richness (*i.e.,* that desirable results of Cantorian set theory can be derived as theorems).

These points are elaborated upon below.

POSTULATES OF AXIOMATIC SET THEORY

The Zermelo–Fraenkel axioms: discussion. The first axiomatization of set theory was given in 1908 by Ernst Zermelo, a German mathematician. From his analysis of the paradoxes, he concluded that they are associated with sets that are "too big," such as the set of all sets in Cantor's paradox. Thus, the axioms that Zermelo formulated are restrictive insofar as the asserting or implying of the existence of sets is concerned. As a consequence, there is no apparent way, in his system, to derive the known contradictions from them. On the other hand, the results of classical set theory short of the paradoxes can be derived. Zermelo's axiomatic theory is here discussed in a form that incorporates modifications and improvements suggested by later mathematicians, principally Thoralf Albert Skolem, a pioneer in metalogic, and Abraham Adolf Fraenkel, an Israeli mathematician. In the literature on set theory, it is called Zermelo–Fraenkel set theory (symbolized ZF), though it would seem historically more correct to call it Zermelo–Fraenkel–Skolem set theory. The 10 axioms are first discussed and then formally listed (see below *The Zermelo-Fraenkel axioms: formal presentation*).

Zermelo's pioneering work

Schemas for generating well-formed formulas. In the axioms that follow, "set" and "$\in$" are undefined terms. Lowercase Latin letters are used for variables; and variables denote sets. Equality (symbolized $=$) is taken as part of the underlying logic.

The first axiom (see axiom 1, below) conveys the idea that, as in classical set theory, a set is determined by its members. It should be noted that this is not merely a logically necessary property of equality but an assumption about the membership relation as well.

The set defined by the second axiom (see 2) is the empty (or null) set $\varnothing$.

For an understanding of the third axiom (see 3) considerable explanation is required. Zermelo's original system included the assumption (*Aussonderung* axiom) that, if a formula $S(x)$ is "definite" for all elements of a set s, then there exists a set the elements of which are precisely those elements x of s for which $S(x)$ holds. This is a version of the principle of abstraction, for it provides for the existence of sets corresponding to formulas. It restricts that principle, however, in two ways. Instead of asserting the existence of sets unconditionally, it can be applied only in conjunction with pre-existing sets. Further, only "definite" formulas (for which Zermelo offered only a vague description) may be used. Clarification was given, however, by Skolem (1922) by way of a precise definition of what will be called simply a formula of ZF. Using tools of modern logic, the definition may be made as follows:

a. For any variables x and y, $x \in y$ and $x = y$ are formulas (such formulas are called atomic).

b. If A and B denote formulas and x is any variable, then each of the following is a formula: If *A*, then *B; A* if and only if *B; A* and *B; A* or *B;* not *A;* for all *x, A;* for some *x, B.*

Formulas are constructed recursively (in a finite number of systematic steps) beginning with the (atomic) formulas of (a) and proceeding via the constructions permitted in (b). "Not ($x \in y$)," for example, is a formula (which is abbreviated to $x \notin y$), and "There exists an *x* such that for every *y*, $y \notin x$" is a formula. A variable is free in a formula if it occurs at least once in the formula without being introduced by one of the phrases "for some *x*" or "for all *x*." Henceforth, a formula *P* in which *x* occurs as a free variable will be called a "condition on *x*" and symbolized *P*(*x*). The formula "For every *y*, $x \in y$," for example, is a condition on *x*. It is to be understood that a formula is a formal expression—*i.e.,* a term without meaning. Indeed, a computer could be programmed to generate atomic formulas and build up from them other formulas of ever-increasing complexity using logical connectives ("not," "and," etc.) and operators ("for all" and "for some"). A formula acquires meaning only when an interpretation of the theory is spelled out; *i.e.,* when (A) a nonempty collection (called the domain of the interpretation) is specified as the range of values of the variables (thus the term set is assigned a meaning, viz., an object in the domain), (B) the membership relation is defined for these sets, (C) the logical connectives and operators are interpreted as in everyday language, and (D) the logical relation of equality is taken to be identity among the objects in the domain.

Intuitive interpretation

The terminology "a condition on *x*" for a formula in which *x* is free is merely suggestive; relative to an interpretation, such a formula does impose a condition on *x*. Thus, the intuitive interpretation of the third axiom schema is: given a set *a* and a condition on *x*, *P*(*x*), those elements of *a* for which the condition holds form a set. It provides for the existence of sets by separating off certain elements of existing sets. Calling the third axiom schema an axiom schema is appropriate, for it is a schema for generating axioms—one for each choice of *P*(*x*).

Axioms for compounding sets. Although the third axiom schema has a constructive quality, further means of constructing sets from existing sets must be introduced if some of the desirable features of Cantorian set theory are to be established. Each of the next three axioms is of this sort.

Using five of the axioms (see 2–6), a variety of basic concepts of classical set theory (*e.g.,* the operations of union, intersection, and Cartesian product; the notions of relation, equivalence relation, ordering relation, and function) can be defined with ZF. Further, the standard results about these concepts that were attainable in classical set theory can be proved as theorems of ZF.

Axioms for infinite and ordered sets. If *I* is an interpretation of an axiomatic theory of sets, the sentence that results from an axiom when a meaning has been assigned to "set" and "∈," as specified by *I*, is either true or false. If each axiom is true for *I*, then *I* is called a model of the theory. If the domain of a model is infinite, this fact does not imply that any object of the domain is an "infinite set." An infinite set in the latter sense is an object *d* of the domain *D* of *I* for which there is an infinity of distinct objects *d′* in *D* such that *d′Ed* holds (*E* standing for the interpretation of ∈). Though the domain of any model of the theory of which the axioms thus far discussed are axioms is clearly infinite, models in which every set is finite have been devised. For the full development of classical set theory, including the theories of real numbers and of infinite cardinal numbers, the existence of infinite sets is needed; thus the seventh axiom (see 7) is included.

The existence of a (unique) minimal set, ω, having properties expressed in the seventh axiom can be proved; its distinct members are Ø, {Ø}, {Ø, {Ø}}, {Ø, {Ø}, {Ø, {Ø}}}, · · · . These elements are denoted by 0, 1, 2, 3, · · · and are called natural numbers. Justification for this terminology rests with the fact that the Peano axioms, which can serve as a base for arithmetic, can be proved as theorems. Thereby the way is paved for the construction within ZF of entities that have all the expected properties of the real numbers.

The origin of the next axiom was Cantor's recognition of the importance of being able to well-order arbitrary sets; *i.e.,* to define an ordering relation for a given set such that each nonempty subset has a least element. The virtue of a well-ordering for a set is that it offers a means of proving that a property holds for each of its elements by a process (transfinite induction) similar to mathematical induction. Zermelo (1904) gave the first proof that any set can be well-ordered. His proof employed a set-theoretic principle that he called the axiom of choice, which, shortly thereafter, was shown to be equivalent to the so-called well-ordering theorem. One form of this principle is expressed as an eighth axiom (see 8).

Well-ordering arbitrary sets

Intuitively, the axiom asserts the possibility of making a simultaneous choice of an element in every nonempty member of any set; this guarantee accounts for its name. The assumption is significant only when the set has infinitely many members. Zermelo was the first to state explicitly the axiom, although it had been used but essentially unnoticed earlier. It soon became the subject of vigorous controversy because of its unconstructive nature. Some mathematicians rejected it totally on this ground. Others accepted it but avoided its use whenever possible. Some changed their minds about it when its equivalence with the well-ordering theorem was proved as well as the assertion that any two cardinal numbers *c* and *d* are comparable (*i.e.,* that exactly one of $c < d$, $d < c$, $c = d$ holds). There are many other equivalent statements, though even today there are a few mathematicians who feel that the use of the axiom of choice is improper. To the vast majority, however, it, or an equivalent assertion, has become an indispensable and commonplace tool.

Schema for transfinite induction and ordinal arithmetic. One more axiom has been added to the list of axioms (with modifications) postulated by Zermelo. When Zermelo's eight were found to be inadequate for a fullblown development of transfinite induction and ordinal arithmetic, Fraenkel and Skolem independently proposed an additional axiom schema to eliminate the difficulty. As modified by John von Neumann, a Hungarian-born U.S. mathematician, it says, intuitively, that if with each element of a set there is associated exactly one set, then the collection of the associated sets is itself a set; *i.e.,* it offers a way to "collect" existing sets to form sets. As an illustration, each of ω, *P*(ω), *P*(*P*(ω)), · · · is a set in the theory based on the first eight axioms. But there appears to be no way to establish the existence of the set having these sets as its members. An instance of the next schema, however, provides for its existence.

Intuitively, the ninth axiom or schema (see 9) is the assertion that if the domain of a function is a set then so is its range. That this is a powerful schema (in respect to the further inferences that it yields) is suggested by the fact that the third axiom can be derived from it and that, when applied in conjunction with the sixth axiom, the axiom of pairing can be deduced. The ninth axiom has played a significant role in developing a theory of ordinal numbers. In contrast to cardinal numbers, which serve to designate the size of a set, ordinal numbers are used to determine positions within a prescribed sequence. Following an approach conceived independently by Zermelo and von Neumann, if *x* is a set, the successor *x′* of *x* is the set obtained by adjoining *x* to the elements of *x* ($x' = x \cup \{x\}$). In terms of this notion the natural numbers, as defined above, are simply the succession 0, 0′, 0″, 0‴, · · · ; *i.e.,* the natural numbers are the sets obtained starting with Ø and iterating the prime operation a finite number of times. The natural numbers are well-ordered by the ∈ relation, and with this ordering they constitute the finite ordinal numbers. The axiom of infinity secures the existence of the set of natural numbers; and this set, ℜ ω, is the first infinite ordinal. Greater ordinal numbers are obtained by iterating the prime operation beginning with ω. An instance of the ninth axiom or schema asserts that ω, ω′, ω″, · · · form a set. The union of this set and ω is the still greater ordinal that is denoted by ω2 (employing notation from ordinal arithmetic). A repetition of this process beginning with ω2 yields the ordinals (ω2)′, (ω2)″, · · · ; next after all of those of this form is ω3. In this way the sequence of ordinals ω, ω2, ω3, · · ·

is generated. An application of the ninth axiom schema then yields the ordinal that follows all of these in the same sense in which ω follows the finite ordinals; using notation from ordinal arithmetic, it is ω^2. At this point the iteration process can be repeated. In summary, the axiom of replacement makes possible the extension of the counting process as far beyond the natural numbers as one chooses.

Cardinal and ordinal numbers

In the ZF system, cardinal numbers are defined as certain ordinals. From the well-ordering theorem (a consequence of the axiom of choice), it follows that every set a is equivalent to some ordinal number. Also, the totality of ordinals equivalent to a can be shown to form a set. Then a natural choice for the cardinal number of a is the least ordinal to which a is equivalent. This is the motivation for defining a cardinal number as an ordinal that is not equivalent to any smaller ordinal. The arithmetics of both cardinal and ordinal numbers have been fully developed. That of finite cardinals and ordinals coincides with the arithmetic of the natural numbers. For infinite cardinals, the arithmetic is uninteresting since, as a consequence of the eighth axiom, the sum and product of two such cardinals are each equal to the maximum of the two. In contrast the arithmetic of infinite ordinals is interesting and presents a wide assortment of oddities.

In addition to the guidelines already mentioned for the choice of axioms of ZF, another guideline is taken into account by some set theorists. For the purposes of foundational studies of mathematics, it is assumed that mathematics is consistent; for otherwise any foundation would fail. It may thus be reasoned that, if a precise account of the intuitive usages of sets by mathematicians is given, an adequate and correct foundation will result. Traditionally, mathematicians deal with the integers, with real numbers, and with functions. Thus an intuitive hierarchy of sets in which these entities appear should be a model of ZF. It is possible to construct such a hierarchy explicitly from the empty set by iterating the operations of forming power sets and unions in the following way.

The intuitive hierarchy of sets

The first level of the hierarchy is composed of the sequence of sets $s_0 = \varnothing, s_1, \cdots, s_n, \cdots$, in which s_{n+1} is the power set of s_n. The second level consists of the sets at the first level together with those sets obtained by iterating the power set operation any finite number of times. The third level has as its members the union of all sets constructed thus far together with those obtainable by iterating the power set operation as before. The hierarchy of sets envisaged, therefore, consists of all sets that can be obtained by proceeding to an arbitrarily large transfinite level. The domain of the intuitive model of ZF is conceived as the union of all sets in the hierarchy. In other words, a set x is in the model if it is an element of some set of the hierarchy.

Axiom for eliminating infinite descending species. From the assumptions that this system is sufficiently comprehensive for mathematics and that it is the model to be "captured" by the axioms of ZF, it may be argued that models of the first nine axioms that differ sharply from this system should be ruled out. The discovery of such a model led to the formulation by von Neumann of the tenth axiom (see 10).

This axiom eliminates from the models of the first nine axioms those in which there exist infinite descending ∈-chains (*i.e.*, sequences $x_1, x_2, x_3, \cdots$ such that $x_2 \in x_1$, $x_3 \in x_2, \cdots$), a phenomenon that does not appear in the heuristic model described above. (The existence of models having such chains was discovered by D. Mirimanoff in 1917.) It also has other attractive consequences; *e.g.*, a simpler definition of the notion of ordinal number is possible. Yet there is no unanimity among mathematicians whether there are sufficient grounds for adopting it as an additional axiom, since it does not have the immediate plausibility that even the axiom of choice has nor has it ever been shown to have any mathematical applications.

The Zermelo–Fraenkel axioms: formal presentation.

(1) *Axiom of extension.* If a and b are sets and if, for all x, $x \in a$ if and only if $x \in b$, then $a = b$.

(2) *Axiom of the empty set.* There exists a set a such that for all x, it is false that $x \in a$.

(3) *Axiom schema of separation.* If a is a set, there exists a set b such that for all x, $x \in b$ if and only if $x \in a$ and $P(x)$. Here, $P(x)$ is any condition on x in which b is not free (it must be bound by a quantifier such as "all" or "some").

(4) *Axiom of pairing.* If a and b are sets, there exists a set (symbolized $\{a, b\}$ and called the unordered pair of a and b) having a and b as its sole members.

(5) *Axiom of union.* If c is a set, there exists a set a such that $x \in a$ if and only if $x \in b$ for some member b of c.

(6) *Axiom of power set.* If a is a set, there exists a set b such that $x \in b$ if and only if $x \in a$.

(7) *Axiom of infinity.* There exists a set a such that $\varnothing \in a$ and, if $x \in a$, then $(x \cup \{x\}) \in a$, in which $x \cup \{x\}$ is the set x with x adjoined as a further member.

(8) *Axiom of choice.* If a is a set the elements of which are nonempty sets, then there exists a function f with domain a such that for member b of a, $f(b) \in b$.

(9) *Axiom schema of replacement.* If a is a set and $B(x, y)$ a formula (in which x and y are free) such that for $x \in a$ there is exactly one y such that $B(x, y)$, then there exists a set b the members of which are the y's determined by $B(x, y)$ as x ranges over a.

(10) *Axiom of restriction.* Every nonempty set a contains an element b such that $a \cap b = \varnothing$; *i.e.*, a and b have no elements in common.

The Neumann–Bernays–Gödel axioms: discussion. The second axiomatization of set theory originated with von Neumann in the 1920s. His formulation differed considerably from ZF because the notion of function, rather than that of set, was taken as primitive. In a series of papers beginning in 1937, however, the Swiss logician Paul Bernays, a collaborator with the formalist David Hilbert, modified the von Neumann approach in a way that put it in much closer contact with ZF. In 1940, the Czech-born logician Kurt Gödel, known for his undecidability proof, further simplified the theory. This version will be called NBG.

For expository purposes it is convenient to adopt two undefined notions for NBG: class and the binary relation, ∈, of membership (though, as is also true in ZF, ∈ suffices). In the axioms, capital Latin letters are used as variables. For the intended interpretation, variables take classes—the totalities corresponding to certain properties—as values. A class is defined to be a set if it is a member of some class; those classes that are not sets are called proper classes. Lowercase Latin letters are used as special restricted variables for sets. For example, "for all x, $A(x)$" stands for "for all X, if X is a set, then $A(X)$"; *i.e.*, the condition holds for all sets. Intuitively, sets are intended to be those classes that are adequate for mathematics, and proper classes are thought of as those collections that are "so big" that, if they were permitted to be sets, contradictions would follow. In NBG, the classical paradoxes are avoided by proving in each case that the collection on which the paradox is based is a proper class—*i.e.*, is not a set.

Comments about the axioms that follow are limited to features that distinguish them from their counterpart in ZF. The axioms are listed later (see 11–20, below *The Neumann–Bernays–Gödel axioms: statement*).

The third axiom or schema (see 13) is presented in a form to facilitate a comparison with the third axiom schema of ZF. In a detailed development of NBG, however, there appears, instead, a list of seven axioms (not schemas) that state that for each of certain conditions there exists a corresponding class of all those sets satisfying the condition. From this finite set of axioms, each an instance of the above schema, the schema (in a generalized form) can be obtained as a theorem. When obtained in this way, the third axiom schema of NBG is called the class existence theorem.

In brief, the fourth to eighth axioms of NBG (see 14–18) are axioms of set existence. The same is true of the next axiom, which for technical reasons is usually phrased in a more general form.

Finally there may appear in a formulation of NBG an analogue (see 20) of the last axiom of ZF.

Comparison of ZF and NBG axiomatizations

A comparison of the two theories that have been formulated is in order. In contrast to the ninth schema of ZF (see 9), that of NBG (see 19) is not an axiom schema but an axiom. Thus, with the comments above about the third axiom in mind, it follows that NBG has only a finite number of axioms. On the other hand, since the ninth axiom or schema of ZF provides an axiom for each formula, ZF has infinitely many axioms—which is unavoidable because it is known that no finite subset yields the full system of axioms. The finiteness of the axioms for NBG makes the logical study of the system simpler. The relationship between the theories may be summarized by the statement that ZF is essentially the part of NBG that refers only to sets. Indeed, it has been proved that every theorem of ZF is a theorem of NBG and that any theorem of NBG that speaks only about sets is a theorem of ZF. Finally, it has been shown that ZF is consistent if and only if NBG is consistent.

The Neumann–Bernays–Gödel axioms: statement.

(11) *Axiom of extension.* If A and B are classes and if, for all (sets) x, $x \in A$ if and only if $x \in B$, then $A = B$.

(12) Same as axiom (2).

(13) *Axiom schema for class formation.* If $P(x)$ is a condition on x in which (a) only set variables are introduced by the phrase "for all" or "for some" and (b) B is not free, then there exists a class B such that $x \in B$ if and only if $P(x)$.

(14) *Axiom of pairing.* Same as axiom (4).

(15) *Axiom of union.* Same as axiom (5).

(16) *Axiom of power set.* Same as axiom (6).

(17) *Axiom of infinity.* Same as axiom (7).

(18) *Axiom of choice.* Same as (8).

(19) *Axiom of replacement.* If (the class) X is a function and a is a set, then there exists a set b such that $y \in b$ if and only if for some x, $(x, y) \in X$ and $x \in a$; *i.e.*, the range of the restriction of a function X to a domain that is a set is also a set.

(20) *Axiom of restriction.* Every nonempty class A contains an element b such that $A \cap b = \emptyset$.

LIMITATIONS OF AXIOMATIC SET THEORY

The fact that NBG avoids the classical paradoxes and that there is no apparent way to derive any one of them in ZF does not settle the question of the consistency of either theory. One method for establishing the consistency of an axiomatic theory is to give a model; *i.e.*, an interpretation of the undefined terms in another theory such that the axioms become theorems of the other theory. If this other theory is consistent, then that under investigation must be consistent. Such consistency proofs are thus relative: the theory for which a model is given is consistent if that from which the model is taken is consistent. The method of models, however, offers no hope for proving the consistency of an axiomatic theory of sets. In the case of set theory and, indeed, of axiomatic theories generally, the alternative is a direct approach to the problem.

If T is the theory of which the (absolute) consistency is under investigation, this alternative means that the proposition "There is no sentence of T such that both it and its negation are theorems of T" must be proved. The mathematical theory (developed by the formalists) to cope with proofs about an axiomatic theory T is called proof theory, or metamathematics. It is premised upon the formulation of T as a formal axiomatic theory; *i.e.*, the theory of inference (as well as T) must be axiomatized. It is then possible to present T in a purely symbolic form; *i.e.*, as a formal language based on an alphabet the symbols of which are those for the undefined terms of T and those for the logical operators and connectives. A sentence in this language is a formula composed from the alphabet according to prescribed rules. The hope for metamathematics was that by using only intuitively convincing, weak number-theoretic arguments (called finitary methods), unimpeachable proofs of the consistency of such theories as axiomatic set theory could be given.

Gödel's theorem and the consistency of S

That hope suffered a severe blow in 1931 from a theorem proved by Gödel about any formal theory S that includes the usual vocabulary of elementary arithmetic. By coding the formulas of such a theory with natural numbers (now called Gödel numbers) and by talking about these numbers, Gödel was able to make the metamathematics of S to become part of the arithmetic of S and hence to be expressible in S. The theorem in question asserts that the formula of S that expresses (via a coding) "S is consistent" in S is unprovable in S if S is consistent. Thus, if S is consistent, then the consistency of S cannot be proved within S; rather, methods beyond those that can be expressed or reflected in S must be employed. Because, in both ZF and NBG, elementary arithmetic can be developed, Gödel's theorem applies to these two theories. Although there remains the possibility of a finitary proof of consistency that cannot be reflected in the foregoing systems of set theory, no hopeful, positive results have been obtained.

Other theorems of Gödel when applied to ZF (and there are corresponding results for NBG) assert that if the system is consistent, then (A) it contains a sentence such that neither it nor its negation is provable (such a sentence is called undecidable), (B) there is no algorithm (or iterative process) for deciding whether a sentence of ZF is a theorem, and (C) these same statements hold for any consistent theory resulting from ZF by the adjunction of further axioms. Apparently ZF can serve as a foundation for all of present-day mathematics because every mathematical theorem can be translated into and proved within ZF, or within extensions obtained by adding suitable axioms. Thus, the existence of undecidable sentences in each such theory (which entails the existence of true sentences of ZF not provable in ZF) points out the hopelessness of any attempt to base all of conceivable mathematics on a single axiomatic theory and hence implies the inadequacy of the axiomatic approach to mathematics, in particular, via axiomatic set theory.

PRESENT STATUS OF AXIOMATIC SET THEORY

The foundations of axiomatic set theory are in a state of significant change as a result of new discoveries. The situation is analogous to the 19th-century revolution in geometry, set off by the discovery of non-Euclidean geometries. It is difficult to predict the ultimate consequences of these late 20th-century findings for set theory, but already they have had profound effects on attitudes about certain axioms and have forced the realization of a continuous search for additional axioms. These discoveries have focussed attention on the concept of the independence of an axiom. If T is an axiomatic theory and S is a sentence (*i.e.*, a formula) of T that is not an axiom, and if $T + S$ denotes the theory that results from T upon the adjunction of S to T as a further axiom, then S is said to be consistent relative to T if $T + S$ is consistent and independent of T whenever both S and $\sim S$ (the negation of S) are consistent relative to T. Thus, assuming that T is consistent, if S is independent of T, then the addition of S or $\sim S$ to T yields a consistent theory. The role of the axiom of restriction (AR) can be clarified in terms of the notion of independence. If ZF′ denotes the theory obtained from ZF by deleting AR and either retaining or deleting the axiom of choice (AC), then it can be proved that if ZF′ is consistent, AR is independent of ZF′.

Relations of axiom of choice to continuum hypotheses

Of far greater significance for the foundations of set theory is the status of AC relative to the other axioms of ZF. The status in ZF of the continuum hypothesis (CH) and its extension, the generalized continuum hypothesis (GCH) are also of profound importance. (If $q(a)$ denotes "There does not exist a set b such that $\overline{\overline{a}} < \overline{\overline{b}} < \overline{\overline{P(a)}}$ the CH is $q(a)$ for $\overline{\overline{a}} = \aleph_0$ and GCH is $q(a)$ for all infinite sets a.) In the following discussion of these questions, ZF denotes Zermelo–Fraenkel set theory without AC. The first finding was obtained by Gödel in 1938. He proved that AC and GCH are consistent relative to ZF (*i.e.*, if ZF is consistent, then so is ZF + AC + GCH), by showing that a contradiction within ZF + AC + GCH can be transformed into a contradiction in ZF. In 1963, Paul J. Cohen, a U.S. mathematician, proved that (1) if ZF is consistent, then so is ZF + AC + ~CH, and (2) if ZF is consistent, then so is ZF + ~AC. Since in ZF + AC it can be demonstrated that GCH implies CH, Gödel's theorem together with Cohen's establishes the independence of AC and CH. For his proofs Cohen introduced a new method (called forcing) of constructing interpretations of ZF + AC. The method of forcing is applicable to many problems in set theory, and since 1963 it has been used to give independence proofs for a wide variety of highly technical propositions. Some of these results have opened new avenues for attacks on important foundational questions.

The current unsettled state of axiomatic set theory can be sensed by the responses that have been made to the ques-

Status of the continuum hypothesis

tion of how to regard CH in the light of its independence from ZF + AC. Someone who believes that set theory deals only with nonexistent fictions will have no concern about the question. But for most mathematicians sets actually exist; in particular, ω and *P*(ω) exist. Further, it should be the case that every nondenumerable subset of *P*(ω) either is or is not equivalent to *P*(ω); *i.e.*, either CH is true or is false. Followers of this faith regard the axioms of set theory as describing some well-defined reality—one in which CH must be either true or false. Thus there is the inescapable conclusion that the present axioms do not provide a complete description of that reality. A search for such axioms is in progress. One who hopes to prove CH as a theorem must look for axioms that restrict the number of sets. There seems to be little hope for this restriction, however, without changing the intuitive notion of the set. Thus the expectations favour the view that CH will be disproved. This disproof requires an axiom that guarantees the existence of more sets; *e.g.*, of sets having cardinalities greater than those that can be proved to exist in ZF + AC. So far, none of the axioms that have been proposed that are aimed in this direction (called "generalized axioms of infinity") serves to prove ~CH. Although there is little supporting evidence, the optimists hope that the status of the continuum hypothesis CH will eventually be settled. (R.R.S.)

BIBLIOGRAPHY. N. BOURBAKI, *Éléments de mathématique, théorie des ensembles* (1966; Eng. trans., *Theory of Sets,* 1968); G. CANTOR, *Beiträge zur Begründung der transfiniten Mengenlehre* (1895–97; Eng. trans., *Contributions to the Founding of the Theory of Transfinite Numbers,* 1915), of historical interest; P.J. COHEN, *Set Theory and the Continuum Hypothesis* (1966), proofs of the independence of AC and CH; K. GOEDEL, *The Consistency of the Axiom of Choice and of the Generalized Continuum-Hypothesis with the Axioms of Set Theory,* rev. ed. (1953); W.S. HATCHER, *Foundations of Mathematics* (1968), an overall view of axiomatic set theory and its relationship to the foundations of mathematics; E. MENDELSON, *Introduction to Mathematical Logic* (1964), a formal development of NBG; J.R. SHOENFIELD, *Mathematical Logic* (1967), a development of ZF and the independence proofs; R.R. STOLL, *Set Theory and Logic* (1963), an informal development of ZF; P. SUPPES, *Axiomatic Set Theory* (1960); AZRIEL LÉVY, *Basic Set Theory* (1979), an explanation for the advanced student; and I. GRATTAN-GUINNESS (ed.), *From the Calculus to Set Theory, 1630–1910: An Introductory History* (1980).

(Jo.Ha./R.R.S.)

Sex and Sexuality

Sex, sexuality, and reproduction are all closely interwoven into the fabric of living things. All relate to the propagation of the race and the survival of the species. Yet there can be sex without sexuality, and reproduction need not be sexual, although for most forms of life sexual reproduction is essential both for propagation and long-term survival.

The article is divided into the following sections:

Animals and plants

SEXUAL AND NONSEXUAL REPRODUCTION

Nonsexual reproduction

Because the life-span of all individual forms of life, from microbes to man, is limited, the first concern of any particular population is to produce successors. This is reproduction, pure and simple. Among lower animals and plants it may be accomplished without involving eggs and sperm. Ferns, for example, shed millions of microscopic, nonsexual spores, which are capable of growing into new plants if they settle in a suitable environment. Many higher plants also reproduce by nonsexual means. Bulbs bud off new bulbs from the side. Certain jellyfish, sea anemones, marine worms, and other lowly creatures bud off parts of the body during one season or another, each thereby giving rise to populations of new, though identical, individuals. At the microscopic level, single-celled organisms reproduce continually by growing and dividing successively to give rise to enormous populations of mostly identical descendants. All such reproduction depends on the capacity of cells to grow and divide, which is a basic property of life. In the case of most animals, however, particularly the higher forms, reproduction by nonsexual means is apparently incompatible with the structural complexity and activity of the individual.

Although nonsexual reproduction is exploited by some creatures to produce very large populations under certain circumstances, it is of limited value in terms of providing the variability necessary for adaptive advantages. Such so-called vegetative forms of reproduction, whether of animals or plants, result in individuals that are genetically identical with the parent. All have exactly the same genetic, or inherited, traits, for better or worse. If some adverse environmental change should occur, all would be equally affected and none might survive. At the best, therefore, nonsexual reproduction can be a valuable and perhaps an essential means of propagation, but it does not exclude the need for sexual reproduction.

Sexual reproduction not only takes care of the need for replacement of individuals within a population but gives rise to populations better suited to survive under changing circumstances. In effect it is a kind of double assurance that the race or species will persist for an indefinite time. The great difference between the two types of reproduction is that individual organisms resulting from nonsexual reproduction have but a single parent and are essentially alike, whereas those resulting from sexual reproduction have two parents and are never exact replicas of either.

Sexual reproduction thus introduces a variability, in addition to its propagative function. Both types of reproduction represent the capacity of individual cells to develop into whole organisms, given suitable circumstances. Sex is therefore something that has been combined with this primary function and is responsible for the capacity of a race to adapt to new environmental conditions.

Gonads and primary sex

Sex cells. The term sex is variously employed. In the broad sense it includes everything from the sex cells to sexual behaviour. Primary sex, which is generally all that distinguishes one kind of individual from another in the case of many lower animals, denotes the capacity of the reproductive gland, or gonad, to produce either sperm cells or eggs or both. If only sperm cells are produced, the reproductive gland is a testis, and the primary sex of the tissue and the individual possessing it is male. If only eggs are produced, the reproductive gland is an ovary, and the primary sex is female. If the gland produces both sperm and eggs, either simultaneously or successively, the condition is known as hermaphroditic. An individual, therefore, is male or female or hermaphrodite primarily according to the nature of the gonad.

As a rule, male and female complement each other at all levels of organization: as sex cells; as individuals with either testes or ovaries; and as individuals with anatomical, physiological, and behavioral differences associated with the complemental roles they play during the whole reproductive process. The role of the male individual is to deliver sperm cells in enormous numbers in the right place and at the right time to fertilize eggs of female individuals of the same species. The role of the female individual is to deliver or otherwise offer eggs capable of being fertilized under precise circumstances. In the case of hermaphrodite organisms, animal or plant, various devices are employed to ensure cross-fertilization, or cross-pollination, so that full advantage of double parentage is obtained. The basic requirement of sexual reproduction is that reproductive cells of different parentage come together and fuse in pairs. Such cells will be genetically different to a significant degree, and it is this feature that is essential to the long-term well-being of the race. The other sexual distinctions, between the two types of sex cell and between two individuals of different sex, are secondary differences connected with ways and means of attaining the end.

Sexuality: complementary mating types. The complementarity of both male and female sex cells and male and female individuals is a form of division of labour. Male sex cells are usually motile cells capable of swimming through liquid, either freshwater, seawater, or body fluids, and they contribute the male cell nucleus but little else to the fertilization process. The female cell also contributes its nucleus, together with a large mass of cell substance necessary for later growth and development following fertilization. The female cell, however, is without any capacity for independent movement.

In other words, small male cells (sperm cells, spermatozoa, or male gametes) are burdened with the task of reaching a female cell (egg, ovum, or female gamete), which is relatively large and awaits fertilization. A full complement of genes is contributed by both nuclei, representing contributions by both parents, but, apart from the nucleus, only the egg is equipped or prepared to undergo development to form a new organism. A comparable division of labour is seen in the distinction between male and female individuals. The male possesses testes and whatever accessory structures may be necessary for spawning or delivery of the sperm, and the female possesses ovaries and what may be needed to facilitate shedding the eggs or to nurture developing young. Accordingly there is the basic sex, which depends on the kind of sex gland present, and sexuality, which depends on the different structures, functions, and activities associated with the sex glands.

The role of each sex

THE ADAPTIVE SIGNIFICANCE OF SEX

When two reproductive cells from somewhat unlike parents come together and fuse, the resulting product of development is never exactly the same as either parent. On the other hand, when new individuals, plant or animal, develop from cuttings, buds, or body fragments, they are exactly like their respective parents, as much alike as identical twins. Any major change in environmental circumstances might exterminate a race since all could be equally affected. When eggs and sperm unite, they initiate development and also establish genetic diversity among the population. This diversity is truly the spice of life and one of the secrets of its success; sex is necessary to its accomplishment.

In each union of egg and sperm, a complete set of chromosomes, representing a complete set of genes, is contributed by each cell to the nucleus of the fertilized egg. Consequently, every cell in the body inherits the double set of chromosomes and genes derived from the two parental cells. Every time a cell divides, during the long period of development and growth, each daughter cell receives exact copies of the original two sets of chromosomes. The process is known as mitosis. Accordingly, any fragment of tissue has the same genetic constitution as the body as a whole and therefore inevitably gives rise to an identical individual if it becomes separated and is able to grow and develop. Only in the case of the tissue that produces the sex cells do cells divide differently, and genetic differences occur as a result.

During the ripening of the sex cells, both male and female, cell divisions occur (known as meiosis) that result in each sperm and egg cell having only a single set of chromosomes. In each case the set of chromosomes is complete—*i.e.*, one chromosome of each kind—but each such set is, in effect, drawn haphazardly from the two sets present in the original cells. In other words, the single set of chromosomes present in the nucleus of any particular sperm or egg, while complete in number and kinds, is a mixture, some chromosomes having come from the set originally contributed by the male parent and some from the female. Each reproductive cell, of either sex, therefore contains a set of chromosomes different in genetic detail from that of every other reproductive cell. When these in turn combine to form fertilized eggs or fertile seeds, the double set of chromosomes characteristic of tissue cells is reestablished, but the genetic constitution of all such cells in the new individual will be the same as that of the fertilized egg—two complete sets of genes, randomly derived from sets contributed by the two different parents. Variation is thus established in two steps. The first is during the ripening of the sex cells, when each sperm or egg receives a single set of chromosomes of mixed ancestry. None of these cells will have exactly the same combination of genes characteristic of the respective parent. The second step occurs at fertilization, when the pair of already genetically unique sex cells fuse together and their nuclei combine, thus compounding the primary variation.

Reproduction and evolution. Sexual reproduction appears to be a process serving two opposing needs. The individuals produced must be almost exactly like their parents if they are to succeed; *i.e.*, to grow and reproduce in turn, under the prevailing circumstances. At the same time they should exhibit a wide range of differences so that some at least can survive under different environmental circumstances. The first business of reproduction is to produce perfect working copies of the parental organism, without any mistakes. The second is to introduce novelties; *i.e.*, new models that make possible other life styles. Extreme conservatism, in either sexual or nonsexual reproduction, may be disastrous to the species in the long run. Extreme variability may also be detrimental, resulting in the production of too high a percentage of misfits. A delicate balance has to be struck. Variability is necessary but must be kept within bounds. Sex is responsible for controlled diversity, without which adaptation and evolution could not take place.

Significance of reproduction

Natural selection operates in two ways on this basic diversity inherent in any particular population or community. In a stable environment, where there is little change during a long period of time, except for the regular changes associated with the seasons and the daily cycle, those individuals most likely to survive and produce offspring are precisely those that are most like their parents at all stages of their existence. The more radical departures from the established types fail either to grow or to compete successfully and consequently do not reproduce. The less radical departures struggle along but leave progeny in proportionately smaller numbers. If, however, a significant long-term change occurs in the environment, the previously established types are likely to suffer, while other types that previously had been weeded out generation by generation now may be favoured. They may become the more successful at surviving and growing and consequently replace themselves more readily than do others. They, in turn, become the establishment, and the older type is jeopardized. A constant interplay persists between a changeable environment and a variable population. This is adaptation. If environmental change continues in the same general direction, adaptation also continues in the initial direction, and eventually significant evolution becomes apparent.

The variability or diversity resulting from sexual reproduction is vital in two ways. It permits the process of natural selection to work and allows a population of organisms to adapt to new conditions. It also serves as a corrective mechanism. During nonsexual reproduction, particularly of single-cell organisms, large populations of virtually identical individuals are readily built up and maintained for a great many generations. Sooner or later, however, more and more abnormalities appear and, usually, a general waning of vigour ensues. When such organisms subsequently fuse together in pairs, equivalent to sexual reproduction, a rejuvenation and reestablishment of healthy strains generally follows.

Alteration of sexual and nonsexual phases

Life cycles adjusted to environmental change. Both sexual and nonsexual reproduction may be exploited or adjusted to meet widely fluctuating environmental conditions, especially those of a regular seasonal character. This phenomenon is particularly striking in the case of the smaller or simpler forms of animal and plant life that have a life-span of a year or less. The seeds of annual plants germinate in the spring, grow and set seed in turn during the summer, and die in the fall. Only the sexually produced seeds persist and represent the species during the long winter season. Certain small, though common, freshwater creatures have a similar cycle. The microscopic eggs of *Hydra* and of *Daphnia,* for example, lie at the bottom of ponds throughout the winter, each within a tough protective case. In late winter or early spring, a new generation of hydras develops, each individual becoming attached to a stone or vegetation and feeding on small crustaceans by means of its long slender tentacles. The daphnias, or so-called water fleas, emerge at about the same time and grow rapidly to maturity. In both cases the growing season, usually from spring until fall, is a time for intensive reproduction by whatever means is most effective. Hydras bud off new hydras continually, each new hydra repeating the process, with the size of populations limited only by available food. Only late in the season, when the food supply drops off and the temperature drops, does the riotous splurge of nonsexual reproduction come to an end. Then each individual ceases to bud and produces either minute ovaries or testes, and in some species, both. Eggs become fertilized, encased, drop into the mud, and await the coming of the following spring, while the parental creatures die as living conditions worsen with approaching winter. Such is a general pattern of life, widely seen among creatures whose individual existence is measured in weeks or months but whose race must persist in some form at all times if extinction is to be avoided.

So it is with *Daphnia* and many others. The *Daphnia* also changes according to the times, but it alternates between one form of sexual reproduction and another. Sexually, it is exquisitely adapted to the little world in which it lives. Under ideal conditions every member of a *Daphnia* community is female. All those first hatching out from winter eggs in the spring are females. Each produces a succession of broods during the month or two of its individual existence, all offspring being females. Each such female, generation after generation, during the spring and summer season, produces eggs that develop at once without need or opportunity of being fertilized. No males in fact are present. Every individual is a self-sufficient breeding female. Population explosions occur wherever environmental circumstances are good. Eventually, however, conditions inevitably change for the worse, either because of effects inherent in any population explosion or because every season comes to an end. Food becomes scarce because of too many consumers; space becomes crowded and in some degree polluted; chilly days succeed the warmth of summer. Whatever it is, and well before disaster can strike, the creatures respond in remarkable ways. On the first signal that conditions may be getting less than good, a certain number of the eggs produced by a population of *Daphnia* develop into males, each with testes in place of ovary, together with certain secondary sexual characteristics. A scattering of males through the virgin paradise, however, is only a first step, a preparedness in case conditions go from bad to worse. If there has been a false alarm, the females continue to produce female-producing eggs that develop parthenogenetically—*i.e.*, without benefit of fertilization—and the males die off without performing any sexual function. But if the environmental signal means the beginning of the end of congenial conditions, a cell in the ovary of each female grows to form a larger egg than usual, and it is of a type that must be fertilized. Then mating between the sexes takes place, and these special, fertilized eggs become thickly encased and alone survive the winter season after becoming separated from the parent.

Parthenogenesis

Wherever small aquatic creatures live in bodies of water that may freeze in winter or dry up in summer, similar adaptations are seen in many forms of life besides hydras and water fleas. Certain small fish, known as the annual fishes, have individual life-spans of about six months. The life-span itself is in fact adapted to the period during which active existence is possible in their particular habitat. When the water holes, swamps, and puddles in which they live begin to dry up, mating takes place, and the fertilized eggs drop into the mud. The parents die, and the eggs remain in a state of suspended development until the next rainy season occurs. The race must go on whatever the circumstances, and all sex is directed toward this end.

THE ORIGIN OF SEX AND SEXUALITY

All sexual reproduction, no matter how large or small the organisms may be, is a performance of single cells. Only at the level of single cells can the essential genetic recombinations be accomplished. So in every generation new life begins with the egg, which is a single cell, however large it may be. Egg and sperm unite at fertilization, but the fertilized egg is as much a single cell as before. When did it all begin? The generally accepted answer is that the fundamental, or molecular, basis of sexuality is an ancient evolutionary development that goes back almost to the beginning of life on earth, several billion years ago, for it is evident among the vast world of single-celled organisms, including bacteria.

In these lowest forms of life, sex and reproduction are distinct happenings. Reproduction is accomplished in most cases entirely by fission, which is simply cell division repeated regularly, as long as the environmental conditions permit. As long as crowding and other adverse changes are avoided, cells divide, and the daughter cells grow and divide again, for weeks or months on end. This process occurs in both plantlike and animal-like single-celled organisms and in bacteria as well. Under certain other conditions, such cell organisms come together and fuse in pairs, a form of sexual behaviour at its primary level and comparable to the fusion of an egg and sperm. In all such case, a combined cell is produced in which nuclear exchange or recombination has occurred. Pairing off of this sort takes place sooner or later in all forms of unicellular life, even where no outwardly distinguishable differences can be detected between the pairing individuals. The lack of discernible differences between the members of mating pairs, however, does not mean that pairing occurs between identical individuals. In the much investigated *Paramecium* and other protozoan organisms, two separate populations of cells may continue to increase almost indefinitely by ordinary cell division of single individuals, but when two such populations are mixed together, mating generally occurs immediately between individuals from the two different sources. The fusion, or pairing, has essentially the same function as the fusion of the male and female nucleus during the process of fertilization of eggs of higher forms. It is the basis of sex, the essential event in all cases being the genetic or chromosomal recombination.

Kinds of gametes

Individual mating cells (*i.e.*, eggs, sperm, or even whole single-celled organisms) may be called gametes whether or not they are distinguishable from one another. Yet even among the varius single-celled organisms, mating commonly occurs between individuals of two different kinds. This kind of mating is seen most often among the single-celled organisms known as flagellates. In some species the gametes may be alike and all are motile, progressing through the water by means of one or more whiplike flagella similar to the tail of a sperm. In other species, all individuals may still be motile, but pairing occurs between individuals of different sizes. In still others, one of the two mating types may be very small and motile, and the other, large, with stored nutritional material, and nonmotile. All degrees of differentition between male and female gametes can be found, and it is probable that the basic and characteristic distinction between the sex cells of both animal and plant life in general was established very early in the course of evolution, during the immense period of time when virtually all living organisms consisted of single cells.

This division of labour between mating types, male and female, respectively, is nature's way of attaining two ends. These are the bringing together of the gametes so that fusion may take place and the accumulation of reserves so that development of a new organism can be accomplished. The first calls for as many motile cells as possible; the second calls for cells as large as possible. These different requirements are practically impossible to satisfy by a single type of cell. Accordingly, and especially in multicelled animals of all sorts, male gametes, or spermatozoa, are extremely small, extremely motile, and are produced in enormous numbers. The larger the number, the greater the likelihood that some will encounter and fertilize eggs. On the other hand, the female gametes, or ova, individually need to be as large as possible since the larger the size and the more condensed the internal nutritional reserves, the farther along the path of embryonic development the egg can travel before hatching must occur and the new organism must fend for itself. Nevertheless, eggs in general are caught between the desirability of being individually as large as they can be and the persisting need to be produced in reasonably large numbers, so that an assortment of differing indivuduals is produced from a single pair of parents. A large number of offspring ensures that a proportion, at least, will survive the environmental hazards faced by all developing organisms in some degree.

Differentiation of the sexes. Animals and plants, apart from microscopic kinds of life, consist of enormous numbers of cells coordinated in various ways to form a single organism, and each consists of many different kinds of cells specialized for performing different functions. Certain tissues are set aside for the production of sexual reproductive cells, male or female as the case may be. Whether they are testes or ovaries or, as in some animals and plants, both together in the same parental individual, they are typically contained within the body, and therefore the sex cells usually need to be passed to the outside in order to function. Only in certain lowly creatures such as hydras is there a simpler state, for in hydras the testes and ovaries form in the outermost layer of cells of the slender, tubular body, and the sex cells when ripe burst directly from the simple, bulging gonads into the surrounding water. With few other exceptions, in all other creatures the gonads are part of the internal tissues and some means of exit is necessary. In some, such as most worms, all that is needed are small openings, or precisely placed pores, in the body wall through which sperm or eggs can escape. In most others, more is needed and a tubular sperm duct or an oviduct leads from each testis or ovary, through which the sex cells pass to the exterior. This is minimal equipment, except where none is needed. The gonad and its duct is accordingly comparable to other glands in the body; that is, the gland is generally a more or less compact mass of cells of a particular, specialized kind, together with a duct for passage of the product of the tissue to the site of action. Gonads secrete—*i.e.*, produce and transmit—sex cells that usually act outside the body.

Differentiation between the sexes exists, therefore, as the primary difference represented by the distinction between eggs and sperm, by differences represented by nature of the reproductive glands and their associated structures, and lastly by differences, if any, between individuals possessing the male and female reproductive tissues, respectively.

The value of cross-fertilization

Sex cells, sexual organs, other sexual structures, and sexual distinction between individuals constitute a series of evolutionary advances connected with various changes and persisting needs in the general evolution of animals and, to some degree, of plants as well. In other words, no matter how large or complex a creature may become, it still needs to deliver functional sex cells to the exterior. This condition is almost always the case for sperm cells. Among aquatic animals, particularly marine animals whose external medium, the ocean, is remarkably similar chemically to the internal body fluid medium of all animals, eggs are also in most cases shed to the exterior, where development of the fertilized eggs can proceed readily. Even so, time and place are important. Starfish, sea urchins, and many others, for instance, accumulate mature eggs and sperm in the oviducts and sperm ducts until an appropriate time when all can be shed at once. When one member of a group of such creatures begins to spawn, chemicals included in the discharge stimulate other members to do the same, so that a mass spawning takes place. One might say that the more they are together the more variable their offspring may be. This situation actually is the crux of the matter for nearly all forms of life, because while it may be possible for a single individual to possess both male and female gonads, producing both sperm and eggs, it remains generally desirable, if not essential, that eggs be fertilized by sperm produced by another individual. Cross-fertilization results in a much greater degree of variability than does self-fertilization. The existence of two types of individuals, male and female, is the common means of ensuring that cross-fertilization will be accomplished, since then nothing else is possible. Where the sexes are separate, therefore, all that is necessary is that members of the opposite sex get together at a time and place appropriate for the initial development of fertilized eggs. Typically, spawning of this sort is a communal affair, with many individuals of each sex discharging sex cells into the surrounding water. This process is only suitable, however, when eggs are without tough protective cases or membranes; that is, only when eggs are readily fertilizable for some time after being shed and while drifting in the sea. In this circumstance there is no need for individuals of the opposite sex to mate in pairs, nor is such mating practiced.

Mating. Mating between two individuals of the opposite sex becomes necessary when eggs must be fertilized at or before the time the eggs are shed. Whenever eggs have a protective envelope of any kind through which sperm cannot penetrate, fertilization must take place before the envelope is formed. The envelope may at first be a gluey liquid, which covers the egg and solidifies as a tough egg case, as in all crustaceans, insects, and related creatures. It may be a thick membrane of protein deposited around the egg, as in fishes generally; or it may be a material that swells up as a mass of jelly surrounding the eggs after the eggs have been shed, as in frogs and salamanders. And finally, it may be a calcified shell, as in birds and reptiles. In all of these organisms the sperm must reach the egg before the protective substance is added, except in those forms in which a small opening or pore persists in the egg membrane through which sperm can enter.

When and how such eggs need to be fertilized depends on the nature of the protective membranes and the time and place of their formation. The jelly surrounding frog and toad eggs, for instance, swells up immediately after the eggs are shed. Mating and fertilization must take place at the time of spawning. Male frogs mount the back of female frogs and each clasps his mate firmly around the body, which not only helps press the egg mass downward but brings the cloacal opening of male and female close together. Eggs and sperm are shed simultaneously, and the eggs are fertilized as they leave the female body. Fish eggs are also fertilized as or shortly after they are shed, although fish have no arms and mating generally is usually no more than a coming together of the two sexes side by side, so that simultaneous shedding of sperm and eggs can be accomplished. In other creatures the mating procedure may be much more complicated, depending on various circumstances. Crustaceans such as crabs and lobsters, for example, mate in somewhat the same manner as frogs, with the male holding on to the female by means of claw-like appendages and depositing sperm at the openings of the oviducts, which are typically situated near the middle of the undersurface of the body.

Mating modifications imposed by the land environment. Greater problems arise on land than in water. Eggs produced by truly terrestrial creatures are either retained in the parental body during their development or must be fully protected from drying up. Protective membranes must be tough indeed. More importantly, however, sperm cells must still be deposited where they can swim toward the eggs, for they cannot survive or function except in a watery solution of dilute salts. In all terrestrial creatures, except those that return to water to breed, sperm can survive only in the body of the male or female organism. All insects, therefore, must mate in order for eggs to be fertilized, and all have appendages at the rear of the body that serve as a copulatory device capable of being used even when in flight. Sperm is injected into the female's duct or storage sac, either for immediate fertilization or for later use. The queens of bees, ants, and termites, in fact, mate once and for all during a nuptial flight and thereafter use the stored sperm to fertilize all the eggs they subsequently produce.

The land vertebrates have to cope with much the same breeding circumstances as the insects. Man is more aware of these procedures because they happen mostly in much larger creatures and also because he has some fellow feeling for them. Reptiles, birds, and even the most primitive surviving mammals—namely, the platypus and spiny anteater of Australasia—produce yolky eggs encased in a more or less rigid calcareous shell. Moreover, within the shell, a thick layer of albumen surrounds the egg proper. Both the albumen and the shell are added after the ovum leaves the ovary and during its passage down the oviduct. Fertilization must take place, if at all, as the eggs enter the oviduct, for neither the albumen nor the shell can be penetrated by spermatozoa. Sperm must therefore be introduced into the female and must be able to make their way up to the end of the oviduct, which is a very long journey for so small a cell. An enormous number must begin the journey to make sure that some will reach the goal.

Necessity of internal fertilization in land animals

Sexual anatomy. In reptiles and birds of both sexes, as in amphibians and fish, a single opening to the exterior serves jointly for both the intestine and reproductive duct. This is the cloaca, or vestibule. Nevertheless, copulation of a sort occurs in all three groups of terrestrial vertebrates: the reptiles, birds, and mammals. With the exception of man, the male always mounts the female from the rear or back, and in both reptiles and birds the cloacal openings are pressed closely together to form a continuous passage from one individual to the other. With one exception, the archaic tuatara (*Sphenodon*) of New Zealand, all present-day reptiles have an erectile penis, derived from the cloacal wall, that delivers the sperm into the proper duct. One mating may serve for a long time, and there are cases known in which female snakes have laid fertile eggs after months and sometimes years of isolation in captivity. On the other hand, a penis of any sort is lacking in most kinds of birds, and the pressing together of the cloacal apertures seems to serve well enough. The most advanced copulatory procedure is that of mammals. In mammals the cloaca has become replaced by separate openings for the reproductive duct and intestine, respectively. Eggs have become microscopic, devoid of shell, yolk, and virtually all albumen, although they still need to be fertilized as they enter the upper end of the oviduct. A well-developed, erectile penis is always present in the male for the ejaculation of stored sperm well up the reproductive passage of the female. Accordingly, the two sexes have become strikingly differentiated anatomically, with regard to delivery of sperm, compared with the seemingly primitive anatomical equipment of birds.

Courtship. The coming together of two members of the opposite sex is a necessary preliminary to mating. It may be accomplished by two individuals independently of any larger congregation, or it may result from two individuals pairing off within a breeding population that may have assembled even from the ends of the earth. In the one the problem is to find one another; in the other the problem is to find the appropriate place, called the staging area. In both cases timing and some sort of navigation are important. Mass assembly appears to be the more effective, although a local crowd of any kind of animal may be an open invitation to predators, human or otherwise, and may on occasion become disastrous.

The searching out of a solitary individual by another of the opposite sex can be a difficult matter. In the dark depths of the ocean, for instance, where fish and other marine life forms are extremely scarce and scattered, the chance of encounter is rare indeed. The small angler fish (*Photocorynus spiniceps*) that cruise around at great depths are most unlikely to meet a member of the opposite sex at a time or place when the female happens to be ready to shed her eggs. As a form of insurance to this end, however, any small, young male that happens to meet a large female, apparently at any time, immediately fastens on to her head or sides by his jaws and thereafter lives a totally parasitic existence sustained by the juices of the female body. Sperm thus becomes available at any time the female may produce eggs to be fertilized.

On land this individual procedure of searching out is common among insects and the more predatory mammals. Male crickets and cicadas sound their familiar signals, by night or by day, which attract any females within hearing distance. More remarkable are those insects and other creatures that produce living light, in some cases for no apparent purpose but in others, such as the firefly, for signalling between the sexes in the dark of summer nights. The male individuals, always more dispensable than females, fly freely at considerable risk, flashing their light at regular intervals. The light of the female, perched more safely on some tall grass, winks back as though it were a landing light, and so they come together. Each of the several species of firefly has its own flash code, or rhythm, and any wasteful attempt at interspecies mixing is avoided. On the same principle, female moths send their personal perfume into the night air, and those males that detect the scent fly toward the source, the winner taking all. Mammals also depend mainly on their sense of smell, being generally colour blind, not too attentive to sound, and, apart from the grazing and browsing creatures, mainly active at night. The scented sex appeal of a cat in heat, whether domestic or wild, excites all the males in the neighbourhood and, with or without the sound of voice, male and female come quickly together in the dark. In all of these, courting is mostly uncalled for since only ready-to-mate individuals are involved in this sexual searching in the dark.

Sexual attractants

Courting is necessary whenever the male is a supplicant. A female may not be ready to mate, and stimulation in the form of dance or song may be required to create the mood; or, as is commonly the case, there is a surplus of available and eager males, and one must be chosen among many. However it may be, courting is most practiced not only when the female is in command of the final outcome but also when the mating procedure presents certain difficulties. A small male spider dances before a larger and ever ravenous female in an effort to induce her sexual interest rather than her hunger. Birds especially, however, depend on courtship as a preliminary to mating. The mating of birds represents copulation in its simplest form, without benefit of significant anatomical devices. Bird wings are a poor substitute for arms in a sexual embrace. Consequently the fullest cooperation between male and female is essential to success. In most birds a long-lasting, often lifetime, bonding becomes established between a male and female, a bonding that is usually reinforced by ritual behaviour at certain intervals, particularly during the onset of each breeding season and on various occasions when the individuals meet after short periods of separation. In some species a new mate may be taken each season or, as in sparrows, a general promiscuity may prevail.

One important aspect of courtship concerns the question of recognition. In gull colonies, for instance, members of the opposite sex look very much alike, and, at least to humans, the various individuals of one sex or the other may appear exactly the same. The advantages, with regard to successful production, incubation, and rearing of eggs and young, of permanent or semipermanent mate selection, however, are as great in gull colonies as elsewhere. The preliminaries to such a mutual selection not only establish a bond, by various posturings, but also establish the many small idiosyncrasies of action that add up to individuality and make one bird distinguishable among many within a colony, at least to its mate.

Many different forms of sex-oriented behaviour have consequently evolved among birds, depending on the character and particular needs of the various species. Penguins apparently not only look alike to human observers but also to themselves. Penguins seemingly have trouble even distinguishing between the sexes. Being unable to dance or sing, though they can make a lot of noise, male penguins can do little more than offer a pebble to a prospective female. If she accepts it as a token contribution to nest making, the match is on. If it is rejected, the suitor may have picked an unready female or even another male. In the case of most birds, however, the male can either sing, particularly the smaller kinds, or can strut and dance, with wings and feathers displayed, and some species, such as the lyre bird, continue to enchant the female by sight and sound together. In general, the need for physical mating has led to courtship and an emotional bonding between mating pairs throughout much of the animal kingdom at the higher level, particularly among birds and mammals. These are primarily utilitarian functions relating to the survival of the species, but in their fullest expression they represent what seem to man to be among the finest attributes of life.

SEX PATTERNS

Since the great value of sex as distinct from reproduction is the reassortment and recombination of genes every generation, sex cells from two separate parents ordinarily give rise to the greatest variaton, unless the parental individuals are themselves too closely related to each other. The presence of male and female individuals, respectively, generally produced in approximately equal numbers, is characteristic of so much of the animal kingdom that it appears to be the natural state. All that is certain, however, is that this condition has evolved as the most effective means to the particular end, and it may have done so independently among the various more or less unrelated groups of animals. The condition of separate sexes is not a universal fact, and two sexes within the same individual is typical of the more sluggish or actually attached kinds of animal life. Earthworms, slugs, land snails, flatworms, tapeworms, barnacles, sea squirts, and some others are all double-sexed individuals, or hermaphrodites. All have ovaries and testes producing mature eggs and sperm at the same time. Nevertheless, cross-fertilization is accomplished, and self-fertilization, even though possible, is generally avoided. Of those kinds of animal life mentioned above, all except the sea squirts have well-encased eggs that need to be fertilized before being laid. Mutual copulation, whereby each member of a mating pair of individuals introduces sperm into the body of the other member, is characteristic of these creatures, with the exception of the sea squirts.

Hermaphrodites

When animals shed sperm and comparatively naked eggs into the surrounding water, as is the case in sea squirts, self-fertilization is difficult to avoid. Most creatures have evolved an effective separation of the sexes between different individuals. Even so, there are more ways than one of accomplishing this. The common means is to produce male and female individuals that are constitutionally different, yet an equally effective procedure is for all individuals to be constitutionally the same but to become mature as male or female at different stages of the growth cycle. The oyster on its rock changes sex from male to female and back again once or twice a year. Certain shrimps also are hermaprodites. Each young shrimp of this kind grows up to be a male and is fully and functionally a male when about half the size of the females. As the next season approaches, his testes shrink, no more spermatozoa are produced, and ovaries begin to enlarge. As full growth is reached, the shrimp that had been a male becomes a typical female, ready to mate again, but this time with a young male of a newer generation. The system works as well as any other and clearly has its points. In fact the hagfish, not a true fish but a more primitive jawless vertebrate, also changes sex regularly, from year to year.

Sex differences in animals. In many animals, sexual differences are apparent in addition to the primary sex differentiation into males with testes and females with ovaries and apart from the accessory structures and tissues associated with the presence of one kind of sex gland or the other. Secondary sex differentiation in sexually distinct individuals is to be seen in many forms. In humans, for example, the beard and deep voice of the male and the enlarged breasts of the female are features of this sort. The great claw of the fiddler crab, the antlers of a moose, the great bulk and strength of a harem master in a fur seal colony, the beautiful fan tail of the peacock, and the bright feathers of other birds, are all distinctively male characteristics, and all are associated with the sexual drive of males. Females, by and large, are of comparatively quiet disposition and relatively drab appearance. Their function is to produce and nurture eggs, as safely and usually as inconspicuously as possible. The male function is to find and fertilize the female, for which both drive and display are generally required.

It is the business of sperm to be active and so find an egg. Similarly it is the busines of males to find a female and mate with her if possible. The male drive, or male eagerness, is a consequence of this special function of males. In nature, males possessing a strong eagerness to mate will find more females and leave more progeny than males lacking in sex drive. The progeny moreover will tend to inherit the drive of the parent. Males therefore are generally competitive with other males, with a premium placed on physical strength and sex drive and also on various devices for the attraction and stimulation of the female. The various exclusively male features already listed are all examples of characteristics of this sort, and they are related to the securing of female mates rather than the actual fertilization of eggs or to the problems of survival and adaptation.

Seasonal or periodic sexual cycles. In most animals sexual reproduction is seasonal or rhythmical, and so is sexual behaviour, whether in the form of courtship, drive, or other activities that lead to mating. In the marine fireworm of the West Indies, for instance, individuals of both sexes live in crevices on the sea floor but come out to breed where their fertilized eggs can drift and develop in the water above. But they can only find one another by means of the luminescence they themselves produce, which is an eerie light visible only in complete darkness. Each spring or summer month they emerge and swim to the surface about one-half hour after sunset when all daylight is gone but only before the moon can rise, a situation that confines them to a monthly breeding period of three or four days after the full of the moon. They follow a lunar rhythm. So do the grunion, a common fish along the southern California coast. Here again mating takes place when all is dark and the tide is high. Pairing occurs in the wash of the waves on the sand; fertilized eggs become immediately buried and there develop until the next high spring tides reach and wash the upper level sand nearly two weeks later. The mysterious biological clocks that apparently all living things possess adjust the rhythms of life to the needs of the particular organism. Some of these timing processes call internal signals on a regular day and night basis; others, on a somewhat longer cycle that keeps pace with the moon rather than the sun; and many, especially in the larger animals, run on a seasonal, or annual, cycle. Many activities are brought into line with the regular changes occurring in the environment. Sex and reproduction, however, are adjusted mainly with regard

Biological control of sexual cycles

to two functions; namely, safety while mating, which is therefore commonly in the dark, and the launching of the new generation at a time or season when circumstances are most favourable.

Birds lay eggs, and most mammals deliver their young in early spring, when the months ahead are warm and food is plentiful. Sex for the most part is adjusted to this end. Among the mammals, for example, the period of development within the womb varies greatly, from less than three weeks in the smallest to almost a year in the largest and certain others. Yet with few exceptions, the time for birth is in the spring. The time for mating in most cases is accordingly adjusted to this event: the larger the offspring at birth, the earlier the mating must take place. The horse and the great whales mate in spring and deliver in spring; roe deer mate in summer and deliver in spring; goat and sheep mate in the fall and deliver in spring. Even the elephant, which has a 22-month pregnancy, delivers in spring but must mate in early summer two years before. In small creatures, however, such as mice, rats, hamsters, and shrews, where the gestation, or pregnancy, period is about three weeks, reproduction is still seasonal, but there is time during the warmer months for several broods to be conceived and raised. In others, expediency may prevail, and mating may occur at a time to suit the convenience of the pairing animals. The little brown bat, for instance, mates in the fall, and yet ovulation does not take place until winter has passed; the spermatozoa survive the winter in the uterus and fertilize the eggs when they in turn arrive there five or six months later. In some other creatures mating occurs at a convenient time, eggs are fertilized, but development itself is suspended at an early stage for a time so that hatching or birthing, depending on the kind of animal, takes place when circumstances are suitable.

In all of this, the time of the mating season is clearly regulated, both with regard to the physiological condition of the animal and to the environmental conditions. The urge and capacity to mate depends on the ripeness of the gonads, male or female. In most animals, the reproductive glands wax and wane according to the seasons; that is, with an annual rhythm or else with a shorter cycle. Hormones are mainly in control of this rhythm. Sex hormones, male or female, respectively, are produced by the gonads themselves and cause or maintain their growth and at the same time cause the various secondary sexual characteristics of the male or female individual to become enhanced. Male hormone increases masculinity, even when injected into a female. Female canaries injected with male hormone no longer behave as females and shortly begin to sing loud and long and commence the courtship activities of a male. A hen thus injected grows a larger comb, starts to crow, and begins to strut.

Hormonal regulation of cycles

The production of these hormones is in turn controlled by hormones of the pituitary gland. Pituitary hormones stimulate ovarian or testicular tissue, which secretes the sex hormones. The sex hormones not only maintain the growth of the sexual tissues generally but inhibit the secretion of pituitary hormones, so that the process does not get out of hand. The pituitary activity, however, is also influenced by external conditions, particularly by stimuli received indirectly from light. The annual growth of ovaries or testes that occurs in late winter and early spring in frogs, reptiles, birds, and mammals is initiated by the steadily increasing period of daylight. In response to this changing day length, female frogs are packed with eggs and male frogs are ready to croak by the time the mating period arrives. The large eggs of reptiles and birds are ready to be fertilized, and the males are showing whatever they may have to display at the proper time. In mammals, the female comes into heat, the uterus undergoes the preparatory changes for taking care of fertilized eggs, and the male usually has but one thought in his mind. But as daylight ceases to lengthen, the sexual drive slowly diminishes.

SEX DETERMINATION

The determination of the sex of an individual, with regard to both the primary sex—*i.e.*, whether the ovaries or the testes develop—and the various secondary sexual characteristics may be rigorously controlled from the start of development or may be subject to later influences of a hormonal or environmental nature. However this may be, in order to appreciate the action of the control systems, the point of departure is that animals were primitively hermaphrodite, that during early stages of evolution every individual probably possessed both male and female gonads. Differentiation into separate sexes, each possessing male or female gonads but not both at the same time, is a device to ensure cross-fertilization of eggs, whether this is accomplished by having the two types of sexual gland mature at different stages of the growth of the individual, as in some shrimp and others, or whether by the production of two distinct types of individuals, as in most species of animals. This point of view is important because the question ceases to be how testes are caused to develop in the male organism and ovaries in the female but how, in a potentially double-sexed organism, the development of one or the other sex is suppressed. That such is the case is seen as clearly as anywhere in the human condition itself. Neither sex is completely male or female. Females have functional, well-developed mammary glands. Males also have mammary glands, undeveloped and nonfunctional although equipped with nipples. Males have a penis for delivering sperm, but females have a small, nonfunctional equivalent—the clitoris. These are secondary sexual features, to be sure, but the difference between the sexes is in the degree of their development, not a matter of absolute presence or absence.

Early development of reproductive systems in young animals

The basis for this is seen in the very beginnings of the development of the reproductive system, in frog, mouse, and man alike. In the young embryo a pair of gonads develop that are indifferent or neutral, showing no indication whether they are destined to develop into testes or ovaries. There are also two different duct systems, one of which can develop into the female system of oviducts and related apparatus and the other into the male sperm duct system. As development of the embryo proceeds, either the male or the female reproductive tissue differentiates in the originally neutral gonad of the mammal.

In the frog and other lower vertebrate animals, the picture is even clearer. The original gonad consists of an outer layer of cells and an inner core of cells. If the individual is to be a male, the central tissue grows at the expense of the outer layer. If it is to be a female, the outer tissue grows at the expense of the central core tissue. If both should grow, which is a possibility although a rare occurrence, the individual will be a hermaphrodite. Anything that influences the direction taken therefore may be said to determine sex.

Sex chromosomes. In most species of animals the sex of individuals is determined decisively at the time of fertilization of the egg, by means of chromosomal distribution. This process is the most clear-cut form of sex determination. When any cell in the body divides, except during the formation of the sex cells, each daughter cell receives the full complement of chromosomes; *i.e.*, copies of the two sets of chromosomes derived from the sperm cell and egg, respectively. The two sets are similar except for one pair of chromosomes. These are the so-called sex chromosomes, and the pair may be exactly alike or they may be obviously different, depending on the sex of the individual. The sex chromosomes are of two types, which are designated X and Y, and the pair of sex chromosomes may consist of two X chromosomes or of an X and Y paired together. In mammals (including man) and flies, the cells of males contain an XY pair and the cells of females contain an XX pair. On the other hand, in butterflies, fishes, and birds, the cells of females contain an XY pair and those of males contain an XX pair. In either case the Y chromosome is generally smaller than the X chromosome and may even be absent. What is most important concerning chromosomal sex determination is whether the cells of the individual contain one X chromosome or two X chromosomes. Human beings, for example, have cells with 22 pairs of nonsexual chromosomes, or autosomes, together with an XX pair or an XY pair. The female has a total of 46 functional chromosomes; the male has 45 plus a Y, which is mainly inert. Sex determination thus becomes a matter of balance. With one X chromosome plus the 44 autosomes in every cell, the whole course of development of primary and secondary sexual characteristics is toward the male; with two X chromosomes plus the autosomes in every cell, the whole system is swung over to the female.

The manipulation of this control system is readily accomplished during the special process of cell division that takes place in the gonads to produce sperm and eggs and their subsequent union at fertilization. In mammals, for example, since all cells in the female contain two X chromosomes, all the eggs will receive a single X chromosome when they are formed. All eggs are accordingly the same in this respect. In contrast, all cells in the male have the XY constitution, and therefore, when the double set of chromosomes is reduced to a single set during the formation of the spermatozoa, half of the spermatozoa will receive an X and half will receive a Y. Consequently, when an egg is fertilized by a sperm, the chances are about equal that the sperm will carry an X or will carry a Y, since the two types are inevitably produced in equal numbers. If it carries an X, the XX female constitution results; if a Y, then the XY male constitution results.

Abnormal chromosome effects. Occasionally, however, the processes of chromosomal reassortment and recombination occurring during sex cell formation and fertilization depart somewhat from the normal course. Sperm and eggs may be produced that are oversupplied or undersupplied with sex chromosomes. Fertilized eggs in humans may, for instance, have abnormal sex chromosome constitutions such as XXX, XXY, or XO. Those with the triple-X chromosome constitution have all the appearance of normal females and are called, in fact, superfemales, although only some will be fertile. Those with the XO (one X, but lacking Y altogether) constitution, a much more common condition, are also feminine in body form and type of reproduction system but remain immature. Individuals with the XXY constitution are outwardly males but have small testes and produce no spermatozoa. Those with the more abnormal and relatively rarer constitutions XXXXY and XXYY are typically mentally defective and in the latter case are hard to manage. Thus abnormal combinations generally result in an infertility on the one hand and an abnormal sexuality in the whole system, for either too little or too much of what is ordinarily good can be disastrous.

Very different kinds of abnormal development resulting from faulty chromosomal distribution are particularly observable in insects. The most common form in flies is an individual that is male on one side, female on the other, with a sharp line of demarcation. In other cases one-quarter of the body may be male and three-quarters female, or the head may be female and the rest of the body, male. These types are known as gynandromorphs, or sexual mosaics, and result from aberration in the distribution of the X chromosomes among the first cells to be formed during the early development of the embryo. This condition is unknown among higher animals.

Sexual mosaics

Parthenogenesis. The unfertilized, ripe egg possesses all the potentiality for full development. The process of fertilization by a spermatozoon introduces the nucleus of the male sex cell into the female egg, a process that increases the differences between parent and offspring and may determine the sex of the new individual and also stimulates the egg to begin development. These two functions are separate. Parthenogenetic development, without benefit of sperm, occurs naturally in various kinds of animals besides the waterflea (*Daphnia*), already described. Artificial, or experimental, parthenogenesis is readily brought about in many other species and by a variety of means. Mature, unfertilized eggs of starfish, sea urchins, various worms, and other marine invertebrate animals can be caused to develop by treatment with a weak organic acid. Unfertilized frog eggs can be readily caused to develop by gentle pricking of the egg surface with the tip of a fine glass needle that has been dipped in lymph. In nature the eggs of various creatures can develop with or without the aid of spermatozoa. The sex of parthenogenetically developed individuals, insofar as it depends on the chromosomal constitution of the developing egg, is consequently affected. Frog eggs developing parthenogenetically become males, since only one X chromosome is present in each cell. In nature, where varying conditions call for various responses, the system is usually more complicated, although based on the general relationship that individuals with the XX constitution will be female and those with a single X will be males. A queen honeybee, for instance, begins her reproductive life with a store of sperm received from a male during her nuptial flight. Throughout spring and summer almost all eggs become fertilized and develop into females (either as nonfertile female workers or as new fertile queens, depending on the nature of food received during growth). Toward the end of summer, when the sperm supply runs low, eggs cease to be fertilized and, when laid, develop into drones, ready to mate with a new queen should occasion arise. In other cases, even parthenogenetically developing eggs may become female individuals through a process of chromosome doubling, which takes place in the mature but unfertilized eggs. Thus certain wasps, waterfleas, and others are able to produce many exclusively female generations in succession.

Effects of environment. Sex chromosomes, however, do not determine sex directly but do so through their control of such cell activities as metabolism and hormone production. Their determinative influence, indirect though it is, may be complete. On the other hand, environmental conditions may play the dominating role. In the case of *Bonellia,* a unique kind of marine worm, all eggs develop into small larvae of a sexually indifferent kind. Those that settle freely on the sea floor grow into comparatively large females, each of which has a long, broad extension, the proboscis, at its front end. Those larvae that happen to settle on the proboscis of a female, however, fail to grow beyond a certain minute size and become dwarf males, permanently attached to the female body. The sex-determining factor appears to be the environmental carbon dioxide tension, which is relatively high at the surface of living tissue.

Hormones. Because in most developing animals the reproductive gland is essentially neutral to begin with, there is generally some possibility that agents external to the gland, particularly chemical agents—*i.e.*, hormones—circulating in the blood system, may override the sex-determining influence of the sex chromosomes. In the chick, for example, the sex can be controlled experimentally by such means until about four hours after hatching. If a female chick is injected on hatching with the male sex hormone, testosterone, it will develop into a fully functional cock. Even when injected at later stages of growth, the male hormone causes extra early growth of the comb, crowing, and aggressive behaviour after being injected in either male or female chicks. Female sex hormones, such as estrogen, on the other hand, stimulate early growth of the oviduct in the female and feminize the plumage and suppress comb growth when injected in the male.

This susceptibility of the reproductive glands, and sexuality in general, to the influence of sex hormones is particularly acute in mammals, where the egg and embryo, unprotected by any shell, develop in the uterus exposed to various chemicals filtering through from the maternal blood stream. A developing embryo eventually produces its own sex hormones, but they are not manufactured in any quantity until the anatomical sex of the embryo is already well established. One of the curious things about sex hormones, however, is that the reproductive glands are not the only tissues that produce them. The placenta, through which all exchange between fetus and mother takes place, itself produces tremendous amounts of female sex hormone, together with some male hormone, which are excreted by the mother during pregnancy. This condition is true of humans, as well as of mice and rats. As a rule these hormones are produced too late to do any harm, but not always. The female embryo is fairly immune inasmuch as additional female hormone merely causes a child to be more feminine than usual at an early age. Male embryos, however, may be seriously affected if the female hormone catches them at an early stage. Boy babies may be born that are truly males but under the impact of the feminizing hormone appear superficially to be

Production and effects of sex hormones in mammals females and are often raised as such. As a rule, even when older, they have more or less sterile, undescended testes; an imperfect penis; well-developed breasts; an unbroken voice; and no beard. One in a thousand may be like this and on occasion may have won in women's Olympic competitions. In other cases, those somewhat less severely affected, during adolescence when the hidden testes begin to secrete their own male hormones in abundance, the falsely female characteristics become suppressed, and the voice, beard, breasts, and sexual interest take on the pattern of the male. What were thought to be girls in their youth change into the men they were meant to be upon reaching maturity. (N.J.B.)

Human beings

Human sexual behaviour may be defined as any activity—solitary, between two persons, or in a group—that induces sexual arousal. There are two major determinants of human sexual behaviour: the inherited sexual response patterns that have evolved as a means of ensuring reproduction and that are a part of each individual's genetic inheritance, and the degree of restraint or other types of influence exerted on the individual by society in the expression of his sexuality. The objective here is to describe and explain both sets of factors and their interaction.

It should be noted that taboos in Western culture and the immaturity of the social sciences for a long time impeded research concerning human sexual behaviour, so that by the early 20th century scientific knowledge was largely restricted to individual case histories that had been studied by such European writers as Sigmund Freud, Havelock Ellis, and Richard, Freiherr von Krafft-Ebing. By the 1920s, however, the foundations had been laid for the more extensive statistical studies that were conducted before World War II in the United States. Of the two major organizations for sex study, one, the Institut für Sexualwissenschaft in Berlin (established in 1897), was destroyed by the Nazis in 1933. The other, the Institute for Sex Research, begun in 1938 by the American sexologist Alfred Charles Kinsey at Indiana University, Bloomington, undertook the study of many aspects of human sexual behaviour. Much of the following discussion rests on the findings of the Institute for Sex Research, which comprise the most comprehensive data available. The only other country for which comprehensive data exist is Sweden.

TYPES OF BEHAVIOUR

Human sexual behaviour may conveniently be classified according to the number and gender of the participants. There is solitary behaviour involving only one individual, and there is sociosexual behaviour involving more than one person. Sociosexual behaviour is generally divided into heterosexual behaviour (male with female) and homosexual behaviour (male with male or female with female; see *Homosexuality* below.) If three or more individuals are involved it is, of course, possible to have heterosexual and homosexual activity simultaneously.

In both solitary and sociosexual behaviour there may be activities that are sufficiently unusual to warrant the label deviant behaviour. The term deviant should not be used as a moral judgment but simply as indicating that such activity is not common in a particular society. Since human societies differ in their sexual practices, what is deviant in one society may be normal in another.

Solitary behaviour. Self-masturbation is self-stimulation with the intention of causing sexual arousal and, generally, orgasm (sexual climax). Most masturbation is done in private as an end in itself but is sometimes practiced to facilitate a sociosexual relationship.

Masturbation, generally beginning at or before puberty, is extremely common in males, particularly among young males, but becomes less frequent or is even abandoned when sociosexual activity is available. Consequently, masturbation is most frequent among the unmarried. Fewer females masturbate; in the United States, roughly one-half to two-thirds have done so, as compared to nine out of ten males. Females also tend to reduce or discontinue masturbation when they develop sociosexual relationships. There is great individual variation in frequency, so that it is impractical to try to define what range could be considered "normal."

Attitudes toward masturbation

The myth persists, despite scientific proof to the contrary, that masturbation is physically harmful. Neither is there evidence that masturbation is immature behaviour; it is common among adults deprived of sociosexual opportunities. While solitary masturbation does provide pleasure and relief from the tension of sexual excitement, it does not have the same psychological gratification that interaction with another person provides; thus, extremely few people prefer masturbation to sociosexual activity. The psychological significance of masturbation lies in how the individual regards it. For some, it is laden with guilt; for others, it is a release from tension with no emotional content; and for others it is simply another source of pleasure to be enjoyed for its own sake.

The majority of males and females have fantasies of some sociosexual activity while they masturbate. The fantasy not infrequently involves idealized sexual partners and activities that the individual has not experienced and even might avoid in real life.

Since the masturbating person is in sole control of the areas that are stimulated, the degree of pressure, and the rapidity of movement, masturbation is often more effective in producing sexual arousal and orgasm than is sociosexual activity, during which the stimulation is determined to some degree by one's partner.

Orgasm in sleep evidently occurs only in humans. Its causes are not wholly known. The idea that it results from the pressure of accumulated semen is invalid because not only do nocturnal emissions sometimes occur in males on successive nights, but females experience orgasm in sleep as well. In some cases orgasm in sleep seems a compensatory phenomenon, occurring during times when the individual has been deprived of or abstains from other sexual activity. In other cases it may result from external stimuli, such as sleeping prone or having night clothing caught between one's legs. Most orgasms during sleep are accompanied by erotic dreams.

A great majority of males experience orgasm in sleep. This almost always begins and is most frequent in adolescence, tending to disappear later in life. Fewer females have orgasm in sleep, and, unlike males, they usually begin having such experience when fully adult.

Orgasm in sleep is generally infrequent, seldom exceeding a dozen times per year for males and three or four times a year for the average female.

Most sexual arousal does not lead to sexual activity with another individual. Humans are constantly exposed to sexual stimuli when seeing attractive persons and are subjected to sexual themes in advertising and the mass media. Response to such visual and other stimuli is strongest in adolescence and early adult life and usually gradually declines with advancing age. One of the necessary tasks of growing up is learning to cope with one's sexual arousal and to achieve some balance between suppression, which can be injurious, and free expression, which can lead to social difficulties. There is great variation among individuals in the strength of sex drive and responsiveness, so this necessary exercise of restraint is correspondingly difficult or easy.

Sociosexual behaviour. By far the greatest amount of sociosexual behaviour is heterosexual behaviour between only one male and one female. Heterosexual behaviour frequently begins in childhood, and, while much of it may be motivated by curiosity, such as showing or examining genitalia, many children engage in sex play because it is pleasurable. The sexual impulse and responsiveness are present in varying degrees in most children and latent in the remainder. With adolescence, sex play is superseded by dating, which is socially encouraged, and dating almost inevitably involves some physical contact resulting in sexual arousal. This contact, labelled necking or petting, is a part of the learning process and ultimately of courtship and the selection of a marriage partner.

Petting and foreplay

Petting varies from hugging, kissing, and generalized caresses of the clothed body to techniques involving genital stimulation. Petting may be done for its own sake as an expression of affection and a source of pleasure, and it may occur as a preliminary to coitus. This last form of petting is known as foreplay. In a minority of cases, but a substantial minority, petting leads to orgasm and may be a substitute for coitus. Excluding foreplay, petting is usually very stereotyped, beginning with hugging and kissing and gradually escalating to stimulation of the breasts and genitalia. In most societies petting and its escalation are initiated by the male more often than by the female, who generally rejects or accepts the male's overtures but refrains from playing a more aggressive role. Petting in some form is a near-universal human experience and is valuable not only in mate selection but as a means of learning how to interact with another person sexually.

Coitus, the insertion of the penis into the vagina, is viewed by society quite differently depending upon the marital status of the individuals. The majority of human societies permit premarital coitus, at least under certain circumstances. In more repressive societies, such as modern Western society, it is more likely to be tolerated (but not encouraged) if the individuals intend marriage. Marital coitus is usually regarded as an obligation in most societies. Extramarital coitus, particularly by wives, is generally condemned and, if permitted, is allowed only under exceptional conditions or with specified persons. Societies tend to be more lenient toward males than females regarding extramarital coitus. This double standard of morality is also seen in premarital life. Postmarital coitus (*i.e.*, coitus by separated, divorced or widowed persons) is almost always ignored. Even societies that try to confine coitus to marriage recognize the difficulty of trying to force abstinence upon sexually experienced and usually older persons.

In the United States and much of Europe, there has been, within the last century, a progressive trend toward an increase in premarital coitus. Currently in the United States, at least three-quarters of the males and over half of the females have experienced premarital coitus. The proportions for this experience vary in different groups and socioeconomic classes. In Scandinavia, the incidence of premarital coitus is far greater, exceeding the 90 percent mark in Sweden, where it is now expected behaviour.

Extramarital coitus continues to be openly condemned but is becoming more tolerated secretly, particularly if mitigating circumstances are involved. In some areas, such as southern Europe and Latin America, extramarital coitus is expected of most husbands and is accepted by society if the behaviour is not too flagrant. The wives do not generally approve but are resigned to what they believe to be a masculine propensity. In the United States, where at least half the husbands and one-quarter of the wives have extramarital coitus at some point in their lives, there have recently developed small organizations or clubs that exist to provide extramarital coitus for married couples. Despite the publicity they have engendered, however, extremely few individuals have belonged to such organizations. Most extramarital coitus is done secretly without the knowledge of the spouse. Most husbands and wives feel very possessive of their spouses and interpret extramarital activity as an aspersion on their own sexual adequacy, as indicating a loss of affection and as being a source of social disgrace.

Human beings are not inherently monogamous but have a natural desire for diversity in their sexuality as in other aspects of life. Some societies have provided a release for these desires by suspending the restraints on extramarital coitus on special occasions or with certain individuals, and in modern Western society a certain amount of extramarital flirtation or mild petting at parties is not considered unusual behaviour.

Sexuality in ceremony and religion

Discussion of sociosexual behaviour would be incomplete without some note of the role it has played in ceremony and religion. While the major religions of today are to varying degrees antisexual, many religions have incorporated sexual behaviour into their rites and ceremonies. The ancient and continuing interest of man in the fertility of food plants, animals, and himself makes such a connection between sex and religion inevitable, particularly among peoples with uncertain food supplies. In most religions the deities were considered to have active sexual lives and sometimes took a sexual interest in humans. In this regard it is noteworthy that in Christianity sexual behaviour is absent in heaven and sexual proclivities are ascribed only to evil supernatural beings: Satan, devils, incubi, and succubi (spirits or demons who seek out sleeping humans for sexual intercourse).

Whether or not a behaviour is interpreted by society or the individual as erotic (*i.e.*, capable of engendering sexual response) depends chiefly on the context in which the behaviour occurs. A kiss, for example, may express asexual affection (a mother kissing a child or a kiss between relatives), respect (as a French officer kissing a soldier after bestowing a medal on him), reverence (kissing the hand or foot of a pope), or it may be a casual salutation and social amenity. Even something as specific as touching genitalia is not construed as sexual if done for medical reasons. In other words, the apparent motivation of the behaviour determines its interpretation.

Individuals are extremely sensitive in judging motivations: a greeting kiss, if protracted more than a second or two, takes on a sexual connotation, and recent studies show that if an adult male at a party stands closer than the length of his hand and forearm to a female, she generally imputes a sexual motive to his proximity. Nudity is construed as erotic or even as a sexual invitation—unless it occurs in a medical context, in a group consisting of but one gender, or in a nudist camp.

PHYSIOLOGICAL ASPECTS

Sexual response. Sexual response follows a pattern of sequential stages or phases when sexual activity is continued. First, there is the excitement phase marked by increase in pulse and blood pressure, an increase in blood supply to the surface of the body resulting in increased skin temperature, flushing, and swelling of all distensible body parts (particularly noticeable in the penis and female breasts), more rapid breathing, the secretion of genital fluids, vaginal expansion, and a general increase in muscle tension. These symptoms of arousal eventually increase to a near maximal physiological level, the plateau phase, which is generally of brief duration. If stimulation is continued, orgasm usually occurs. Orgasm is marked by a feeling of sudden intense pleasure, an abrupt increase in pulse rate and blood pressure, and spasms of the pelvic muscles causing vaginal contractions in the female and ejaculation by the male. Involuntary vocalization may also occur. Orgasm lasts for a few seconds (normally not over ten), after which the individual enters the resolution phase, the return to a normal or subnormal physiological state. Up to the resolution phase, males and females are the same in their response sequence, but, whereas males return to normal even if stimluation continues, continued stimulation can produce additional orgasms in females. In brief, after one orgasm a male becomes unresponsive to sexual stimulation and cannot begin to build up another excitement phase until some period of time has elapsed, but females are physically capable of repeated orgasms without the intervening "rest period" required by males.

Genetic and hormonal factors. While all normal individuals are born with the neurophysiology necessary for the sexual-response cycle described above, inheritance determines the intensity of their responses and their basic "sex drive." There is great variation in this regard: some persons have the need for frequent sexual expressions; others require very little; and some persons respond quickly and violently, while others are slower and milder in their reactions. While the genetic basis of these differences is unknown and while such variations are obscured by conditioning, there is no doubt that sexual capacities, like all other physiological capacities, are genetically determined. It is unlikely, however, that genes control the sexual orientation of normal humans in the sense of individuals being predestined to become homosexual or heterosexual. Some severe genetic abnormality can, of course, profoundly affect intelligence, sexual capacity, and physical appearance and hence the entire sexual life.

While the normal female has 44 autosomes plus two X-chromosomes (female) and the normal male 44 autosomes plus one X-chromosome and one Y-chromosome

Genetic combinations in the sexes

(male), many genetic abnormalities are possible. There are females, for example, with too many X-chromosomes (44+XXX) or too few (44+X) and males with an extra female chromosome (44+XXY) or an extra male chromosome (44+XYY). No 44+YY males exist—an X-chromosome is necessary for survival, even in the womb.

One's genetic makeup determines one's hormonal status and the sensitivity of one's body to these hormones. While a disorder of any part of the endocrine system can adversely affect sexual life, the hormones most directly influencing sexuality are the androgens (male sex hormones), produced chiefly in the testicles, and the estrogens (female sex hormones), produced chiefly in the ovaries. In early embryonic life there are neither testicles nor ovaries but simply two undifferentiated organs (gonads) that can develop either into testicles or ovaries. If the embryo has a Y-chromosome, the gonads become testicles; otherwise, they become ovaries. The testicles of the fetus produce androgens, and these cause the fetus to develop male anatomy. The absence of testicles results in the development of female anatomy. Animal experiments show that, if the testicles of a male fetus are removed, the individual will develop into what seems a female (although lacking ovaries). Consequently, it has been said that humans are basically female.

After birth and until puberty, the ovaries and testicles produce comparatively few hormones, and little girls and boys are much alike in size and appearance. At puberty, however, these organs begin producing in greater abundance, with dramatic results. The androgens produced by boys cause changes in body build, greater muscular development, body and facial hair, and voice change. In girls the estrogens cause breast development, menstruation, and feminine body build. A boy castrated before puberty does not develop masculine physical characteristics and manifests in adult life more of a feminine body build, lack of masculine body and facial hair, less muscular strength, a high voice, and small genitalia. A girl who has her ovaries removed before puberty is less markedly altered but retains a childlike body build, does not develop breasts, and never menstruates. Castrated individuals or persons producing insufficient hormones can be restored to a normal condition by administration of appropriate hormones.

Beyond their role in developing the secondary sexual characteristics of the body, the hormones continue to play a role in adult life. An androgen deficiency causes a decrease in a man's sexual responsiveness, and an estrogen deficiency adversely affects a woman's fertility and causes atrophy of the genitalia. A loss of energy may also result in both men and women.

Role of androgen and estrogen

Androgen seems linked in both males and females with aggressiveness and strength of sexual drive. When androgen is given to a female in animal experiments, she becomes more aggressive and displays behaviour more typical of males—by mounting other animals, for example. Estrogen increases her sexual responsiveness and intensifies her female behaviour. Androgen given to a male often increases his sexual behaviour, but estrogen diminishes his sex drive.

In humans the picture is more complex, since human sexual behaviour and response is less dependent on hormones once adulthood has been reached. Removing androgen from an adult male reduces his sexual capacity; but this occurs gradually, and sometimes the reduction is small. Giving androgen to a normal human male generally has little or no effect since he is already producing all he can use. Giving him estrogen reduces his sex drive. Administration of androgen to an adult human female often increases her sex drive, enlarges her clitoris, and promotes the growth of facial hair. Giving estrogen to a normal woman before menopausal age generally has no effect whatsoever—probably because human females, unlike other female mammals, do not have hormonally controlled periods of "heat" (estrus).

Hormones have no connection with the sexual orientation of humans. Male homosexuals do not have more estrogens than normal males (who have a little) nor can their preferences be altered by giving them androgen.

Nervous system factors. The nervous system consists of the central nervous system and the peripheral nervous system. The brain and spinal cord constitute the central system, while the peripheral system is composed of (1) the cerebrospinal nerves that go to the spinal cord (afferent nerves), transmitting sensory stimuli and those that come from the cord (efferent nerves) transmitting impulses to activate muscles, and (2) the autonomic system, the primary function of which is the regulation and maintenance of the body processes necessary to life, such as heart rate, breathing, digestion, and temperature control. Sexual response involves the entire nervous system. The autonomic system controls the involuntary responses; the afferent cerebrospinal nerves carry the sensory messages to the brain; the efferent cerebrospinal nerves carry commands from the brain to the muscles; and the spinal cord serves as a great transmission cable. The brain itself is the coordinating and controlling centre, interpreting what sensations are to be perceived as sexual and issuing appropriate "orders" to the rest of the nervous system.

The parts of the brain thought to be most concerned with sexual response are the hypothalamus and the limbic system, but no specialized "sex centre" has been located in the human brain. Animal experiments indicate that each individual has coded in its brain two sexual response patterns, one for mounting (masculine) behaviour and one for mounted (feminine) behaviour. The mounting pattern can be elicited or intensified by male sex hormone and the mounted pattern by female sex hormone. Normally, one response pattern is dominant and the other latent but capable of being called into action when suitable circumstances occur. The degree to which such inherent patterning exists in humans is unknown.

While the brain is normally in charge, there is some reflex (*i.e.*, not brain-controlled) sexual response. Stimulation of the genital and perineal area can cause the "genital reflex": erection and ejaculation in the male, vaginal changes and lubrication in the female. This reflex is mediated by the lower spinal cord, and the brain need not be involved. Of course, the brain can override and suppress such reflex activity—as it does when an individual decides that a sexual response is socially inappropriate.

Development and change in the reproductive system. One's anatomy and sexuality change with age. The changes are rapid in intra-uterine life and around puberty but are much slower and gradual in other phases of the life cycle.

Intra-uterine sexual development

The reproductive organs first develop in the same form for both males and females: internally there are two undifferentiated gonads and two pairs of parallel ducts (Wolffian and Müllerian ducts); externally there is a genital protrusion with a groove (urethral groove) below it, the groove being flanked by two folds (urethral folds). On either side of the genital protrusion and groove are two ridgelike swellings (labioscrotal swellings). Around the fourth week of life the gonads differentiate into either testes or ovaries. If testes develop, the hormone they secrete causes the Müllerian duct to degenerate and almost vanish and causes the Wolffian duct to elaborate into the sperm-carrying tubes and related organs (the vas deferens, epididymis, and seminal vesicles, for example). If ovaries develop, the Wolffian duct deteriorates, and the Müllerian duct elaborates to form the fallopian tubes, uterus, and part of the vagina. The external genitalia simultaneously change. The genital protrusion becomes either a penis or clitoris. In the female the groove below the clitoris stays open to form the vulva, and the folds on either side of the groove become the inner lips of the vulva (the labia minora). In the male these folds grow together, converting the groove into the urethral tube of the penis. The ridgelike swellings on either side remain apart in the female and constitute the large labia (labia majora), but in the male they grow together to form the scrotal sac into which the testes subsequently descend.

At the time of birth both male and female have all the neurophysiological equipment necessary for sexual response, although the reproductive system is not at this stage functional. Sexual interests, sexual behaviour, and sexual response are seen with increasing frequency in most children from infancy on. Even newborn males have penile erections, and babies of both sexes seem to find some pleasure in genital stimulation. What appears to be orgasm has been observed in infant boys and girls, and, later in childhood, orgasm definitely can occur in self-masturbation or sex play.

Puberty may be defined as that short period of time (generally two years) during which the reproductive system matures and the secondary sexual characteristics appear. The ovaries and testes begin producing much larger amounts of hormones, pubic hair appears, female breasts develop, the menstrual cycle begins in females, spermatozoa and viable eggs are produced, and males experience voice change and a sudden acceleration in growth. Puberty generally occurs in females around age 12–13 and in males at about 13–14, but there is much individual variation. With puberty there is generally an intensification or the first appearance of sexual interest. Puberty marks the beginning of adolescence.

Adolescence, from a physical viewpoint, is that period between puberty and the attainment of one's maximum height. By the latter point, which occurs around age 16 in females and 18 in males, the individual has adult anatomy and physiology. In late adolescence the majority of individuals are probably at their peak in terms of sexual capacity: the ability to respond quickly and repeatedly. During this period the sex drive is at its maximum in males, although it is difficult to say whether this is also true of females, since female sexuality, in many societies, is frequently suppressed during adolescence.

Following adolescence there are about three decades of adult life during which physiological changes are slow and gradual. While muscular strength increases for a time, the changes may best be described as slow deterioration. This physical decline is not immediately evident in sexual behaviour, which often increases in quantity and quality as the individual develops more social skills and higher socio-economic status and loses some of the inhibitions and uncertainties that often impede adolescent sexuality. Indeed, in the case of the United States female, the deterioration is more than offset by her gradual loss of sexual inhibition, and the effect of age is not clear until menopausal symptoms begin. In the male, however, there is no such masking of deterioration, and the frequency of sexual activity and the intensity of interest and response slowly, but inexorably, decline.

Sexual changes in later life

If one must arbitrarily select an age to mark the beginning of old age, 50 is appropriate. By then, most females have experienced menopausal symptoms, and most males have been forced to recognize their increasing physical limitations. With menopause, the female genitalia gradually begin to atrophy and the amount of vaginal secretion diminishes—this is the direct consequence of the cessation of ovarian function and can be prevented, or the symptoms reversed, by administering estrogen. If a female has had a good sexual adjustment prior to menopause and if she does not believe in the fallacy that it spells the end of sexual life, menopause will have no adverse effect on her sexual and orgasmic ability. There is reason to believe that if a woman remains in good health and genital atrophy is prevented, she could enjoy sexual activity regardless of age. Males in good health are also capable of continuing sexual activity, although with an ever-decreasing frequency, throughout old age. The male has more difficulty in achieving erection, cannot maintain erection as long, and must have longer and longer "rest periods" between sexual acts. The amount of ejaculate becomes less, but most old males are still fertile. The Cowper's gland secretion (called "precoital mucus") diminishes or disappears entirely. According to Kinsey's data, about one-quarter of males are impotent by age 65, one-half by age 75, and three-quarters by age 80. One must remember, however, that some unknown but certainly substantial proportion of this impotence may be attributed to poor health.

In general, the female withstands the onslaughts of age better than the male. The reduction in the frequency of marital intercourse or even its abandonment is more often than not the result of male deterioration.

PSYCHOLOGICAL ASPECTS

Effects of early conditioning. Physiology sets only very broad limits on human sexuality; most of the enormous variation found among humans must be attributed to the psychological factors of learning and conditioning.

The human infant is born simply with the ability to respond sexually to tactile stimulation. It is only later and gradually that the individual learns or is conditioned to respond to other stimuli, to develop a sexual attraction to males or females or both, to interpret some stimuli as sexual and others as nonsexual, and to control in some measure his or her sexual response. In other words, the general and diffuse sexuality of the infant becomes increasingly elaborated, differentiated, and specific.

The early years of life are, therefore, of paramount importance in the development of what ultimately becomes adult sexual orientation. There appears to be a reasonably fixed sequence of development. Before age five, the child develops a sense of gender identity, thinks of himself or herself as a boy or girl, and begins to relate to others differently according to their gender. Through experience the child learns what behaviour is rewarded and what is punished and what sorts of behaviour are expected of him or her. Parents, peers, and society in general teach and condition the child about sex not so much by direct informational statements and admonitions as by indirect and often unconscious communication. The child soon learns, for example, that he can touch any part of his body or someone else's body except the anal–genital region. The child rubbing its genitals finds that this quickly attracts adult attention and admonishment or that adults will divert him or her from this activity. It becomes clear that there is something peculiar and taboo about this area of the body. This "genital taboo" is reinforced by the great concern over the child's excretory behaviour: bladder and bowel control is praised; loss of control is met by disappointment, chiding, and expressions of disgust. Obviously, the anal–genital area is not only a taboo area but a very important one as well. It is almost inevitable that the genitalia become associated with anxiety and shame. It is noteworthy that this attitude finds expression in the language of Western civilizations, as in "privates" (something to be kept hidden) and the German word for the genitals, *Scham* ("shame").

Development of genital taboo

While all children in Western civilizations experience this antisexual teaching and conditioning, a few have, in addition, atypical sexual experiences, such as witnessing or hearing sexual intercourse or having sexual contact with an older person. The effects of such atypical experiences depend upon how the child interprets them and upon the reaction of adults if the experience comes to their attention. Seeing parental coitus is harmless if the child interprets it as playful wrestling but harmful if he considers it as hostile, assaultive behaviour. Similarly, an experience with an adult may seem merely a curious and pointless game, or it may be a hideous trauma leaving lifelong psychic scars. In many cases the reaction of parents and society determines the child's interpretation of the event. What would have been a trivial and soon-forgotten act becomes traumatic if the mother cries, the father rages, and the police interrogate the child.

Some atypical developments occur through association during the formative years. A child may associate clothing, especially underclothing, stockings, and shoes with gender and sex and thereby establish the basis for later fetishism or transvestism. Others, having been spanked or otherwise punished for self-masturbation or childhood sex play, form an association between punishment, pain, and sex that could escalate later into sadism or masochism. It is not known why some children form such associations whereas others with apparently similar experience do not.

Around the age of puberty, parents and society, who more often than not refuse to recognize that children have sexual responses and capabilities, finally face the inescapable reality and consequently begin inculcating children with their attitudes and standards regarding sex. This campaign by adults is almost wholly negative—the child is told what not to do. While dating may be encouraged, no form of sexual activity is advocated or held up as model behaviour. The message usually is "be popular" (*i.e.*, sexually attractive), but abstain from sexual activity. This antisexualism is particularly intense regarding young females and is reinforced by reference to pregnancy, venereal disease, and, most importantly, social disgrace. To this list religious families add the concept of the sinfulness of premarital sexual expression. With young males the double standard of morality still prevails. The youth receives a double message, "don't do it, but we expect that you will." No such loophole in the prohibitions is offered young girls. Meanwhile, the young male's peer group is exerting a prosexual influence, and his social status is enhanced by his sexual exploits or by exaggerated reports thereof.

Effects of the double standard on the sexual relations of the young

As a result of this double standard of sexual morality, the relationship between young males and females often becomes a ritualized contest, the male attempting to escalate the sexual activity and the female resisting his efforts. Instead of mutuality and respect, one often has a struggle in which the female is viewed as a reluctant sexual object to be exploited, and the male is viewed as a seducer and aggressor who must succeed in order to maintain his self-image and his status with his peers. This sort of pathological relationship causes a lasting attitude on the part of females: men are not to be trusted; they are interested only in sex; a girl dare not smile or be friendly lest males interpret it as a sign of sexual availability, and so forth. Such an aura of suspicion, hostility, and anxiety is scarcely conducive to the development of warm, trusting relationships between males and females. Fortunately, love or infatuation usually overcomes this negativism with regard to particular males, but the average female still maintains a defensive and skeptical attitude toward men.

Western society is replete with attitudes that impede the development of a healthy attitude toward sex. The free abandon so necessary to a full sexual relationship is, in the eyes of many, an unseemly loss of self-control, and self-control is something one is urged to maintain from infancy onward. Panting, sweating, and involuntary vocalization are incompatible with the image of dignity. Worse yet is any substance once it has left the body: it immediately becomes unclean. The male and female genital fluids are generally regarded with disgust—they are not only excretions but sexual excretions. Here again, societal concern over excretion is involved, for sexual organs are also urinary passages and are in close proximity to the "dirtiest" of all places—the anus. Lastly, many individuals in society regard menstrual fluid with disgust and abstain from sexual intercourse during the four to six days of flow. This attitude is formalized in Judaism, in which menstruating females are specifically labelled as ritually unclean.

In view of all of these factors working against a healthy, rational attitude toward sex and in view of the inevitable disappointments, exploitations, and rejections that are involved in human relationships, one might wonder how anyone could reach adulthood without being seriously maladjusted. The sexual impulse, however, is sufficiently strong and persistent and repeated sexual activity gradually erodes the inhibitions and any sense of guilt or shame. Further, all humans have a deep need to be esteemed, wanted, and loved. Sexual activity with another is seen as proof that one is attractive, desired, valued, and possibly loved—a proof very necessary to self-esteem and happiness. Hence, even among the very inhibited or those with weak sex drive, there is this powerful motivation to engage in sociosexual activity.

Most persons ultimately achieve at least a tolerable sexual adjustment. Some unfortunates, nevertheless, remain permanently handicapped, and very few completely escape the effects of society's antisexual conditioning. While certain inhibitions and restraints are socially and psychologically useful—such as deferring gratification until circumstances are appropriate and modifying behaviour out of regard for the feelings of others—most people labour under an additional burden of useless and deleterious attitudes and restrictions.

Sexual problems. Sexual problems may be classified as physiological, psychological, and social in origin. Any given problem may involve all three categories; a physiological problem, for example, will produce psychological effects, and these may result in some social maladjustment.

Physiological problems of a specifically sexual nature are rather few. Only a small minority of people suffer from diseases of or deficient development of the genitalia or that part of the neurophysiology governing sexual response. A large number of persons, however, experience, at some point in life, sexual problems that are byproducts of other pathologies or injuries. Vaginal infections, for example, retroverted uteri, prostatitis, adrenal tumours, diabetes, senile changes of the vagina, and cardiovascular conditions may cause disturbance of the sexual life. In brief, anything that seriously interferes with normal bodily functioning generally causes some degree of sexual trouble. Fortunately, the great majority of physiological sexual problems are solved through medication or surgery. Generally, only those problems involving damage to the nervous system defy therapy.

Psychological problems constitute by far the largest category. They are not only the product of socially induced inhibitions, maladaptive attitudes, and ignorance but also of sexual myths held by society. An example of the latter is the idea that good, mature sex must involve rapid erection, protracted coitus, and simultaneous orgasm. Magazines, marriage books, and general sexual folklore reinforce these demanding ideals, which cannot always be met and hence give rise to anxiety, guilt, and feelings of inadequacy.

Premature ejaculation and its therapy

Premature ejaculation is a common problem, especially for young males. Sometimes this is not the consequence of any psychological problem but the natural result of excessive tension in a male who has been sexually deprived. In such cases, more frequent coitus solves the problem. Premature ejaculation is difficult to define. The best definition is that offered by the American sexologists, William Howell Masters and Virginia Eshelman Johnson, who say that a male suffers from premature ejaculation if he cannot delay ejaculation long enough to induce orgasm in a sexually normal female at least half the time. This generally means that vaginal penetration with some movement (although not continuous) must be maintained for more than one minute. The average United States male ejaculates in two or three minutes after vaginal penetration, a coital duration sufficient to cause orgasm in most females the majority of the time. Various methods of preventing premature ejaculation have been tried. One is for the male to excite the female more during the foreplay so that she reaches orgasm more rapidly after penetration, but this technique often excites the male as well and defeats its purpose. Another common method is for the male to think of nonsexual matters, which may prove effective but reduces his pleasure. The most effective therapy is that advocated by Masters and Johnson in which the female brings the male nearly to orgasm and then prevents the male's orgasm by briefly compressing the penis between her fingers just below the head of the penis. The couple come to realize that premature ejaculation can thus be easily prevented, their anxiety disappears, and ultimately they can achieve normal coitus without resorting to this squeeze technique.

Erectile impotence is almost always of psychological origin in males under 40; in older males physical causes are more often involved. Fear of being impotent frequently causes impotence, and, in many cases, the afflicted male is simply caught up in a self-perpetuating problem that can only be solved by achieving a successful act of coitus. In other cases, the impotence may be the result of disinterest in the sexual partner, fatigue, distraction because of nonsexual worries, intoxication, or other causes—such occasional impotency is common and requires no therapy. Some males, however, are chronically impotent and require psychotherapy or behaviour therapy. Such impotency is thought to be the result of deep-seated causal factors such as unconscious feelings of hostility, fear, inadequacy, or guilt. Primary impotence, the inability to ever have achieved erection sufficient for coitus, is more difficult to treat than the far more common secondary impotence, which is impotence in a male who was formerly potent.

Ejaculatory impotence, the inability to ejaculate in coitus, is quite rare and is almost always of psychogenic origin.

It seems associated with ideas of contamination or with memories of traumatic experiences. Occasional ejaculatory inability may be expected in older men or in any male who has exceeded his sexual capacity and should cause no concern.

Vaginismus and dyspareunia

Vaginismus is a powerful spasm of the pelvic musculature constricting the vagina so that penetration is painful or impossible. It seems wholly due to antisexual conditioning or psychological trauma and serves as an unconscious defense against coitus. It is treated by psychotherapy and by gradually dilating the vagina with increasingly large cylinders.

Dyspareunia, painful coitus, is generally physical rather than psychological. It is mentioned here only because some inexperienced females fear they cannot accommodate a penis without being painfully stretched. This is a needless fear since the vagina is not only highly elastic but enlarges with sexual arousal, so that even a small female can, if aroused, easily receive an exceptionally large penis.

Disparity in sexual desire constitutes the most common sexual problem. It is to some extent inescapable, since differences in the strength of the sexual impulse and the ability to respond are based on neurophysiological differences. Much disparity, however, is the result of inhibition or of one person having been subjected to more sexual stimuli during the day than the other. The husband who has been seeing attractive females periodically during his work day and who may have had an opportunity to relax on his way back from the office or store is naturally more interested in coitus than his harried wife who has remained at home caring for children and doing housework. Another cause of disparity is a difference in viewpoint. Often a male will anticipate coitus as a palliative to compensate for the trials and tribulations of life, whereas many females are interested in sex only if the preceding hours have been reasonably problem-free and happy. Even in cases of neurophysiological differences in sex drive, the less-motivated partner can be trained to a higher level of interest, since most humans operate well below their sexual capacities.

Psychological fatigue, a growing disinterest in sexual behaviour with a particular partner, sometimes constitutes a problem. Humans are subject to monotony, and coitus may become routine or even a chore. Lessening frequencies of marital coitus are more often the result of this than of age. The solution lies in varying the time, the setting, and in breaking away from habitual techniques and positions.

Preferences for or antipathies toward particular position, techniques, or times not infrequently cause trouble. One partner may desire mouth–genital contact or anal stimulation that the other partner finds disagreeable or perverse. Some wish to have coitus in the light, others insist upon darkness; some prefer morning, others evening. The possibilities of disagreement are legion. Even if disagreements stemming from needless inhibition are overcome, there still remain disparities in preference, and these should be met by the philosophy that, by giving pleasure to another, one obtains pleasure. Needless to say, no partner should insist upon that which is abhorrent to the other after the latter has made honest attempts to cooperate.

Anorgasmy and frigidity

Lack of female orgasm, anorgasmy, is a very frequent problem. One should differentiate between females who become sexually aroused but do not reach orgasm and those who do not become aroused. Only the latter merit the label frigid. It is common for females not to achieve orgasm during the first weeks or months of coital activity. It is almost as though many females must learn how to have orgasm, for after having had one, they respond with increasing frequency. In some cases, the female initially has no idea how to copulate effectively and simply lies passive, expecting the male to bring her to orgasm. Other females resist orgasm because the feeling of being swept away and losing control is frightening. In most cases, however, anorgasmy is simply the result of years of inhibition—having been trained since childhood to avoid yielding to the sexual impulse, it is difficult to metamorphose into a responsive and orgasmic being. In the final analysis, anorgasmy is psychological in origin; few, if any, females lack the neurophysiology necessary for orgasm, and anthropology shows that in sexually permissive societies virtually all females have little difficulty in attaining orgasm in coitus.

Anorgasmy is treated by removing inhibitions, teaching coital techniques, and by inducing orgasm through noncoital methods. The effective therapist should also impress upon the female that not reaching orgasm is no sign of failure or inadequacy on her part or her partner's and that sexual activity is very pleasurable to both, even if orgasm does not ensue. Indeed, some females derive great pleasure and satisfaction without orgasm, a fact that should be made known to anxious husbands. Too great a concern over orgasm defeats itself. As Kinsey once pointed out, thinking is the enemy of sexual pleasure, and a female can scarcely have orgasm if she is worrying about whether she will attain it or not and if she senses that her partner is mentally turning the pages of a marriage manual.

Lastly, sexual problems are often perpetuated by the inability of the partners to communicate freely their feelings to one another. There is a curious and unfortunate reticence about informing one's partner as to what does or does not contribute to one's pleasure. The partner must function on a trial and error basis, ever on the alert for signs indicating the efficacy of his or her efforts. This muteness is even more pronounced when it comes to an individual making suggestions to the partner. Many persons feel that a suggestion or request would be interpreted by the partner that he or she had been inept or at least remiss. As with any other problems, sexual problems can be overcome or ameliorated only if the individuals concerned communicate effectively.

SOCIAL AND CULTURAL ASPECTS

The effects of societal value systems on human sexuality are, as has already been mentioned, profound. The U.S. anthropologist George P. Murdock summarized the situation, saying:

> All societies have faced the problem of reconciling the need of controlling sex with that of giving it adequate expression, and all have solved it by some combination of cultural taboos, permissions, and injunctions. Prohibitory regulations curb the socially more disruptive forms of sexual competition. Permissive regulations allow at least the minimum impulse gratification required for individual well-being. Very commonly, moreover, sex behavior is specifically enjoined by obligatory regulations where it appears directly to subserve the interests of society.

The historical heritage is, of course, the foundation upon which the current situation rests. Western civilizations are basically Greco-Roman in social organization, philosophy, and law, with a powerful admixture of Judaism derived from Christianity. This historical mixture contained incompatible elements: individual freedom was cherished, yet there was a great emphasis on law and proper procedure; the pantheism of the Greeks and Romans clashed with Judeo-Christian monotheism; and the sexual permissiveness of the Greeks met the fanatical antisexuality of early Christianity.

Effects of Christianity on sexual attitudes

In terms of sex, the most important factor was Christianity. While other vital aspects of human life, such as government, property rights, kinship, and economics, were influenced to varying degrees, sexuality was singled out as falling almost entirely within the domain of religion. This development arose from an ascetic concept shared by a number of religions, the concept of the good spiritual world as opposed to the carnal materialistic world, the struggle between the spirit and the flesh. Since sex epitomizes the flesh, it was obviously the enemy of the spirit. While Judaism subscribed to this dichotomous philosophy, it did have the saving grace of largely exempting marriage from its antisexuality. This was not the case with early Christianity, in which sex in any form outside of marriage was unmitigated evil and, within marriage, an unfortunate necessity for procreation rather than pleasure. The powerful antisexuality of the early Christians (note that neither God nor Christ has a wife and that marriage does not exist in heaven) was in part due to their apocalyptic vision of life: they anticipated that the end of the world and the Last Judgment would soon be upon them.

There was no time for a gradual weaning away from the flesh; an immediate and drastic approach was necessary. Indeed, such excessive antisexuality developed that the church itself was finally moved to curb some of its more extreme forms.

As it became evident that human existence was going to continue for some unforeseeable length of time and as occasional intelligent theologians made themselves felt, antisexuality was ameliorated to some extent but still remained a foundation stone of Christianity for centuries. This attitude was particularly unfortunate for women, to whom most of the sexual guilt was assigned. Women, like the original temptress Eve, continued to attract men to commit sin. They were spiritually weak creatures prone to yield to carnal impulses. This is, of course, a classic example of projecting one's own guilty desires upon someone else.

Ultimately, legal control over sexual behaviour passed from the church to the state, but, in most instances, the latter simply perpetuated the attitudes of the former. Priests and clergymen frequently continued to exert powerful extralegal control: denunciations from the pulpit can be as effective as statute law in some cases. Although religion has weakened as a social control mechanism, even today, liberalization of sex laws and relaxation of censorship have often been successfully opposed by religious leaders. On the whole, however, Christianity has become progressively more permissive, and sexuality has come to be viewed not as sin but as a God-given capacity to be used constructively.

Apart from religion, the state sometimes imposes purely secular restrictions. The more totalitarian a government, the more likely it is to restrict or direct sexual behaviour. In some instances, this is simply the consequence of a powerful individual (or individuals) being in a position to impose his (or their) ideas upon the public. In other instances, one cannot escape the impression that sex, being a highly personal and individualistic matter, is recognized as antithetical to the whole idea of strict governmental control and supervision of the individual. This may help explain the rigid sexual censorship exerted by most totalitarian regimes. It is as though such a government, being obsessed with power, cannot tolerate the power the sexual impulse exerts on the population.

Social control of sexual behaviour. Societies differ remarkably in what they consider socially desirable and undesirable in terms of sexual behaviour and consequently differ in what they attempt to prevent or promote. There appear, however, to be four basic sexual controls in the majority of human societies. First, to control endless competition, some form of marriage is necessary. This not only removes both partners from the competitive arena of courtship and assures each of a sexual partner, but it allows them to devote more time and energy to other necessary and useful tasks of life. Despite the beliefs of earlier writers, marriage is not necessary for the care of the young; this can be accomplished in other ways.

Four major areas of sexual control

Second, control of forced sexual relationships is necessary to prevent anger, feuding, and other disruptive retribution.

Third, all societies exert control over whom one is eligible to marry or have as a sexual partner. Endogamy, holding the choice within one's group, increases group solidarity but tends to isolate the group and limit its political strength. Exogamy, forcing the individual to marry outside the group, dilutes group loyalty but increases group size and power through new external liaisons. Some combination of endogamy and exogamy is found in most societies. All have incest prohibitions. These are not based on genetic knowledge. Indeed, many incest taboos involve persons not genetically related (father–stepdaughter, for example). The prime reason for incest prohibition seems to be the necessity for preventing society from becoming snarled in its own web: every person has a complex set of duties, rights, obligations, and statuses with regard to other people, and these would become intolerably complicated or even contradictory if incest were freely permitted.

Fourth, there is control through the establishment of some safety-valve system: the formulation of exceptions to the prevailing sexual restrictions. There is the recognition that humans cannot perpetually conform to the social code and that well-defined exceptions must be made. There are three sorts of exceptions to sexual restrictions: (1) Divorce: while all societies encourage marriage, all realize that it is in the interest of society and the individual to terminate marriage under certain conditions. (2) Exceptions based on kinship: many societies permit or encourage sexual activity with certain kin, even after marriage. Most often these kin are a brother's wife or a wife's sister. In addition, sexual "joking relationships" are often expected between brothers-in-law, sisters-in-law, and cousins. While coitus is not involved, there is much explicit sexual banter, teasing, and humorous insult. (3) Exceptions based on special occasions, ranging from sexual activity as a part of religious rites to purely secular ceremonies and celebrations wherein the customary sexual restrictions are temporarily lifted.

Turning to particular forms of sexual behaviour, one learns from anthropology and history that extreme diversity in social attitude is common. Most societies are unconcerned over self-masturbation since it does not entail procreation or the establishment of social bonds, but a few regard it with disapprobation. Sexual dreams cause concern only if they are thought to be the result of the nocturnal visitation of some spirit. Such dreams were once attributed to spirits or demons known as incubi and succubi, who sought out sleeping humans for sexual intercourse.

Petting among most preliterate societies is done only as a prelude to coitus—as foreplay—rather than as an end in itself. In some parts of sub-Saharan Africa, however, petting is used as a premarital substitute for coitus in order to preserve virginity and avoid pregnancy. There is great variation in petting and foreplay techniques. Kissing is by no means universal, as some groups view the mouth as a biting and chewing orifice ill-suited for expressing affection. While some societies emphasize the erotic role of the female breast, others—such as the Chinese—pay little attention to it. Still others regard oral stimulation of the breast unseemly, being too akin to infantile suckling. Although manual stimulation of the genitalia is nearly universal, a few peoples abstain because of revulsion toward genital secretions. Not much information exists on mouth–genital contact, and one can say only that it is common among some peoples and rare among others.

A considerable number of societies manifest scratching and biting in conjunction with sexual activity, and most of this is done by the female. Sadomasochism in any other form, however, is conspicuous by its absence in preliterate societies.

An enumeration of the societies that permit or forbid premarital coitus is complicated not only by the double standard but also by the fact that such prohibition or permission is often qualified. As a rough estimate, however, 40 to 50 percent of preliterate or ancient societies allowed premarital coitus under certain conditions to both males and females. If one were to count as permissive those groups that theoretically disapprove but actually condone such coitus, the percentage would rise to perhaps 70.

In marital coitus, when sexual access is not only permitted but encouraged, one would expect considerable uniformity in frequency of coitus. This expectation is not fulfilled: social conditioning profoundly affects even marital coitus. On one Irish island reported upon by a researcher, for example, marital coitus is best measured in terms of per year, and among the Cayapas of Ecuador, a frequency of twice a week is something to boast of. The coital frequencies of other groups, on the other hand, are nearer to human potential. In one Polynesian group, the usual frequency of marital coitus among individuals in their late 20s was 10 to 12 per week, and in their late 40s the frequency had fallen to three to four. The African Bala, according to one researcher, had coitus on the average of once or twice per day from young adulthood into the sixth decade of life.

Frequency of marital coitus

Marital coitus is not unrestricted. Coitus during menstruation or after a certain stage of pregnancy is generally taboo. After childbirth a lengthy period of time must often elapse before coitus can resume, and some peoples abstain for magical reasons before or during warfare, hunting expeditions, and certain other important events or ceremonies. In modern Western society one finds menstrual, pregnancy, and postpartum taboos perpetuated under an aesthetic or medical guise, and coaches still attempt to force celibacy upon athletes prior to competition.

Extramarital coitus provides a striking example of the double standard: it is expected, or tolerated, in males and generally prohibited for females. Very few societies allow wives sexual freedom. Extramarital coitus with the husband's consent, however, is another matter. Somewhere between two-fifths and three-fifths of preliterate societies permit wife lending or allow the wife to have coitus with certain relatives (generally brothers-in-law) or permit her freedom on special ceremonial occasions. The main concern of preliterate societies is not one of morality, but of more practical considerations: does the act weaken kinship ties and loyalty? Will it damage the husband's social prestige? Will it cause pregnancy and complicate inheritance or cause the wife to neglect her duties and obligations? Most foreign of all to Western thinking is that of those peoples whose marriage ceremony involves the bride having coitus with someone other than the groom, yet it is to be recalled that this practice existed to a limited extent in medieval Europe as jus primae noctis, the right of the lord to the bride of one of his subjects.

Sexual deviations and sex offenses are, of course, social definitions rather than natural phenomena. What is normative behaviour in one society may be a deviation or crime in another. One can go through the literature and discover that virtually any sexual act, even child–adult relations or necrophilia, has somewhere at some time been acceptable behaviour. Homosexuality is permitted in perhaps two-thirds of human societies. In some groups it is normative behaviour, whereas in others it is not only absent but beyond imagination. Generally, it is not an activity involving most of the population but exists as an alternative way of life for certain individuals. These special individuals are sometimes transvestites—that is, they dress and behave like the opposite sex. Sometimes they are regarded as curiosities or ridiculed, but more often they are accorded respect and magical powers are attributed to them. It is noteworthy, however, that aside from these transvestites, exclusive homosexuality is quite rare in preliterate societies.

In conclusion, the cardinal lesson of anthropology is that no type of sexual behaviour or attitude has a universal, inherent social or psychological value for good or evil—the whole meaning and value of any expression of sexuality is determined by the social context within which it occurs.

Class distinctions. Differences in sexual behaviour between classes within technologically developed societies are very marked. Civilizations are made up of class hierarchies, and the different subgroups normally develop their own value systems. Most of the knowledge of the sexual behaviour and attitudes of ancient cultures is that of the upper or ruling class; the behaviour and feelings of the slaves and peasants were seldom recorded. There is the impression—probably a correct one—that throughout history the lower socio-economic class was the most permissive. Sex has always been one of the few pleasures of the poor and oppressed. On the other hand, one must not overlook the fact that a fanatical Puritanism can also flourish at the bottom of the social scale, and, hence, one can never assume that low status and sexual permissiveness are inevitably linked.

Breakdown of class differences in sexual behaviour

The Kinsey studies showed considerable social class differences in sexuality in the United States, chiefly in that the lower class was more tolerant of nonmarital coitus. More recent studies indicate that these class differences have rapidly broken down. Increased literacy and the influence of mass media have made the population more homogeneous in sexual attitudes. One can find, moreover, reversals of the previous pattern: a lower class person on the way up the social ladder may be quite conservative in his sexual views, feeling that this facilitates upward mobility, whereas the person secure in his or her high social status often feels that he or she can afford to flout convention. Actually, the most sexually liberal are those at the very bottom, who have nothing to lose, and those at the very top, who are beyond social retribution.

The great middle class remains the bastion of traditionalism, and it is here that the double standard of morality is most prominent. The intellectualized liberalism of the upper level seeps down only slowly, and the pragmatic egalitarianism of the lower level does not penetrate far upward.

Economic influences. Systems of production and distribution have had a growing influence on sexual behaviour since the Industrial Revolution. The old family pattern was inexorably disrupted by the rise of the industrial state. Children were no longer kept at home to share in the work and be economic assets but left for school or for nonfamily employment, and the degree of parental control diminished. The "working wife" employed outside the home, once found only among the impoverished, has gradually become the typical wife. With her enhanced economic power and her greater association with people outside the home, she became less a chattel. As the population left the family farm and tight-knit small communities for anonymous big-city existence, not only parental but societal controls over behaviour were weakened. Society became increasingly nomadic with improved transportation and job opportunities. Cultural and ethnic subgroups that formerly would have had little contact were thrown together in the same schools, factories, offices, and neighbourhoods.

All of this vast uprooting and rearranging naturally altered sexual attitudes and behaviour. The individual no longer had the option of choosing to conform or depart from a rather clear-cut sexual moral code but instead was faced with a multiplicity of choices of varying degrees of social acceptability. The major sexual change—one still in progress—was the emancipation of women, which brought with it an increasing acceptance of premarital sexual activity, the concept of woman as a human being with her own sexual needs and rights, and the possibility of terminating an unhappy marriage without incurring serious social censure. A second major change was the erosion of simplistic value systems: with increased mobility and social mixing, the individual learned that the values and attitudes he or she had unquestioningly accepted were not necessarily shared by neighbours and co-workers. As a result, life became not only more complex but more permissive. This growing tolerance has in recent decades extended, to a limited extent, to homosexuality. There is no evidence that homosexuality or other deviant behaviour has measurably increased as a result of society's urbanization and technological progress, but one gains the impression of an increase simply because these topics, previously unmentionable, are now openly discussed in the mass media.

While the old monolithic value systems broke down and individuals were accorded a wider variety of choices in terms of sexual life, there developed a paradoxical trend toward homogeneity as a result of mobility, the mass media, and increasing economic parity. Geographical and social-class differences in sexual attitudes and behaviour have steadily lessened. The plumber's family and the banker's family are now indistinguishable in terms of dress; both have automobiles; their offspring attend the same schools; and they share the same newspapers, magazines, and television programs. One might summarize by saying that society is homogeneous in that everyone now has available a wide diversity of sexual attitudes and activities.

Legal regulation. Sex laws, the origins of which, as mentioned above, are found within the church, are unique in one important respect. Whereas all other laws are basically concerned with the protection of person or property, the majority of sex laws are concerned solely with maintaining morality. The issue of morality is minimal in other laws: one can legitimately evict an impoverished old couple from their mortgaged home or sentence a hungry man for stealing food. Only in the realm of sex is there a consistent body of law upholding morality.

Early examples of sex laws

The earliest sex laws of which there is knowledge are from the Near East and date back to the 2nd millennium BC. They are remarkable in three respects: there are great omissions—certain acts are not mentioned whereas others receive detailed attention; some laws seem almost contradictory; and penalties are often extraordinarily severe. One obtains the distinct impression that these laws were case law—that is, laws formulated upon specific cases as they arose rather than being the result of lengthy judicial deliberation done in advance. These laws influenced Judaic and, hence, Christian thinking, and some were immortalized in the Bible, chiefly in Leviticus.

As mentioned earlier, when secular law replaced religious law, there was rather little change in content. In Europe the Napoleonic Code represented a break with tradition and introduced some measure of sexual tolerance, but in England and the United States there was no such rift with the past. In the latter country, as each new state joined the union, its sex laws simply duplicated, to a great extent, those of pre-existing states; legislators were disinclined to debate sexual issues or to risk losing votes by discarding or weakening sex laws.

Sex laws may be grouped in three categories: (1) Those concerned with protection of person. These are based on the element of consent. These otherwise logical laws become problematic when society deems that minors, mental retardates, and the insane are incapable of giving consent—hence, coitus with them is rape. (2) Those concerned with preventing offense to public sensibilities. Statutes preclude public sexual activity, exhibitionism, and offensive solicitation. (3) Those concerned with maintaining sexual morality. These constitute the majority of sex laws, covering such items as premarital coitus, extramarital coitus, incest, homosexuality, prostitution, peeping, nudity, animal contact, transvestism, censorship, and even specific sexual techniques—chiefly oral or anal. Laws relating to sexual conduct and morality are generally far more extensive in the United States than in western Europe and most other areas of the world.

In recent years, in Europe and the United States, a number of highly respected legal, medical, and religious organizations have deliberated on the whole issue of the legal control of human sexuality. They have been unanimous in the conclusion that, while the laws protecting person and public sensibilities should be retained, the purely moral laws should be dropped. Specifying what consenting adults do sexually in private, it is argued, should not be subject to legal control.

In the final analysis, sexuality, like any other vital aspect of human life, must be dealt with on an individual or societal level with a combination of rationality, sensitivity, and tolerance if society is to avoid personal and social problems arising from ignorance and misconception. (P.H.Ge./Ed.)

HOMOSEXUALITY

Homosexual behaviour consists of choosing a partner of the same sex for sexual interaction. The term homosexual is used to characterize individuals who prefer romantic attachments and sexual interaction with the same sex and typically are maximally aroused by same-sex erotic imagery. The term is used for both sexes, although female homosexuality is often referred to as lesbianism. The label homosexual is sometimes inaccurately applied to individuals who predominantly experience heterosexual arousal but who have experienced occasional or fleeting desire or fantasy of same-sex erotic activity.

Because of the difficulty in classification, sex researcher Alfred C. Kinsey and his associates created a seven-point scale that included exclusively heterosexual behaviour at one extreme and exclusively homosexual behaviour at the other. Erotic fantasy could be similarly scaled. For practical purposes, the critical factors in classifying an individual as homosexual, heterosexual, or bisexual are: (1) which sex or sexes one desires to form romantic bonds with and (2) how easily one responds sexually to the chosen partners.

Kinsey data

Statistics on homosexuality. According to the extensive data gathered by Kinsey and his associates for his landmark work *Sexual Behavior in the Human Male* (1948), a study of 5,300 males, approximately 50 percent had a same-sex genital experience before puberty. Twenty-five percent had more than incidental homosexual experience for at least three years between the ages of 16 and 55 years, and 37 percent had at least one homosexual experience leading to orgasm after puberty. Ten percent were exclusively homosexual for a period of at least three years between the ages of 16 and 55.

By age 30, one-quarter of the 5,940 females interviewed by Kinsey (*Sexual Behavior in the Human Female*, 1953) had recognized erotic responses to the same sex. About 20 percent reported some same-sex erotic experience; but, because these figures included "casual" contacts that elicit erotic response, the more revealing figure is same-sex experience to orgasm: 13 percent of women had a homosexual experience to orgasm prior to age 45. Approximately 2 to 3 percent reported having exclusive homosexual experience.

More recent data from a study funded by the Playboy Foundation and from the Institute for Sex Research support the earlier Kinsey findings on homosexual experiences and indicate that the incidence figures have remained constant since the 1950s. It is likely, therefore, that, although sexual liberation movements of the late 20th century brought greater freedom to enjoy various sexual acts, including homosexual experiences, with less guilt and shame, there was not a measurable increase in the number of individuals having homosexual experiences. What increased significantly was the candour with which sexual preference could be discussed and displayed.

Prenatal and environmental influences. The "biologic versus learned" dichotomy regarding the etiology of homosexual preference—whether it is determined by inborn factors or by environmental influences—continues to muddle scientific investigation of the interacting prenatal and postnatal factors that contribute to all sexual orientation. There is now a substantial body of evidence to suggest that human beings are born with a sexual potential and that heterosexual, homosexual, bisexual, or asexual preferences unfold during the experiences of childhood and adolescence. This is not to say that prenatal and genetic factors are unimportant. A useful analogy given by psychologist John A. Money of Johns Hopkins University is that of "handedness." A human is born not right- or left-handed but bipotential. Experiences in the first few years of life define whether the individual will develop the right, the left, or both cerebral hemispheres—and handedness. There might be predispositions to handedness, which are supported by the fact that most individuals are right-handed; however, postnatal experiences probably override such predispositions. Similarly, there is some evidence that suggests possible predispositions to erotic preferences. Such predispositions, however, do not preordain homosexual or heterosexual preferences; postnatal events can override this influence.

Evidence that humans are born bipotential with regard to erotic orientation comes from several sources. The first is the study of clinical cases of intersexuality, that is, individuals born with ambiguous genitalia and raised in the opposite gender of their genetic, or chromosomal, sex. (Males have an X and a Y sex chromosome; females have two X sex chromosomes. The combination that an individual possesses can be determined by testing.) Approximately 60 individuals who were raised in the sex opposite of their genetic sex were studied by Money and his associates. In addition, in a bizarre case, chromosomal male identical twins were reared as opposite sexes owing to a circumcision accident in which the penis of one infant was cut off. The infant was subsequently castrated and a functional vagina was made through plastic surgery. In each of the intersexuality cases, as well as the twins case, the gender identity and gender role—that is, the individual's personal sense of masculinity or femininity and behavioral manifestations—that developed was compatible with the sex of rearing regardless of the genetic and prenatal classification. At puberty these individuals were generally attracted to the sex opposite of their sex of rearing. This suggests that sexual orientation is primarily established in postnatal experiences.

Acceptance of homosexual behaviour

The second source of information on the influences of nature and nurture is anthropologic research. In almost two-thirds of the 76 societies that Clellan S. Ford and Frank A. Beach reviewed in *Patterns of Sexual Behavior* (1951), homosexual activities were considered acceptable under certain circumstances. For example, homosexual activity might be ritualized during childhood and adolescence. In none of these societies was there record of exclusive homosexuality; rather, heterosexual pair-bonding occurred during adulthood, and homosexuality either stopped in adulthood or was permitted under special circumstances. So it can be seen that exclusive homosexuality—and heterosexuality—are atypical in many cultures. Polarizations to homosexuality and to heterosexuality appear to be a product of individual civilizations.

Programming of erotic orientation. The programming of erotic orientation is subsumed under the development of gender identity and role. Children learn to imitate the parent of the same sex and complement the opposite-sexed parent. A child can reliably distinguish gender by age five. Boys, for example, are aware of the sexual anatomy of both sexes as well as the stereotypic characteristics of the masculine sex role. Also, by age five, they are aware that they probably will someday marry a woman. The acculturation of heterosexuality is quite overt. C.A. Tripp, author of *The Homosexual Matrix* (1975), stated, "All the religious and social traditions directly support family living and the kinds of mateships that comprise heterosexuality. Certainly there is nothing mysterious in how family life communicates itself."

Similarly, there is nothing mysterious about the discouragement of homosexuality. Homosexual behaviour is subject to extreme prejudice in most of Western society—it has been considered "perverted and sick" by a large portion of the population. As with most prejudices, homosexuals are seen in terms of various stereotypes, such as effeminacy, characteristic gestures and mannerisms, and promiscuity. The programming that encourages heterosexuality and discourages homosexuality, even close physical contact with the same sex, generally results in the stunting of any budding homosexual fantasies or behaviour in most prepubescents. Heterosexuality is left to unfold.

Even though homosexual behaviour is not sanctioned, homosexual erotic imagery may persist. Data published by sex researchers Virginia Johnson and William Masters and by author Nancy Friday on selected samples suggested that many heterosexuals use homosexual imagery to enhance their sexual arousal when masturbating or having intercourse but do not actually have an interest in having a homosexual experience. Similarly, many homosexuals use heterosexual imagery to enhance their arousal but do not desire heterosexual experiences. Research on broader samples is necessary; unfortunately, it is unlikely that a wide cross-section of individuals of all ages and backgrounds would be willing to discuss their erotic fantasies openly.

Why some individuals develop homosexual orientations despite contrary socialization is not yet completely known. There are obviously both prenatal and postnatal contributing factors that interact to cause this particular sexual orientation, but thus far only some of the labyrinth has been explored.

Prenatal factors. There are now extensive data from lower mammals and primates, as well as indirect data from human clinical investigations, that show that the presence of differential amounts of the hormone androgen in the fetus during a short prenatal period influences the acquisition and expression of sexually dimorphic behaviour. Androgen is secreted by the fetal gonads and is present in widely differing amounts in all male and female fetuses. Males have about seven times the amount of androgen as do females.

Postnatal androgenization in females

Numerous experiments have reported that female animals exposed to excess fetal androgen behave more like males in various tests of sexually dimorphic behaviour, including a mating preference for the same sex. John Money and Anke Ehrhardt (of Columbia University) found that human females exposed to excess fetal androgen as the result of a condition known as the adrenogenital syndrome behaved throughout childhood with a high degree of energy in rough outdoor play, preferred boys as playmates, and had little interest in stereotypic activities of girls or verbalized interest in marriage or childbirth. One study found that during adolescence these girls maintained their intense interest in athletics but had difficulty finding female peers with similar interests. Their interest in dating began late; and of those who eventually did begin their sex lives in young adulthood, a significant number found themselves bisexually aroused and eventually a number had homosexual experiences. These findings support an earlier report that found that 48 percent of women who were exposed to excess prenatal androgens and developed postnatal androgenization (and therefore were partially virilized; *i.e.*, developed some male secondary sex characteristics) had experienced bisexual imagery and 18 percent had had homosexual experiences.

One approach to understanding prenatal influences on sexual orientation is to study the release of hormones from the brains of individuals who are homosexual and transsexual (individuals who, in addition to having homosexual interaction, desire to change their gender). Normally, male and female brains differ, because the female produces hormones in a cyclic manner as part of the menstrual cycle. Günter Dörner in East Germany reported a positive or male-like hormone feedback action in some homosexual women. Similar findings were noted by Lloyd Seyler at the University of Connecticut, who found a male-like hormone response in transsexual women.

Finally, some investigators have found elevated or diminished gonadal hormone levels in some homosexual men and women as compared with those of heterosexuals. The presence of variant hormone levels does not mean that raising or lowering hormone levels in homosexuals will change their sexual preference. Changing the hormone levels would usually have no effect, although if levels were lowered enough, this would raise the threshold for sexual arousal. Rather, a likely explanation for the different hormone levels is that prenatal androgen may have changed the organization of brain pathways, thereby making some individuals less susceptible to postnatal programming toward heterosexuality.

A third source of information about prenatal factors is from the study of boyhood effeminacy. For as long as 22 years Money studied nine boys who showed effeminate behaviour, including gestures, speech, play patterns, and cross-dressing, in early childhood. All nine eventually became homosexual. The appearance of this gender-divergent behaviour at an early age may be suggestive of prenatal influences.

Family patterns

Postnatal factors. Why do certain individuals move into full or partial homosexuality despite society's programming of heterosexuality and its condemnation of homosexuality? Irving Bieber reported family interactions commonly described by male homosexuals coming to him for psychotherapy because of dissatisfaction with their homosexuality. The mothers were conceived as close-binding, overprotective, and seductive, while the fathers were conceived as distant and hostile. This background has been noted by other psychotherapists for homosexuals. The reasons why such a background in particular makes the individual have difficulties in sexual interaction with women are difficult to verify. Some lesbians have been found to have similar patterns in their backgrounds, showing a history of seductive fathers and distant mothers.

The reports by various psychotherapists treating disturbed homosexuals produced the commonly held theory that homosexuality was a disease state resulting from pathological influences. It followed that treatment of such pathological influences would result in the spontaneous appearance of the nondisease state, heterosexuality.

The disease model has been criticized for various reasons. The clinical evidence has been viewed as biassed since only troubled homosexuals came to the psychotherapists' offices. Working with less biassed samples of homosexuals, Evelyn Hooker, Alan P. Bell, and Martin S. Weinberg reported very divergent family constellations.

In addition, the cross-cultural, intersexual, and sociologic data previously mentioned have made many sexologists aware that the human child is born bipotential with regard to erotic orientation. Thus, many began to catalog factors contributing to or inhibiting the unfolding of heterosexuality and homosexuality, rather than labelling the latter a pathological disease state.

Finally, aversive conditioning of the homosexual response (behaviour modification that tries to foster negative associations with erotic response to homosexual stimuli) has not led to the spontaneous appearance of heterosexuality, as the disease model would suggest. Changing individuals' sexual orientations has met with variable success and, when successful, requires that an individual's stunted heterosexual responsiveness be allowed and encouraged to unfold.

The rejection of the disease model by many sexologists was most dramatically marked by the vote of the American Psychiatric Association in 1973 to remove "ego-syntonic homosexuality" (meaning the condition of a person content with his or her homosexuality) from the categorization of psychiatric illness. Regardless of the controversies about whether or not homosexuality is a disease, there is increased scientific recognition that the same prenatal and postnatal factors that contribute to heterosexuality also contribute to homosexuality; the difference is the relative contribution of each in an individual's biography.

Self-labelling

A likely influence in the development of homosexuality is self-labelling. Some individuals discover that they have homosexual imagery, which, as previously noted, might be quite a common occurrence; however, some individuals then label themselves homosexual and polarize themselves away from heterosexuality. Milton Diamond and Arno Karlen in their book *Sexual Decisions* (1980) noted, "Like other minorities homosexuals respond by flaunting their deviance and living up to the majority's expectations." Once labelled, some male homosexuals may develop effeminate mannerisms (eloquently described by Tripp in *The Homosexual Matrix*). Individuals with low self-esteem from childhood trauma might be particularly susceptible to such labelling.

Polarization is reinforced by sexual performance concerns—Can I get an erection? Will I be orgasmic? Can I satisfy my partner?—that characterize new heterosexual relationships. Of the homosexual men with heterosexual experience, Marcel Saghir and Eli Robins, in their *Male and Female Homosexuality* (1973), found that about 30 percent were impotent with females and that 50 percent of the heterosexually inexperienced were afraid of females and were fearful of being impotent with them. Similarly, 61 percent of the 57 lesbians studied reported that they did not experience orgasm in their heterosexual experiences. Concerns about sexual performance are dramatically decreased in homosexual interaction. Thus individuals who find themselves having homosexual fantasies and who are not at ease with the opposite sex or are concerned about performance may find homosexual activity easier. The repeated conditioning of pleasure with orgasm eventually results in the fantasies becoming more unvarying, and polarization is further enhanced.

Homosexual life-styles. In an attempt to provide further empirical data, two major investigations of homosexuality were conducted in the 1960s and '70s, the results of which were published in 1978 and in 1979. The researchers were Bell and Weinberg of the Institute for Sex Research and Masters and Johnson.

Bell and Weinberg interviewed a nonrepresentative sample of some 1,500 persons, including homosexuals and heterosexuals, males and females, from the San Francisco Bay area. The major conclusion from their work, *Homosexualities: A Study of Diversity Among Men and Women* (1978), is that it is impossible to predict the social and psychologic adjustment of homosexuals or heterosexuals by identifying erotic preferences. Both groups show great diversity in a variety of characteristic categories, including life-style, occupation, type of romantic attachment, mental health, happiness, and sexual technique. In some of the characteristic categories, the differences among homosexuals are as wide as or wider than those between homosexuals and heterosexuals.

In particular, 14 percent of male homosexuals were "close-coupled" in monogamous bonds. The majority were not involved in monogamous relationships, and 25 percent called "open-coupled" lived with a special partner but did occasional "cruising" of singles' bars or "gay baths" for a sexual partner. A minority of 21 percent were described as "functional" and organized much of their life around sexual activity and cruising. Approximately 18 percent, described as "dysfunctional," were not content with their homosexuality. The last group, 23 percent, classified as "asexual," were not coupled and had little sexual activity.

Levels of activity

Homosexual men were more sexually active than homosexual women, and less emphasis in the females' thinking was placed on sexual contact. Lesbian public cruising was infrequent, and women tended to be less interested in impersonal sexual encounters than male homosexuals were. Almost 40 percent of males did not cruise or did so infrequently, and fewer than 20 percent of the females engaged in occasional cruising of gay bars and private parties. The females had fewer sexual partners in their lives and were more inclined toward stable monogamous relationships.

Males tended to have several sexual experiences soon after puberty, whereas the females discovered their homosexuality later, usually after a romantic relationship with one female. These data are similar to those of Saghir and Robins, who noted that most lesbians had heterosexual relationships during adolescence that were generally less psychosexually satisfying than those the heterosexual females reported. Generally, the lesbians were more accepting and less guilty about their homosexuality than were the men studied.

Masters and Johnson's text, *Homosexuality in Perspective* (1979), is rich in detail about human sexual behaviour regardless of sexual orientation. Their research documented that the sexual response and physiological capacity to respond to sexual stimuli were not different between homosexuals and heterosexuals. In addition, heterosexuals might have something to learn from the sexual techniques of homosexuals. The latter were generally less goal-oriented in their sexual practices. Male homosexuals in particular were less likely to have the performance demands experienced by heterosexual men for erection to accomplish vaginal penetration. Homosexuals more often concentrated on full body contact. Some heterosexuals were equally effective in stimulating their partners, but communication of desired stimulation was generally found to be of higher quality within the sexes than between the sexes.

In addition, Masters and Johnson reported on their treatment of 57 homosexual male couples and 27 lesbian couples with sexual dysfunction, using techniques similar to those used to treat heterosexual couples. Failure rates in short-term intensive therapy were similarly low as compared with heterosexual couples.

Finally, the Masters and Johnson publication demonstrated that for the 20 percent of homosexuals who were distressed with their homosexuality, short-term psychotherapy could be effective in "converting" or "reverting" them to heterosexuality. There were 54 males and 13 females requesting such therapy, and the failure rate after a five-year follow-up was 28 percent.

Knowing that a person prefers homosexual eroticism tells little else about him or her. Scientific research has confirmed what to some may be obvious—stereotypes of homosexuals are inaccurate and indeed create much of the difficulty that some homosexuals experience. Meanwhile, researchers continue to search for the multitude of determinants that contribute to various individual differences, including gender identity and sexual orientation. Having a close-binding mother and a neglecting father may be associated with relationship difficulties in adulthood and may even result in total rejection of members of the opposite sex as sexual partners. This is not the cause of homosexuality; but, in susceptible individuals, it can be an influence. Other factors such as self-labelling and anxiety over sexual performance also contribute to difficulties in forming a satisfying relationship with members of the opposite sex.

It is likely that in the future attitudes toward homosexuals in the community will become more tolerant. The majority of homosexuals will be able to live contented lives in the thriving homosexual subculture, particularly in certain metropolitan areas such as San Francisco, London, or Amsterdam. The minority who continue to be discontented will be able to seek sex therapy and change their erotic preference. (M.F.S./Ed.)

BIBLIOGRAPHY. C.N. ARMSTRONG and A.J. MARSHAL (eds.), *Intersexuality in Vertebrates Including Man* (1964), a collection of articles by various authors dealing with all aspects of intersexuality in many kinds of animals, including humans,

from physiological and genetic viewpoints; C.R. AUSTIN, *Fertilization* (1965), a concise general account of the nature of reproductive cells, the basic process of fertilization of the egg, and the various adaptive devices employed by animals and plants to ensure successful and normal fertilization, written for biology students; M. BASTOCK, "The Physiology of Courtship and Mating Behavior," *Advances in Reproductive Physiology,* 2:9–51 (1967), a review intended for biologists, but of general interest; N.J. BERRILL, *Sex and the Nature of Things* (1953), an award-winning inclusive account of sex, its diversity, and its significance in relation to evolution and animal behaviour, for the lay reader; *The Person in the Womb* (1968), a treatment of human development from conception to infancy, including the determination of sex at conception and the influence of hormones on sexuality during pregnancy; W.S. BULLOUGH, *The Vertebrate Reproductive Cycles,* 2nd ed. (1961), a short, concise account of reproduction, particularly in birds and mammals, with emphasis on hormonal and seasonal controls of sexual maturity and reproduction; L.J. and M.J.G. MILNE, *The Mating Instinct* (1954, paperback ed. 1968), very inclusive and detailed descriptions of the courting and mating behaviour of animals, extensively illustrated photographically; G. PINCUS, *The Control of Fertility* (1965), a highly scientific discussion, by a pioneering authority, of the reproductive process in man and other mammals, with emphasis on the action of hormones on sex cell production, on the first stages of development, and on the problems of fertility; H. WENDT, *The Sex Life of Animals* (1965), a very descriptive account of courting and mating, and sexuality generally, throughout the animal kingdom, translated from an earlier German edition; G.C. WILLIAMS, *Adaptation and Natural Selection* (1966), an illuminating and thoughtful discussion of sexual reproduction in relation to the processes of biological evolution and adaptation.

An anthropological overview of sex is given in DONALD S. MARSHALL and ROBERT C. SUGGS (eds.), *Human Sexual Behavior: Variations in the Ethnographic Spectrum* (1971); GEORGE P. MURDOCK, *Social Structure* (1949); and CLELLAN S. FORD and FRANK A. BEACH, *Patterns of Sexual Behavior* (1951); the latter also covers mammalian behaviour. Data on sexual behaviour and attitudes in the United States are in ALFRED C. KINSEY *et al., Sexual Behavior in the Human Male* (1948) and *Sexual Behavior in the Human Female* (1953); while IRA L. REISS, *The Social Context of Premarital Sexual Permissiveness* (1967), provides more recent information on attitudes. General information may be found in JAMES L. MCCARY, *Human Sexuality* (1967), a book that also serves as an excellent sex manual. The best single volume on homosexuality is that of DONALD J. WEST, *Homosexuality* (1968). Genetic, hormonal, and gender matters are summarized in two technical volumes: RICHARD GREEN and JOHN MONEY (eds.), *Transsexualism and Sex Reassignment* (1969); and ROBERT J. STOLLER, *Sex and Gender: On the Development of Masculinity and Femininity* (1968). Sexual physiology is comprehensively treated in WILLIAM H. MASTERS and VIRGINIA E. JOHNSON, *Human Sexual Response* (1966), while their second volume, *Human Sexual Inadequacy* (1970), deals with behaviour therapy. The history of Judeo-Christian attitudes toward sex is presented in detail in L.M. EPSTEIN, *Sex Laws and Customs in Judaism* (1968); and DERRICK S. BAILEY, *Sexual Relation in Christian Thought* (1959). Sex laws are covered by three volumes: RALPH SLOVENKO (ed.), *Sexual Behavior and the Law* (1965); EDWARD SAGARIN and DONAL E.J. MACNAMARA, *Problems of Sex Behavior* (1968); and the AMERICAN LAW INSTITUTE, *Model Penal Code,* 2 vol. (1955–57). BENJAMIN B. WOLMAN (ed.), *Handbook of Human Sexuality* (1980), is a collection of essays covering a wide spectrum of topics for the general reader; G. MITCHELL, *Human Sex Differences* (1981), a primotologist's investigation of the causes of gender-related differences in physiology, pathology, aging, etc.

Homosexuality is discussed in J. MARMOR (ed.), *Sexual Inversion: The Multiple Roots of Homosexuality* (1965); MARCEL SAGHIR and ELI ROBINS, *Male and Female Homosexuality: A Comprehensive Investigation* (1973); C.A. TRIPP, *The Homosexual Matrix* (1975); ALAN P. BELL and MARTIN S. WEINBERG, *Homosexualities: A Study of Diversity Among Men and Women* (1978); WILLIAM H. MASTERS and VIRGINIA JOHNSON, *Homosexuality in Perspective* (1979); MILTON DIAMOND and ARNO KARLEN, *Sexual Decisions* (1980); ALAN P. BELL *et al., Sexual Preference: Its Development in Men and Women* (1981); JACK BABUSIO, *We Speak for Ourselves: Experience in Homosexual Counselling* (1976); VERN L. BULLOGH, *Homosexuality: A History* (1979); JONATHAN KATZ, *Gay American History: Lesbians and Gay Men in the U.S.A.: A Documentary* (1976); JEFFREY WEEKS, *Coming Out: Homosexual Politics in Britain from the Nineteenth Century to the Present* (1977); and GEORGE WEINBERG, *Society and the Healthy Homosexual* (1975).

Shakespeare

Widely regarded as the greatest writer of all time, William Shakespeare (or Shakspere) occupies a position unique in world literature. Other poets, such as Homer and Dante, and novelists, such as Leo Tolstoy and Charles Dickens, have transcended national barriers; but no writer's living reputation can seriously compare with that of Shakespeare, whose plays, written in the late 16th and early 17th centuries for a small repertory theatre, are now performed and read more often and in more countries than ever before. The prophecy of his great contemporary, the poet and dramatist Ben Jonson, that Shakespeare "was not of an age, but for all time," has been marvellously fulfilled. It may be audacious even to attempt a definition of his greatness, but it is not so difficult to describe the variety of gifts that enabled him to create imaginative visions of pathos and mirth that, whether read in the book or witnessed in the theatre, fill the mind and linger there. He is a writer of great intellectual rapidity, perceptiveness, and poetic power. Other writers have had these qualities. But with Shakespeare the keenness of mind was applied not to abstruse or remote subjects but to human beings and their complete range of emotions and conflicts. Other writers have applied their keenness of mind in this way. But Shakespeare is astonishingly clever with words and images, so that his mental energy, when applied to intelligible human situations, finds full and memorable expression, convincing and imaginatively stimulating. As if this were not enough, the art form into which his creative energies went was not remote and bookish but involved the vivid stage impersonation of human beings, commanding sympathy and inviting vicarious participation. Thus many of Shakespeare's great merits can survive translation into other languages and into cultures remote from that of Elizabethan England.

This article is divided into the following sections:

Shakespeare the man

LIFE

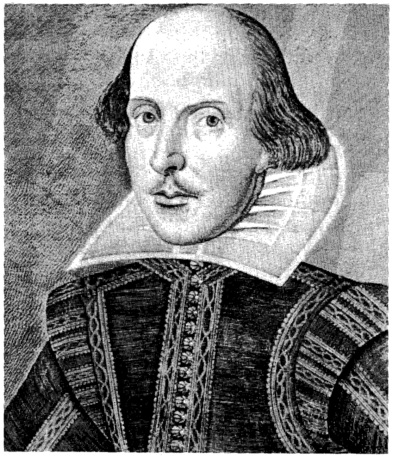
Although the amount of factual knowledge available about Shakespeare is surprisingly large for one of his station in life, many find it a little disappointing, for it is mostly gleaned from documents of an official character. Dates of baptisms, marriages, deaths, and burials; wills, conveyances, legal processes, and payments by the court—these are the dusty details. There is, however, a fair number of contemporary allusions to him as a writer, and these add a reasonable amount of flesh and blood to the biographical skeleton.

Early life in Stratford. The parish register of Holy Trinity Church, Stratford-upon-Avon, Warwickshire, shows that he was baptized there on April 26, 1564; his birthday is traditionally celebrated on April 23. His father, John Shakespeare, was a burgess of the borough, who in 1565 was chosen an alderman and in 1568 bailiff (the position corresponding to mayor, before the grant of a further charter to Stratford in 1664). He was engaged in various kinds of trade and appears to have suffered some fluctuations in prosperity. His wife, Mary Arden, of Wilmcote, Warwickshire, came from an ancient family and was the heiress to some land. (Given the somewhat rigid social distinctions of the 16th century, this marriage must have been a step up the social scale for John Shakespeare.)

Education at Stratford's grammar school

Stratford enjoyed a grammar school of good quality, and the education there was free, the schoolmaster's salary being paid by the borough. No lists of the pupils who were at the school in the 16th century have survived, but it would be absurd to suppose the bailiff of the town did not send his son there. The boy's education would consist

By courtesy of the Folger Shakespeare Library, Washington, D.C.



Shakespeare, first proof of an engraved portrait by Martin Droeshout, from the frontispiece of the First Folio edition of Shakespeare's plays, 1623. In the Folger Shakespeare Library, Washington, D.C.

mostly of Latin studies—learning to read, write, and speak the language fairly well and studying some of the classical historians, moralists, and poets. Shakespeare did not go on to the university, and indeed it is unlikely that the tedious round of logic, rhetoric, and other studies then followed there would have interested him.

Instead, at the age of 18 he married. Where and exactly when are not known, but the episcopal registry at Worcester preserves a bond dated November 28, 1582, and executed by two yeomen of Stratford, named Sandells and Richardson, as a security to the bishop for the issue of a license for the marriage of William Shakespeare and "Anne Hathaway of Stratford," upon the consent of her friends and upon once asking of the banns. (Anne died in 1623, seven years after Shakespeare. There is good evidence to associate her with a family of Hathaways who inhabited a beautiful farmhouse, now much visited, two miles from Stratford.) The next date of interest is found in the records of the Stratford church, where a daughter, named Susanna, born to William Shakespeare, was baptized on May 26, 1583. On February 2, 1585, twins were baptized, Hamnet and Judith. (The boy Hamnet, Shakespeare's only son, died 11 years later.)

How Shakespeare spent the next eight years or so, until his name begins to appear in London theatre records, is not known. There are stories—given currency long after his death—of stealing deer and getting into trouble with a local magnate, Sir Thomas Lucy of Charlecote, near Stratford; of earning his living as a schoolmaster in the country; of going to London and gaining entry to the world of theatre by minding the horses of theatregoers; it has also been conjectured that Shakespeare spent some time as a member of a great household and that he was a soldier, perhaps in the Low Countries. In lieu of external evidence, such extrapolations about Shakespeare's life have often been made from the internal "evidence" of his writings. But this method is unsatisfactory: one cannot conclude, for example, from his allusions to the law that Shakespeare was a lawyer; for he was clearly a writer, who without difficulty could get whatever knowledge he needed for the composition of his plays.

Career in the theatre. The first reference to Shakespeare in the literary world of London comes in 1592, when a fellow dramatist, Robert Greene, declared in a pamphlet written on his deathbed:

Greene's attack on the "upstart" dramatist

> There is an upstart crow, beautified with our feathers, that with his *Tygers heart wrapt in a Players hide* supposes he is as well able to bombast out a blank verse as the best of you; and, being an absolute *Johannes Factotum,* is in his own conceit the only Shake-scene in a country.

It is difficult to be certain what these words mean; but it is clear that they are insulting and that Shakespeare is the object of the sarcasms. When the book in which they appear (*Greenes, groats-worth of witte, bought with a million of Repentance,* 1592) was published after Greene's death, a mutual acquaintance wrote a preface offering an apology to Shakespeare and testifying to his worth. This preface also indicates that Shakespeare was by then making important friends. For, although the puritanical city of London was generally hostile to the theatre, many of the nobility were good patrons of the drama and friends of actors. Shakespeare seems to have attracted the attention of the young Henry Wriothesley, the 3rd earl of Southampton; and to this nobleman were dedicated his first published poems, *Venus and Adonis* and *The Rape of Lucrece.*

One striking piece of evidence that Shakespeare began to prosper early and tried to retrieve the family fortunes and establish its gentility is the fact that a coat of arms was granted to John Shakespeare in 1596. Rough drafts of this grant have been preserved in the College of Arms, London, though the final document, which must have been handed to the Shakespeares, has not survived. It can scarcely be doubted that it was William who took the initiative and paid the fees. The coat of arms appears on Shakespeare's monument (constructed before 1623) in the Stratford church. Equally interesting as evidence of Shakespeare's worldly success was his purchase in 1597 of New Place, a large house in Stratford, which as a boy he must have passed every day in walking to school.

It is not clear how his career in the theatre began; but from about 1594 onward he was an important member of the Lord Chamberlain's Company of players (called the King's Men after the accession of James I in 1603). They had the best actor, Richard Burbage; they had the best theatre, the Globe; they had the best dramatist, Shakespeare. It is no wonder that the company prospered. Shakespeare became a full-time professional man of his own theatre, sharing in a cooperative enterprise and intimately concerned with the financial success of the plays he wrote.

Unfortunately, written records give little indication of the way in which Shakespeare's professional life molded his marvellous artistry. All that can be deduced is that for 20 years Shakespeare devoted himself assiduously to his art, writing more than a million words of poetic drama of the highest quality.

Private life. Shakespeare had little contact with officialdom, apart from walking—dressed in the royal livery as a member of the King's Men—at the coronation of King James I in 1604. He continued to look after his financial interests. He bought properties in London and in Stratford. In 1605 he purchased a share (about one-fifth) of the Stratford tithes—a fact that explains why he was eventually buried in the chancel of its parish church. For some time he lodged with a French Huguenot family called Mountjoy, who lived near St. Olave's Church, Cripplegate, London. The records of a lawsuit in May 1612, due to a Mountjoy family quarrel, show Shakespeare as giving evidence in a genial way (though unable to remember certain important facts that would have decided the case) and as interesting himself generally in the family's affairs.

Shakespeare's financial interests

No letters written by Shakespeare have survived, but a private letter to him happened to get caught up with some official transactions of the town of Stratford and so has been preserved in the borough archives. It was written by one Richard Quiney and addressed by him from the Bell Inn in Carter Lane, London, whither he had gone from Stratford upon business. On one side of the paper is inscribed: "To my loving good friend and countryman, Mr. Wm. Shakespeare, deliver these." Apparently Quiney thought his fellow Stratfordian a person to whom he could apply for the loan of £30—a large sum in Elizabethan money. Nothing further is known about the transaction, but, because so few opportunities of seeing into Shakespeare's private life present themselves, this begging letter becomes a touching document. It is of some interest, moreover, that 18 years later Quiney's son Thomas became the husband of Judith, Shakespeare's second daughter.

Shakespeare's will (made on March 25, 1616) is a long and detailed document. It entailed his quite ample property on the male heirs of his elder daughter, Susanna. (Both his daughters were then married, one to the aforementioned Thomas Quiney and the other to John Hall, a respected physician of Stratford.) As an afterthought, he bequeathed his "second-best bed" to his wife; but no one can be certain what this notorious legacy means. The testator's signatures to the will are apparently in a shaky hand. Perhaps Shakespeare was already ill. He died on April 23, 1616. No name was inscribed on his gravestone in the chancel of the parish church of Stratford-upon-Avon. Instead these lines, possibly his own, appeared:

Good friend, for Jesus' sake forbear
To dig the dust enclosed here.
Blest be the man that spares these stones,
And curst be he that moves my bones.

EARLY POSTHUMOUS DOCUMENTATION

Shakespeare's family or friends, however, were not content with a simple gravestone, and, within a few years, a monument was erected on the chancel wall. It seems to have existed by 1623. Its epitaph, written in Latin and inscribed immediately below the bust, attributes to Shakespeare the worldly wisdom of Nestor, the genius of Socrates, and the poetic art of Virgil. This apparently was how his contemporaries in Stratford-on-Avon wished their fellow citizen to be remembered.

The tributes of his colleagues. The memory of Shakespeare survived long in theatrical circles, for his plays remained a major part of the repertory of the King's Men until the closing of the theatres in 1642. The greatest of Shakespeare's great contemporaries in the theatre, Ben Jonson, had a good deal to say about him. To William Drummond of Hawthornden in 1619 he said that Shakespeare "wanted art." But, when he came to write his splendid poem prefixed to the Folio edition of Shakespeare's plays in 1623, he rose to the occasion with stirring words of praise:

> Triumph, my Britain, thou hast one to show
> To whom all scenes of Europe homage owe.
> He was not of an age, but for all time!

Besides almost retracting his earlier gibe about Shakespeare's lack of art, he gives testimony that Shakespeare's personality was to be felt, by those who knew him, in his poetry—that the style was the man. Jonson also reminded his readers of the strong impression the plays had made upon Queen Elizabeth I and King James I at court performances:

> Sweet Swan of Avon, what a sight it were
>   To see thee in our waters yet appear,
> And make those flights upon the banks of Thames
>   That so did take Eliza and our James!

Popularity of the plays

Shakespeare seems to have been on affectionate terms with his theatre colleagues. His fellow actors John Heminge and Henry Condell (who, with Burbage, were remembered in his will) dedicated the First Folio of 1623 to the Earl of Pembroke and the Earl of Montgomery, explaining that they had collected the plays "... without ambition either of self-profit or fame; only to keep the memory of so worthy a friend and fellow alive as was our Shakespeare, ... "

Anecdotes and documents. Seventeenth-century antiquaries began to collect anecdotes about Shakespeare, but no serious life was written until 1709, when Nicholas Rowe tried to assemble information from all available sources with the aim of producing a connected narrative. There were local traditions at Stratford: witticisms and lampoons of local characters; scandalous stories of drunkenness and sexual escapades. About 1661 the Vicar of Stratford wrote in his diary: "Shakespeare, Drayton, and Ben Jonson had a merry meeting, and it seems drank too hard; for Shakespeare died of a fever there contracted." On the other hand, the antiquary John Aubrey wrote in some notes about Shakespeare: "He was not a company keeper; lived in Shoreditch; wouldn't be debauched, and, if invited to, writ he was in pain." Richard Davies, archdeacon of Lichfield, reported, "He died a papist." How much trust can be put in such a story is uncertain. In the early 18th century, a story appeared that Queen Elizabeth had obliged Shakespeare "to write a play of Sir John Falstaff in love" and that he had performed the task (*The Merry Wives of Windsor*) in a fortnight. There are other stories, all of uncertain authenticity and some mere fabrications.

When serious scholarship began in the 18th century, it was too late to gain anything from traditions. But documents began to be discovered. Shakespeare's will was found in 1747 and his marriage license in 1836. The documents relating to the Mountjoy lawsuit already mentioned were found and printed in 1910. It is possible that further documents of a legal nature may yet be discovered, but as time passes the hope becomes more remote. Modern scholarship is more concerned to study Shakespeare in relation to his social environment, both in Stratford and in London. This is not easy, because the author and actor lived a somewhat detached life: a respected tithe-owning country gentleman in Stratford, perhaps, but a rather rootless artist in London.

Portraits. Curiosity about what Shakespeare looked like has not been adequately satisfied. Two representations have indisputable claims to authenticity: a half-length bust in the Stratford-upon-Avon parish church, which was in position within a few years of his death; and an engraving used as a frontispiece to the Folio edition of his plays in 1623. The family must have approved the memorial statue, and presumably his friends must have thought the engraving a reasonable likeness. The bust shows Shakespeare with a pen in his hand; the face is rather round, but the high forehead is striking. The engraving is by Martin Droeshout, a member of a family of Flemish artists resident in London, and it has provided the image by which Shakespeare is now best known. The "Chandos" portrait (so called because the Duke of Chandos once owned it) in the National Portrait Gallery, London, seems to have been owned by Sir William Davenant. It resembles the bust and the engraving to a certain extent and could have been a portrait painted in Shakespeare's lifetime. Many other portraits alleged to be of Shakespeare began to appear in the later 17th century and in the 18th century. These are artists' "creations," fabrications, or downright forgeries.

The Droeshout engraving

The poet and dramatist

THE INTELLECTUAL BACKGROUND

Shakespeare lived at a time when ideas and social structures established in the Middle Ages still informed men's thought and behaviour. Queen Elizabeth I was God's deputy on earth, and lords and commons had their due places in society under her, with responsibilities up through her to God and down to those of more humble rank. The order of things, however, did not go unquestioned. Atheism was still considered a challenge to the beliefs and way of life of a majority of Elizabethans, but the Christian faith was no longer single—Rome's authority had been challenged by Martin Luther, John Calvin, a multitude of small religious sects, and, indeed, the English church itself. Royal prerogative was challenged in Parliament; the economic and social orders were disturbed by the rise of capitalism, by the redistribution of monastic lands under Henry VIII, by the expansion of education, and by the influx of new wealth from discovery of new lands.

An interplay of new and old ideas was typical of the time: official homilies exhorted the people to obedience, the Italian political theorist Niccolò Machiavelli was expounding a new practical code of politics that caused Englishmen to fear the Italian "Machiavillain" and yet prompted them to ask what men do, rather than what they should do. In *Hamlet,* disquisitions—on man, belief, a "rotten" state, and times "out of joint"—clearly reflect a growing disquiet and skepticism. The translation of Montaigne's *Essays* in 1603 gave further currency, range, and finesse to such thought, and Shakespeare was one of many who read them, making direct and significant quotations in *The Tempest.* In philosophical inquiry the question "how?" became the impulse for advance, rather than the traditional "why?" of Aristotle. Shakespeare's plays written between 1603 and 1606 unmistakably reflect a new, Jacobean distrust. James I, who, like Elizabeth, claimed divine authority, was far less able than she to maintain the authority of the throne. The so-called Gunpowder Plot (1605) showed a determined challenge by a small minority in the state; James's struggles with the House of Commons in successive Parliaments, in addition to indicating the strength of the "new men," also revealed the inadequacies of the administration.

POETIC CONVENTIONS AND DRAMATIC TRADITIONS

The Latin comedies of Plautus and Terence were familiar in Elizabethan schools and universities, and English translations or adaptations of them were occasionally performed by students. Seneca's rhetorical and sensational tragedies, too, had been translated and often imitated, both in structure and rhetoric. But there was also a strong native dramatic tradition deriving from the medieval miracle plays, which had continued to be performed in various towns until forbidden during Elizabeth's reign. This native drama had been able to assimilate French popular farce, clerically inspired morality plays on abstract themes, and interludes or short entertainments that made use of the "turns" of individual clowns and actors. Although Shakespeare's immediate predecessors were known as "university wits," their plays were seldom structured in the manner of those they had studied at Oxford or Cambridge; instead, they used and developed the more popular narrative forms. Their subplots, for example, amplified the main action and theme with a freedom and awareness of hierarchical correspondences that are medieval rather than classical.

The contemporary drama of Shakespeare's time

Changes in language. The English language at this time was changing and extending its range. The poet Edmund Spenser led with the restoration of old words, and schoolmasters, poets, sophisticated courtiers, and travellers all brought further contributions from France, Italy, and the Roman classics, as well as from farther afield. Helped by the growing availability of cheaper, printed books, the language began to become standardized in grammar and vocabulary and, more slowly, in spelling. Ambitious for a European and permanent reputation, the essayist and philosopher Francis Bacon wrote in Latin as well as in English; but, if he had lived only a few decades later, even he might have had total confidence in his own tongue.

Shakespeare's literary debts. In Shakespeare's earlier works his debts stand out clearly: to Plautus for the structure of *The Comedy of Errors;* to the poet Ovid and to Seneca for rhetoric and incident in *Titus Andronicus;* to morality drama for a scene in which a father mourns his dead son, and a son his father, in *Henry VI;* to Marlowe for sentiments and characterization in *Richard III* and *The Merchant of Venice;* to the Italian popular tradition of commedia dell'arte for characterization and dramatic style in *The Taming of the Shrew;* and so on. But he did not then reject these influences; rather, he made them his own, so that soon there was no line between their effects and his. In *The Tempest* (which is perhaps the most original of all his plays in form, theme, language, and setting) folk influences may also be traced, together with a newer and more obvious debt to a courtly diversion known as the masque, as developed by Ben Jonson and others at the court of King James.

THEATRICAL CONDITIONS

The Globe and its predecessor, the Theatre, were public playhouses run by the Chamberlain's Men (later the King's Men), a leading theatre company of which Shakespeare was a member. To these playhouses almost all classes of citizens, except the Puritans, came for afternoon entertainment. The players were also summoned to court, to perform before the monarch and assembled nobility. In the summer they toured the provinces, and on occasion they performed at London's Inns of Court (associations of law students), at universities, and in great houses. Popularity led to an insatiable demand for plays: repertories were always changing, so that early in 1613 the King's Men could present "fourteen several plays." The theatre soon became fashionable too, and in 1608–09 the King's Men started to perform on a regular basis at the Blackfriars, a "private" indoor theatre where high admission charges assured the company a more select and sophisticated audience for their performances.

Shakespeare's first associations with the Chamberlain's Men seem to have been as an actor. He is not known to have acted after 1603, and tradition gives him only secondary roles, such as the ghost in *Hamlet* and Adam in *As You Like It,* but his continuous association must have given him direct working knowledge of all aspects of theatre: like Aeschylus, Molière, Bertolt Brecht, or Harold Pinter, Shakespeare was able to work with his plays in rehearsal and performance and to know his actors and his audiences and all the different potentialities of theatres and their equipment. Numerous passages in Shakespeare's plays show conscious concern for theatre arts and audience reactions. Prospero in *The Tempest* speaks of the whole of life as a kind of "revels," or theatrical show, that, like a dream, will soon be over. The Duke of York in *Richard II* is conscious of how

> . . . in a theatre, the eyes of men,
> After a well-graced actor leaves the stage
> Are idly bent on him that enters next,
> Thinking his prattle to be tedious.

And Hamlet gives expert advice to visiting actors in the art of playing.

The Elizabethan playhouse

In Shakespeare's day, there was little time for group rehearsals, and actors were given the words of only their own parts. The crucial scenes in Shakespeare's plays, therefore, are between two or three characters only, or else are played with one character dominating a crowded stage. Female parts were written for young male actors or boys, so Shakespeare did not often write big roles for them or keep them actively engaged on stage for lengthy periods. Writing for the clowns of the company—who were important popular attractions in any play—presented the problem of allowing them to use their comic personalities and tricks and yet have them serve the immediate interests of theme and action.

Theatre is a collaborative art, only occasionally yielding the right conditions for individual genius to flourish and develop, and Shakespeare's achievement must at least in part be due to his continuous association with the Chamberlain's and King's Men, who were as practiced in acting together as he was to become in writing with their ensemble skills in mind.

CHRONOLOGY OF SHAKESPEARE'S PLAYS

Despite much scholarly argument, it is often impossible to date a given play precisely. But there is a general consensus, especially for plays written 1585–1601, 1605–07, and 1609 onward. The following list of first performances is based on external and internal evidence, on general stylistic and thematic considerations, and on the observation that an output of no more than two plays a year seems to have been established in those periods when dating is rather clearer than others.

1589–92	*1 Henry VI, 2 Henry VI, 3 Henry VI*
1592–93	*Richard III, The Comedy of Errors*
1593–94	*Titus Andronicus, The Taming of the Shrew*
1594–95	*The Two Gentlemen of Verona, Love's Labour's Lost, Romeo and Juliet*
1595–96	*Richard II, A Midsummer Night's Dream*
1596–97	*King John, The Merchant of Venice*
1597–98	*1 Henry IV, 2 Henry IV*
1598–99	*Much Ado About Nothing, Henry V*
1599–1600	*Julius Caesar, As You Like It*
1600–01	*Hamlet, The Merry Wives of Windsor*
1601–02	*Twelfth Night, Troilus and Cressida*
1602–03	*All's Well That Ends Well*
1604–05	*Measure For Measure, Othello*
1605–06	*King Lear, Macbeth*
1606–07	*Antony and Cleopatra*
1607–08	*Coriolanus, Timon of Athens*
1608–09	*Pericles*
1609–10	*Cymbeline*
1610–11	*Winter's Tale*
1611–12	*The Tempest*
1612–13	*Henry VIII, The Two Noble Kinsmen*

Shakespeare's two narrative poems, *Venus and Adonis* and *The Rape of Lucrece,* can be dated with certainty to the years when the Plague stopped dramatic performances in London, in 1592 and 1593–94, respectively, just before their publication. But the sonnets offer many and various problems; they cannot have been written all at one time, and most scholars set them within the period 1593–1600. "The Phoenix and the Turtle" can be dated 1600–01.

PUBLICATION

Pirated versions of plays

During Shakespeare's early career, dramatists invariably sold their plays to an actor's company, who then took charge of them, prepared working promptbooks, and did their best to prevent another company or a publisher from getting copies; in this way they could exploit the plays themselves for as long as they drew an audience. But some plays did get published, usually in small books called quartos. Occasionally plays were "pirated," the text being dictated by one or two disaffected actors from the company that had performed it or else made up from shorthand notes taken surreptitiously during performance and subsequently corrected during other performances; parts 2 and 3 of the *Henry VI* (1594 and 1595) and *Hamlet* (1603) quartos are examples of pirated, or "bad," texts. Sometimes an author's "foul papers" (his first complete draft) or his "fair" copy—or a transcript of either of these—got into a publisher's hands, and "good quartos" were printed from them, such as those of *Titus Andronicus* (1594), *Love's Labour's Lost* (1598), and *Richard II* (1597). After the publication of "bad" quartos of *Hamlet* and *Romeo and Juliet* (1597), the Chamberlain's Men probably arranged for the release of the "foul papers" so that second—"good"—quartos could supersede the garbled versions already on the market. This company had powerful friends at court, and in 1600 a special order was entered in the Stationers' Register to "stay" the publication of *As You Like It, Much Ado About Nothing,* and *Henry V,* possibly in order to assure that good texts were available. Subsequently *Henry V* (1600) was pirated, and *Much Ado About Nothing* was printed from "foul papers"; *As You Like It* did not appear in print until it was included in *Mr. William Shakespeares Comedies, Histories & Tragedies,* published in folio (the reference is to the size of page) by a syndicate in 1623 (later editions appearing in 1632 and 1663).

The only precedent for such a collected edition of public theatre plays in a handsome folio volume was Ben Jonson's collected plays of 1616. Shakespeare's folio included 36 plays, 22 of them appearing for the first time in a good text. (For the Third Folio reissue of 1664, *Pericles* was added from a quarto text of 1609, together with six apocryphal plays.) The First Folio texts were prepared by John Heminge and Henry Condell (two of Shakespeare's fellow sharers in the Chamberlain's, now the King's Men), who made every effort to present the volume worthily. Only about 230 copies of the First Folio are known to have survived.

The following list gives details of plays first published individually and indicates the authority for each substantive edition. Q stands for Quarto: Q2, Q3, Q4, etc., stand for reprints of an original quarto. F stands for the First Folio edition of 1623.

2 Henry VI Q 1594: a reported text. F from revised fair copies, edited with reference to Q.

Titus Andronicus Q 1594: from foul papers. F from a copy of Q, with additions from a manuscript that had been used as a promptbook.

3 Henry VI Q 1595: a reported text. F as for *2 Henry VI.*

Richard III Q 1597: a reconstructed text prepared for use as a promptbook. F from reprints of Q, edited with reference to foul papers and containing some 200 additional lines.

Love's Labour's Lost Q is lost. Q2 1598: from foul papers, and badly printed. F from Q2.

Romeo and Juliet Q 1597: a reported text. Q2 from foul papers, with some reference to Q. F from a reprint of Q2.

Richard II Q 1597: from foul papers and missing the abdication scene. Q4 1608, with reported version of missing scene. F from reprints of Q, but the abdication scene from an authoritative manuscript, probably the promptbook (of which traces appear elsewhere in F).

1 Henry IV Q 1598: from foul papers. F from Q5, with some literary editing.

A Midsummer Night's Dream Q 1600: from the author's fair copy. F from Q2, with some reference to a promptbook.

The Merchant of Venice Q 1600: from foul papers. F from Q, with some reference to a promptbook.

2 Henry IV Q 1600: from foul papers. F from Q, with reference to a promptbook.

Much Ado About Nothing Q 1600: from the author's fair papers. F from Q, with reference to a promptbook.

Henry V Q 1600: a reported text. F from foul papers (possibly of a second version of the play).

The Merry Wives of Windsor Q 1602: a reported (and abbreviated) text. F from a transcript, by Ralph Crane (scrivener of the King's Men), of a revised promptbook.

Hamlet Q 1603: a reported text, with reference to an earlier play. Q2 from foul papers, with reference to Q. F from Q2, with reference to a promptbook, with theatrical and authorial additions.

King Lear Q 1608: from an inadequate transcript of foul papers, with use made of a reported version. F from Q, collated with a promptbook of a shortened version.

Troilus and Cressida Q 1609: from a fair copy, possibly the author's. F from Q, with reference to foul papers, adding 45 lines and the Prologue.

Pericles Q 1609: a poor text, badly printed with both auditory and graphic errors.

Othello Q 1622: from a transcript of foul papers. F from Q, with corrections from another authorial version of the play.

The plays published for the first time in the First Folio of 1623 are:

All's Well That Ends Well From the author's fair papers, or a transcript of them.

Antony and Cleopatra From an authorial fair copy.

1 Henry VI

As You Like It From a promptbook, or a transcript of it.

The Comedy of Errors From foul papers.

Coriolanus From an authorial fair copy, edited for the printer.

Cymbeline From an authorial copy, or a transcript of such, imperfectly prepared as a promptbook.

Henry VIII From a transcript of a fair copy, made by the author, prepared for reading.

Julius Caesar From a transcript of a promptbook.

King John From an authorial fair copy.

Macbeth From a promptbook of a version prepared for court performance.

Measure for Measure From a transcript, by Ralph Crane, of very imperfect foul papers.

The Taming of the Shrew From foul papers.

The Tempest From an edited transcript, by Ralph Crane, of the author's papers.

Timon of Athens From foul papers, probably unfinished.

Twelfth Night From a promptbook, or a transcript of it.

The Two Gentlemen of Verona From a transcript, by Ralph Crane, of a promptbook, probably of a shortened version.

The Winter's Tale From a transcript, by Ralph Crane, probably from the author's fair copy.

The texts of *Venus and Adonis* (1593) and *The Rape of Lucrece* (1594) are remarkably free from errors. Shakespeare presumably furnished a fair copy of each for the printer. He also seems to have read the proofs. The sonnets were published in 1609, but there is no evidence that Shakespeare oversaw their publication (see below *Understanding Shakespeare: The contribution of textual criticism*).

POETIC AND DRAMATIC POWERS

The early poems. Shakespeare dedicated the poem *Venus and Adonis* to his patron, Henry Wriothesley, 3rd earl of Southampton, whom he further promised to honour with "some graver labour"—perhaps *The Rape of Lucrece,* which appeared a year later and was also dedicated to Southampton. As these two poems were something on which Shakespeare was intending to base his reputation with the public and to establish himself with his patron, they were displays of his virtuosity—diploma pieces. They were certainly the most popular of his writings with the reading public and impressed them with his poetic genius. Seven editions of *Venus and Adonis* had appeared by 1602 and 16 by 1640; *Lucrece,* a more serious poem, went through eight editions by 1640; and there are numerous allusions to them in the literature of the time. But after that, until the 19th century, they were little regarded. Even then the critics did not know what to make of them: on the one hand, *Venus and Adonis* is licentiously erotic (though its sensuality is often rather comic); while *Lucrece* may seem to be tragic enough, the treatment of the poem is yet somewhat cold and distant. In both cases the poet seems to be displaying dexterity rather than being "sincere." But Shakespeare's detachment from his subjects has come to be admired in more recent assessments.

Shakespeare's bid for fame

Above all, the poems give evidence for the growth of Shakespeare's imagination. *Venus and Adonis* is full of vivid imagery of the countryside; birds, beasts, the hunt, the sky, and the weather, the overflowing Avon—these give freshness to the poem and contrast strangely with the sensuous love scenes. *Lucrece* is more rhetorical and elaborate than *Venus and Adonis* and also aims higher. Its disquisitions (upon night, time, opportunity, and lust, for example) anticipate brilliant speeches on general themes in the plays—on mercy in *The Merchant of Venice,* suicide in *Hamlet,* and "degree" in *Troilus and Cressida.*

There are a few other poems attributed to Shakespeare. When the *Sonnets* were printed in 1609, a 329-line poem, "A Lovers complaint," was added at the end of the volume, plainly ascribed by the publisher to Shakespeare. There has been a good deal of discussion about the authorship of this poem. Only the evidence of style, however, could call into question the publisher's ascription, and this is conflicting. Parts of the poem and some lines are brilliant, but other parts seem poor in a way that is not like Shakespeare's careless writing. Its narrative structure is remarkable, however, and the poem deserves more attention than it usually receives. It is now generally thought to be from Shakespeare's pen, possibly an early poem revised by him at a more mature stage of his poetical style. Whether the poem in its extant form is later or earlier than *Venus and Adonis* and *Lucrece* cannot be decided. No one could doubt the authenticity of "The Phoenix and the Turtle," a 67-line poem that appeared with other "poetical essays" (by John Marston, George Chapman, and Ben Jonson) appended to Robert Chester's poem *Loves Martyr* in 1601. The poem is attractive and memorable, but very obscure, partly because of its style and partly because it contains allusions to real persons and situations whose identity can now only be guessed at.

The sonnets. In 1609 appeared *SHAKE-SPEARES SONNETS. Never before Imprinted.* At this date Shakespeare was already a successful author, a country gentleman, and an affluent member of the most important theatrical enterprise in London. How long before 1609 the sonnets were written is unknown. The phrase "never before imprinted" may imply that they had existed for some time but were now at last printed. Two of them (nos. 138 and 144) had in fact already appeared (in a slightly different form) in an anthology, *The Passionate Pilgrime* (1599). Shakespeare had certainly written some sonnets by 1598, for in that year Francis Meres, in a "survey" of literature, made reference to "his sugared sonnets among his private friends," but whether these "sugared sonnets" were those eventually published in 1609 cannot be ascertained—Shakespeare may have written other sets of sonnets, now lost. Nevertheless, the sonnets included in *The Passionate Pilgrime* are among his most striking and mature, so it is likely that most of the 154 sonnets that appeared in the 1609 printing belong to Shakespeare's early 30s rather than to his 40s—to the time when he was writing *Richard II* and *Romeo and Juliet* rather than when he was writing *King Lear* and *Antony and Cleopatra.* But, of course, some of them may belong to any year of Shakespeare's life as a poet before 1609.

Contemporary reference to Shakespeare's "sugared sonnets"

The order of the poems. Elizabethan sonnet sequences (following the example of their Italian and French models) were generally in some kind of narrative order. Sir Philip Sidney's *Astrophel and Stella,* moreover, and Edmund Spenser's *Amoretti* each tells a reasonably well authenticated story.

Shakespeare's sonnets, however, do not give the impression of an ordered sequence as it exists in Sidney, Spenser, and others. It is only at times that a narrative can be sensed, frequently breaking off, then resuming later or reverting to an earlier stage of the "story." It is therefore often argued that there was an original order, which has been lost, and many efforts have been made to recover, by ingenious analysis, Shakespeare's "intended" order. Although some interesting observations about associations between widely separated sonnets have been the result, none of the schemes carries any conviction. Most critics feel that it is hopeless to try to replace the order of the 1609 edition with anything convincingly better, arguing that the sonnets have no pretensions to be complete or adequate as a narrative. They are mixed in mood, in quality, and in distinction. Some seem open, addressed to all the world. Some seem too cryptic and personal ever to be intelligible.

It is equally uncertain whether the 1609 arrangement bears any relation to the order in which the sonnets were actually written. It may reasonably be supposed that the printer followed, more or less, the order in which the sonnets appeared in manuscript (or manuscripts). It is quite likely, however, that single leaves of the manuscript got out of order (for the numbers attached to each poem could well be a printer's addition), which, if so, might explain why some groups or pairs of sonnets seem to be oddly separated.

Sonnets 1 to 17 are variations on one theme. A handsome young man is being persuaded to marry and beget offspring who will preserve his beauty in a new generation, though he himself will lose it as he grows old. Gradually this theme gives place to the idea that the beloved will survive through the poet's verse. (There is no discussion of the relative merits of the two kinds of immortality.)

Sonnets 18 to 126 are on a variety of themes associated with a handsome young man (who is presumably, but not necessarily, the youth of 1 to 17). The poet enjoys his friendship and is full of admiration, promising to bestow immortality on the young man by the poems he writes in his honour. But sometimes the young man seems cold. Sometimes he provokes jealousy by his admiration of another poet. The climax of the series comes when the young man seduces the poet's woman. But eventually the poet reconciles himself to the situation and realizes that his love for his friend is greater than his desire to keep the woman.

The "dark lady" of the late sonnets

Sonnet 126 seems a kind of concluding poem (and is not in sonnet form). Then begins a new series, principally about a dark lady by whom the poet is enthralled, though well aware of her faults. At one point, she is stolen from him by his best friend. This faithlessness of both friend and woman wounds the poet deeply. He nevertheless tries to rise above his disappointment. (The two concluding sonnets are impersonal translations of a familiar Renaissance theme about Cupid.)

Artistic invention or real experience. Various persons are addressed or referred to in the poems, though whether they are real people or fictions of Shakespeare's dramatic imagination is not entirely clear. But if the "story" of the sonnets is an invention, then it is badly invented, showing nothing of the skill in storytelling that Shakespeare elsewhere reveals. The relationships between "characters," moreover, are so obscure, so irritatingly cryptic, that it is difficult to believe Shakespeare was devising the story for artistic purposes. The very clumsiness and obscurity of the narrative, if it is considered a fiction, are a strong argument that the sonnets are close to real experience; yet a degree of fictionalization in the sonnets would be in accord with Shakespeare's lifelong devotion to writing plays and with the pervasive "impersonality" of his art as a dramatist.

Some critics have been tempted to declare that it does not matter who the young man and the dark lady were—that the poetry is the important thing. Though as a gesture of impatience this mood is understandable, the world will continue to be curious about the circumstances of these poems, and, if historical documents should be discovered bringing hard facts to bear on the matter, the sonnets would surely be reinterpreted in the light of those facts. But for now, the sonnets on the whole retain an obstinate privacy that is a bar to enjoyment and therefore must be judged a fault, one that would hinder altogether the appreciation of any poems less brilliant. The sonnets do not quite "create a world" within which they can be apprehended. There is the sense of a missing (or unascertainable) body of experience and reference that falls short of poetical mystery.

Attempts to discover details of his life in the poems

Human experience in the poems. From the beginning of the 19th century, explorations of Shakespeare's personality have constantly been made by studying the sonnets. William Wordsworth proclaimed that "with this key Shakespeare unlocked his heart." But many readers feel that Shakespeare the man is elusive in the sonnets, just as he is in the plays. It has been natural to look in the poems for "personal details" about the author. One can observe allusions to his insomnia, to his disapproval of false hair and painted cheeks, to his love of music, and, according to some, to his bisexuality. It does not amount to very much.

The experiences of love and friendship, as related in the sonnets, must be described as disheartening and disenchanted. On the plane of human experience, they are full of disappointment, separation, anxiety, estrangement, self-accusation, and failure. The triumph, or near triumph, of death and of time is deeply felt. Only on the transcendental plane and in the faith in the permanence of poetry does a positive or affirmative attitude assert itself and compensate for the outcast state of the poet and the dateless night of death. On the whole they are quieter and closer to normal human experience than are the plays. The storms of passion are absent. Instead, there is a refined analysis of feeling, somehow more characteristic of the method of Jane Austen and Henry James than of Shakespeare the playwright. There are, of course, many moments of comparable moral scrupulousness in the plays, but they come incidentally and sometimes a little irrelevantly, perhaps, or awkwardly. In the sonnets Shakespeare brilliantly controls the shifting texture of the words in accordance with the variations in mood and tone. Generally, a careful analysis of the words reveals the tone in which a sonnet should be read, though, it must be admitted, this can be ambiguous.

The attractions of the sonnets are indeed very great. They win the admiration of readers by a variety of virtues. They express strong feeling, but they preserve artistic control. They have a density of thought and imagery that makes them seem the quintessence of the poetical experience. They delight by a felicity of phrase and verse movement, no less memorable than that familiar in the plays. They have, in recent years, received more exegesis than any of Shakespeare's plays, except perhaps *Hamlet* and *King Lear.* This is not necessarily a testimony to their value, for they have a tantalizing quality that encourages continued commentary. But it would be ungenerous not to relate this keen interest in the sonnets to an appreciation of their poetic power, of their unembarrassed exploration of intimate human relationships, and of their sensitivity to the tragedy of human aspirations and the triumph of time.

The early plays. Although the record of Shakespeare's early theatrical success is obscure, clearly the newcomer soon made himself felt. His brilliant two-part play on the Wars of the Roses, *The Whole Contention between the two Famous Houses, Lancaster and Yorke,* was among his earliest achievements. He showed, in *The Comedy of Errors,* how hilariously comic situations could be shot through with wonder and sentiment. In *Titus Andronicus* he scored a popular success with tragedy in the high Roman fashion. *The Two Gentlemen of Verona* was a new kind of romantic comedy. The world has never ceased to enjoy *The Taming of the Shrew. Love's Labour's Lost* is an experiment in witty and satirical observation of society. *Romeo and Juliet* combines and interconnects a tragic situation with comedy and gaiety. All this represents the probable achievement of Shakespeare's first half-dozen years as a writer for the London stage, perhaps by the time he had reached 30. It shows astonishing versatility and originality.

Henry VI, 1, 2, and 3. In *The Contention,* a two-part chronicle play (called in the First Folio *2 Henry VI; 3 Henry VI*), Shakespeare seems to have discovered the theatrical excitement that can be generated by representing recent history on the stage—events just beyond living memory but of great moment in the lives of present generations. The civil wars (popularly known as the Wars of the Roses) resulted from the struggle of two families, York and Lancaster, for the English throne. They had ended in 1485 with Richard III's defeat at the Battle of Bosworth, when Henry Tudor, as Henry VII, established a secure dynasty. Queen Elizabeth I was the granddaughter of Henry VII, so the story of York and Lancaster was of great interest to Shakespeare's contemporaries. In *2 Henry VI* the power struggle turns around the ineffective King Henry VI, until gradually the Duke of York emerges as contender for the throne. The climaxes of *3 Henry VI* include the murder of the Duke of York by the Lancastrians and, in the final scene, the murder of King Henry by Richard (York's son and the future Richard III). Shakespeare already showed himself a master of tragic poetry, notably in the speech of the captured York (wounded, mocked by a paper crown on his head, and awaiting death under the cruel taunts of Queen Margaret) and in the meditation of the King on the miseries of civil war. The vigorous and comic scenes with Jack Cade, a rebel leader, and his followers anticipate the kind of political comment that Shakespeare handled with greater subtlety in introducing the mobs of plebeians in *Julius Caesar* and *Coriolanus.*

His early mastery of tragic poetry

A third play, *1 Henry VI,* about the early part of the reign of King Henry VI, concerns events preceding the opening of the first part of *The Contention.* This is less successful, and it is uncertain whether it was a first effort at a historical play, written before *The Contention,* or a preparatory supplement to it, written subsequently and less inspired. It was printed in the 1623 Folio as the first part of *King Henry VI; The Contention* appeared as the second and third parts of *King Henry VI,* on what authority is not known.

The Comedy of Errors. The title of this, Shakespeare's shortest play, speaks for itself (though the opening scene is, unexpectedly, full of pathos). The play is based on Plautus' *Menaechmi,* a play of the comic confusions deriving from the presence of twin brothers, unknown to each other, in the same town; but Shakespeare has added twin servants, and he fills the play with suspense, surprise, expectation, and exhilaration as the two pairs weave their way through quadruple misunderstandings. The play already reveals Shakespeare's mastery of construction.

Titus Andronicus. This play was highly popular and held the stage for many years. Its crude story, its many savage incidents, and its poetic style have led some critics to think it not by Shakespeare. But the tendency of recent criticism is to regard the play as wholly or essentially his, and, indeed, when considered on its own terms as a "Roman tragedy," it displays a uniformity of tone and reveals a consistency of dramatic structure as a picture of the decline of the ancient world (though, by the standards of Shakespeare's later Roman plays, this picture is much confused).

The Two Gentlemen of Verona. Shakespeare took this play's story from a long Spanish prose romance called *Diana,* by Jorge de Montemayor. He added new characters—including Valentine, one of the "Two Gentlemen," whose "ideal" friendship with Proteus is so developed that the plot is more than a love story; indeed, the play glorifies friendship to an extent that, by modern conventions, is absurd. The abrupt last scene suggests that something has gone wrong with the text, and certainly Shakespeare was never again so ready to abandon common sense in motivating the behaviour of his lovers. But it is also clear that Shakespeare is here feeling his way toward a new kind of high comedy, later to find expression in *The Merchant of Venice* and *Twelfth Night.*

The Taming of the Shrew. Often played as a boisterous farce, this play is actually a comedy of character, with implications beyond the story of the wooing, wedding, and taming of Katharina, the "shrew," by Petruchio, a man with a stronger will than her own. Shakespeare arouses more interest in these two than farce permits. They gain, for example, by contrast with the tepid, silly, or infatuated lovers (Bianca, Lucentio, Hortensio, and Gremio), and their relationship is given an admirable vitality and energy; while in the play's last scene Katharina's discourse on wifely submission—if spoken with sincerity and genuine tenderness and without irony—has a moving quality in performance. The Italianate play about the shrew taming is set inside another play (concerning a trick played upon Christopher Sly, a drunken tinker), which gave Shakespeare an opportunity for some brilliant English country scenes. Originally, Sly was made the "audience" of the shrew play, a device that is abandoned after a little while (that is, in the text of the 1623 Folio). Probably the players' company came to abandon the Christopher Sly framing because the Katharina and Petruchio story was too strong not to be acted directly at the real audience in a theatre.

Early version of *The Taming of the Shrew*

Love's Labour's Lost. Once regarded as obsolete, depending too much upon temporary and irrecoverable allusions, this play has come to life in the theatre only during the past 50 years. Its rejection by the theatre was a background for 18th- and 19th-century critics such as John Dryden, Dr. Johnson, and William Hazlitt, all of whom had severe things to say about it. But, once the play had been recovered for the theatre, it was discovered that it is full of humanity, exploring the consequences of man being made of flesh and blood. The central comic device is that of four young men, dedicated to study and to the renunciation of women, meeting four young women;

inevitably they abandon their absurd principles. For variety, and as an escape from the pretty, gay, young royalty and courtiers, there is an entertaining band of eccentrics who are allowed their "vaudeville" turns: Sir Nathaniel the curate, Holofernes the schoolmaster, Dull the constable, Costard the clown, and Jaquenetta the country girl; linking both groups is Don Adriano de Armado the ineffable (who begins by being a bore but becomes interesting as he becomes pathetic). Toward the end, the play takes on a new dramatic vitality through a brilliant *coup de théâtre:* the sudden arrival of Mercade as the messenger of death and the herald of responsibility. The deliberate abstention from the customary conclusion of comedy is remarkable: "Jack hath not Jill," but he will have her after a twelvemonth, when he has done something to deserve her. Thus the play ends with hope—perhaps the best kind of happy ending.

Romeo and Juliet. The most complex work of art among these early plays of Shakespeare, *Romeo and Juliet* is far more than "a play of young love" or "the world's typical love-tragedy." Weaving together a large number of related impressions and judgments, it is as much about hate as love. It tells of a family and its home as well as a feud and a tragic marriage. The public life of Verona and the private lives of the Veronese make up the setting for the love of Juliet and Romeo and provide the background against which their love can be assessed. It is not the deaths of the lovers that conclude the play but the public revelation of what has happened, with the admonitions of the Prince and the reconciliation of the two families.

Shakespeare enriched an already old story by surrounding the guileless mutual passion of Romeo and Juliet with the mature bawdry of the other characters—the Capulet servants Sampson and Gregory open the play with their fantasies of exploits with the Montague women; the tongues of the Nurse and Mercutio are seldom free from sexual matters—but the innocence of the lovers is unimpaired.

Romeo and Juliet made a strong impression on contemporary audiences. It was also one of Shakespeare's first plays to be pirated; a very bad text appeared in 1597. Detestable though it is, this version does derive from a performance of the play, and a good deal of what was seen on stage was recorded. Two years later another version of the play appeared, issued by a different, more respectable publisher, and this is essentially the play known today, for the printer was working from a manuscript fairly close to Shakespeare's own. Yet in neither edition did Shakespeare's name appear on the title page, and it was only with the publication of *Love's Labour's Lost* that publishers had come to feel that the name of Shakespeare as a dramatist, as well as the public esteem of the company of actors to which he belonged, could make an impression on potential purchasers of playbooks.

Shakespeare's sources for the history plays

The histories. For his plays on subjects from English history, Shakespeare primarily drew upon Raphael Holinshed's *Chronicles,* which appeared in 1587, and on Edward Hall's earlier account of *The union of the two noble and illustre famelies of Lancastre and York* (1548). From these and numerous secondary sources he inherited traditional themes: the divine right of royal succession, the need for unity and order in the realm, the evil of dissension and treason, the cruelty and hardship of war, the power of money to corrupt, the strength of family ties, the need for human understanding and careful calculation, and the power of God's providence, which protected his followers, punished evil, and led England toward the stability of Tudor rule.

The Tragedy of King Richard III. In this play, the first history to have a self-contained narrative unity, Shakespeare accentuated the moment of death as a crisis of conscience in which man judges himself and is capable of true prophecy. He centred the drama on a single figure who commits himself to murder, treason, and dissimulation with an inventive imagination that an audience can relish even as it must condemn it; and in defeat Richard discovers a valiant fury that carries him beyond nightmare fear and guilt to unrepentant, crazed defiance.

The Tragedy of King Richard II. In the group of histories written in the late 1590s, Shakespeare developed themes similar to those of *Richard III* but introduced counter-statements, challenging contrasts, and more deeply realized characters. The first of this group, *Richard II,* concentrates on the life and death of the King, but Bolingbroke, his adversary, is made far more prominent than Richmond had been as Richard III's adversary. The rightful king is isolated and defeated by Act III, and in prison he hammers out the meaning of his life in sustained soliloquy and comes to recognize his guilt and responsibility. From this moment of truth, he rediscovers pride, trust, and courage, so that he dies with an access of strength and an aspiring spirit. After the death of Richard, a scene shows Bolingbroke, now Henry IV, with the corpse of his rival in a coffin; and then Bolingbroke, too, recognizes his own guilt, as he sits in power among his silent nobles.

1 Henry IV; 2 Henry IV. In the two plays that bear his name, Henry IV is often in the background. The stage is chiefly dominated by his son, Prince Hal (later Henry V), by Hotspur the young rebel, and by Sir John Falstaff. The secondary characters are numerous, varying from prostitutes and country bumpkins to a Lord Chief Justice and country gentlemen. There is a tension underlying the two-part play that sounds in the King's opening lines:

> So shaken as we are, so wan with care,
> Find we a time for frighted peace to pant,
> And breathe short-winded accents of new broils
> To be commenced in stronds afar remote.

When the Earl of Warwick counsels hope, the King sees how "chances mock, and changes fill the cup of alteration":

> O, if this were seen,
> The happiest youth, viewing his progress through,
> What perils passed, what crosses to ensue,
> Would shut the book, and sit him down and die.
> (Pt. 2, Act III, scene 1, 51–56)

Yet the two plays of *Henry IV* are full of energy. In Falstaff—that "reverend vice . . . that father ruffian, that vanity in years"—Shakespeare has created a character who becomes a substitute father of license and good fellowship for Prince Hal and who comments on the political situation with inglorious, reckless, egotistical good sense.

Falstaff's characterization

Falstaff is Shakespeare's major introduction into English history. His characterization is wholly original, for, although Shakespeare uses something of the earlier "vice" figure from early tragedies and comedies, something of the glutton and coward from allegorical, or morality, plays, something of the braggart soldier and the impotent old lecher from neoclassical comedy, he also studied life for this character of an out-of-work soldier, a knight without lands or alliances, a childless man whose imagination far outruns his achievement.

King John. Already in *King John,* Shakespeare had developed a subsidiary character to offset kings and princes. Here the Bastard, the son of Sir Robert Faulconbridge, is a supporter of the King and yet has soliloquies, asides, and speeches that mock political and moral pretensions. King John provides the central focus of the play, which ends with his death, but Shakespeare presents him on a rapidly changing course, surrounded by many contrasting characters—each able to influence him, each bringing irresolvable and individual problems into dramatic focus—so that the King's unsteady mind seems no more than one small element in an almost comic jumble of events.

Henry V. Henry V is the last of this group of history plays and the last until *Henry VIII* at the end of Shakespeare's career. Structurally the King is again central and dominant, but the subsidiary characters far outnumber those of the earlier plays. In the first two acts Henry is shown in peace and war, politic, angry, confident, sarcastic, and then vowing to weep for another man's revolt. There is an account of Falstaff's death, and, after scenes of military achievement, there is a nervous watch before the Battle of Agincourt when the King walks disguised among his fearful soldiers and prays for victory while acknowledging his own worthless repentance for his father's treason. The presentation of the battle avoids almost all fighting on stage, but recruits, professional soldiers, and dukes and princes are all shown making preparation for meeting defeat or victory. The King's speech to his troops before battle on St. Crispin's Day is famous for its evocation of a brotherhood in arms, but Shakespeare has placed it in a context full of ironies and challenging contrasts. The picture of two nations at war is full of deeply felt individual responses: a boy comes to realize that his masters are cowards; a herald is almost at a loss for words; a common soldier has to justify his heart before his king. Just before the conclusion, the King woos the French princess; she knows almost no English, and so he is forced to plead on the merits only of his simplest words, a kiss, and the plain fact of himself and his "heart." There is no doubt that Kate marries Henry because he has won a battle and peace is necessary, but Shakespeare has developed the comedy and earnestness of their wooing so that the need for human trust and acceptance is also evident. This play is presented by a chorus, which speaks in terms of heroism, pride, excitement, fear, and national glory. But at the end, when the King prays that the oaths of marriage and peace may "well kept and prosperous be," the chorus speaks for the last time, reminding the audience that England was to be plunged into civil war during the reign of Henry V's son. The last of a great series of history plays thus concludes with a reminder that no man's story can bring lasting success.

The Roman plays. *Julius Caesar.* After the last group of English history plays, Shakespeare chose to write about Julius Caesar, who held particular fascination for the Elizabethans. He was soldier, scholar, and politician (Francis Bacon held him in special regard for the universality of his genius); he had been killed by his greatest friend (Shakespeare alluded to the "bastard hand" of Brutus in *2 Henry VI*); and he was seen as the first Roman to perceive and, in part, to achieve the benefits of a monarchial state. His biography had appeared in Sir Thomas North's translation, via a French version, of Plutarch's *Parallel Lives,* published as *The Lives of the noble Grecians and Romanes* in 1579, which Shakespeare certainly read. To all of this, Shakespeare's response was surprising: Caesar appears in three scenes and then is murdered, before the play is half finished. But a variety of characters respond to and reflect upon the central fact of the great man. This is the dramatic strategy of an ironist, or of a writer who wishes to question human behaviour and to observe interactions and consequences. In fact, Caesar influences the whole play, for he appears after his death as a bloodstained corpse and as a ghost before battle. Both Brutus and Cassius die conscious of Caesar and even speak to him as if he were present. And then his heir takes command, to "part the glories" of what is for him a "happy day."

The Elizabethans' fascination with the figure of Julius Caesar

In other ways *Julius Caesar* is shaped differently from the histories and tragedies that precede it, as if in manner as in subject matter Shakespeare was making decisive changes. The scene moves only from Rome to the battlefield, and with this new setting language becomes more restrained, firmer and sharper. Extensive descriptive images are few, and single words such as "Roman," "honour," "love," "friend," and proper names are repeated as if to enforce contrasts and ironies. In performance this sharp verbal edge, linked with commanding performances and the various excitements of debate, conspiracy, private crises, political eloquence, mob violence, supernatural portents, personal antagonisms, battle, and deaths, holds attention. The play has popular appeal and intellectual fascination.

For six or seven years Shakespeare did not return to a Roman theme, but, after completing *Macbeth* and *King Lear,* he again used Plutarch as a source for two more Roman plays, both tragedies that seem as much concerned to depict the broad context of history as to present tragic heroes.

Antony and Cleopatra. The language of *Antony and Cleopatra* is sensuous, imaginative, and vigorous. "*Feliciter audax* ["happily bold"] is the motto for its style comparatively with his other works," said the poet Samuel Taylor Coleridge. Almost every character seems to talk of kingdoms and to envision heroic deeds: Dolabella, the Roman soldier, says that his "love makes religion to obey" Cleopatra in her last imprisonment; Antony's servant is called Eros, who kills himself before his "great chief"; his soldiers have seen his eyes "glow like plated Mars"; his enemies say that, even in defeat, he "continues still a Jove." Octavius knows, as he closes for the kill, that great issues are at stake. Yet, while the issues are thus enlarged (or inflated), the protagonists do not reveal themselves to an audience as they do in *Hamlet, Othello, Lear,* or *Macbeth.* Antony has soliloquies only in defeat, and then he addresses the sun or fortune, false hearts or his queen, rather than seeking to hammer out his thoughts or to explore his own response. In the last scene, however, the focus concentrates intensely on a single character, when Cleopatra, prepared for death in robe and crown, believing in immortality, and hearing the dead Antony mock "the luck of Caesar," seems indeed to be transfigured:

The tragic intensity of Cleopatra's death scene

> . . . Husband, I come:
> Now to that name my courage prove my title!
> I am fire, and air; my other elements
> I give to baser life. . . .
>
> (Act IV, scene 2, 278 ff.)

Coriolanus. The hero of *Coriolanus* has still fewer soliloquies: one in rhyme as he arrives, a renegade in Antium, and another of a few sentences when he stands alone in the marketplace waiting for the citizens to express their choice of him for consul. Through many guises, the audience sees the same man: as a young nobleman in peacetime, as a soldier going to battle, as bloodstained fighter and then victor, as candidate for consul in the "napless garment of humility," as a banished renegade, and then as leader of the Volscians, enemies of Rome.

In this play Shakespeare's customary contrasts and ironies, which lead an audience to discover meanings for themselves, are replaced by repetitions within the single narrative line: there are three Forum scenes, four family scenes, a succession of fights, four scenes of mob violence, and continual attempts to argue with Coriolanus and deflect him from his chosen course. The language is sometimes elaborate, but does not have the poetic richness of *Antony and Cleopatra;* images are compact, sharply effective. Those moments when the audience is drawn most intently into the drama are strangely silent or understated. The citizens "steal away" instead of volunteering to fight for their country; their tribunes stay behind to organize their own responses. When the banished Coriolanus returns at the head of an opposing army, he says little to Menenius, the trusted family friend and politician, or to Volumnia, his mother, who have come to plead for Rome. His mother's argument is long and sustained, and for more than 50 lines he listens silent, until his resolution is broken from within: then, as a stage direction in the original edition testifies, he "holds her by the hand, silent." In his own words, he has "obeyed instinct" and betrayed his dependence; he cannot

> stand
> As if a man were author of himself
> And knew no other kin.
>
> (Act V, scene 3, 35 ff.)

His desire for revenge is defeated, and the army retreats. Volumnia is hailed as "patronness and life of Rome," but she is silent while the drums, trumpets, and voices greet her. Coriolanus is seen only once more, in the enemy city where he is accused of treachery and is assassinated in the kind of mob violence he has previously withstood. So *Coriolanus* finishes with the audience observing a hero helpless to prevent his own death and with patterns of political, social, and personal behaviour repeated without hope of change. Nowhere has Shakespeare shown an aspect so severe as in the silent moments of this Roman tragedy, during which the actor is at a loss for words.

The "great," or "middle," comedies. The comedies written between 1596 and 1602 have much in common and are as well considered together as individually. With the exception of *The Merry Wives of Windsor,* all are set in some "imaginary" country. Whether called Illyria, Messina, Venice and Belmont, Athens, or the Forest of Arden, the sun shines as the dramatist wills. A lioness, snakes, magic caskets, fairy spells, identical twins, disguise of sex, the sudden conversion of a tyrannous duke or the defeat offstage of a treacherous brother can all change the course of the plot and bring the characters to a conclusion in which almost all are happy and just deserts are found. Lovers are young and witty and almost always rich. The action concerns wooing; and its conclusion is marriage, beyond which the audience is scarcely concerned. Whether Shakespeare's source was an Italian novel (*The Merchant of Venice* and *Much Ado About Nothing*), an English pastoral tale (*As You Like It*), an Italian comedy (the Malvolio story in *Twelfth Night*), or something of his own invention (probably *A Midsummer Night's Dream,* and parts of each), always in his hands story and sentiments are instinct with idealism and capable of magic transformations. *The Merry Wives of Windsor* differs from the other comedies in that it is set not in an imaginary country but in Windsor and the rural life of Shakespeare's own day. Fantasy occurs at the end, however, when the characters enter a land of make-believe around the folk fertility symbol of Herne, the Hunter's oak in the forest; and, as they leave to "laugh this sport o'er by a country fire," quarrels are forgotten. The more overtly fantastic plays, in their turn, contain observations of ordinary and (usually) country life.

The blend of fantasy and everyday experience in the comedies

Much Ado About Nothing is distinctive in many ways, for there is no obvious magic or disguise of sex. But many misunderstandings arise in a masked dance, and the play concludes with a pretended death and a simulated resurrection from behind yet another mask; moreover, the two main characters, Beatrice and Benedick, are transformed from within by what is called "Cupid, the only matchmaker."

In some ways these are intellectual plays. Each comedy has a multiple plot and moves from one set of characters to another, between whom Shakespeare invites his audience to seek connections and explanations. Despite very different classes of people (or immortals) in different strands of the narrative, the plays are unified by Shakespeare's idealistic vision and by an implicit judgment of human relationships, and all their characters are brought together—with certain significant exceptions—at, or near, the end.

The "outsider." The plays affirm truth, good order, and generosity, without any direct statement; they are shapely and complicated like a dance or like a game of chess. Yet at the general resolution all is not harmony: some characters are held back from full participation. At the end of *A Midsummer Night's Dream,* for example, the lovers' more callow comments on the rustics' play of *Pyramus and Thisbe* mark them as irresponsive to the imaginative world of Bottom and his fellows, who project themselves into their play's heroics almost without fear of failure; they are also distinct from the duke, Theseus, who says of the amateur performers:

> The best in this kind are but shadows: and the worst are no worse, if imagination amend them.

In *The Merchant of Venice,* most of the unresolved elements in the comedy are concentrated in the person of Shylock, a Jew who attempts to use justice to enforce a terrible, murderous revenge on Antonio, the Christian merchant, but is foiled by Portia, in disguise as a lawyer, who turns the tables on the Jew by a legal quibble and has him at the mercy of the court. This strange tale is realized with exceptionally credible detail: Shylock is a moneylender, like many in Shakespeare's London, and a Jew of pride and deep religious instincts; the Christians treat him with contempt and distrust, and, when one of them causes his daughter to elope and steal his money and jewels, he suffers with an intensity equalled only by that of his murderous hatred of all Christians. In a scene written in prose that gives at one time both histrionic power and sensitive, personal feeling, Shylock identifies his cause with basic human rights:

Shylock

> I am a Jew . . . Hath not a Jew eyes? hath not a Jew hands, organs, dimensions, senses, affections, passions? fed with the same food, hurt with the same weapons, subject to the same diseases, healed by the same means, warmed and cooled by the same winter and summer, as a Christian is? If you prick us, do we not bleed? if you tickle us, do we not laugh? if you poison us, do we not die? and if you wrong us, shall we not revenge?
>
> (Act III, scene 1, 50 ff.)

The happiness that follows the thwarting of his revenge, however, cannot be celebrated with full unison while there are reminders that he never reached "the beautiful mountain," Belmont.

In *As You Like It,* the melancholy character Jaques leaves the play before the concluding dances. He has seen, and voiced, the limitations of each of the pairs of lovers and is determined to hear more and learn more from the tyrant duke, so mysteriously converted to a "religious life." In *Twelfth Night,* Malvolio the steward is gulled by practical jokes that take advantage of his self-esteem; he is the last character to come onstage in the final scene, and he refuses reconciliation; he leaves after a noticeable silence with only these words: "I'll be revenged on the whole pack of you." This comedy has yet other "outsiders"; alone at the end, Feste the fool sings a strange song in which the whole of life is reduced to a melancholy tale sung by a knowing idiot. *Twelfth Night* is probably the last of the "great" comedies, and it is the saddest.

Wit and ambiguity. Incidental images, too, strike deep into the audience's remembrance of pain, fear, and suffering. In *Twelfth Night,* barren mountains, salt sea, and the smoke of war, storms, imprisonment, death, and madness are all invoked. When the king and queen of fairies quarrel in *A Midsummer Night's Dream,* Titania's speech evokes a world of chaos:

> The ox hath therefore stretched his yoke in vain,
> The plowman lost his sweat, and the green corn
> Hath rotted ere his youth attained a beard:
> The fold stands empty in the drowned field,
> And crows are fatted with the murrion flock . . .
>
> (Act II, scene 1, 81 ff.)

Yet the poetry also celebrates happiness and joy with a clearer note and quicker interplay of thought than had been achieved in English before. So, in *Twelfth Night,* from Orsino:

> Then let thy love be younger than thyself,
> Or thy affection cannot hold the bent:
> For women are as roses, whose fair flower,
> Being once displayed doth fall that very hour.

From Olivia:

> Cesario, by the roses of spring,
> By maidhood, honour, truth, and everything,
> I love thee so that, maugre all thy pride,
> Nor wit nor reason can my passion hide.

And from Viola:

> And all those sayings will I over-swear,
> And all those swearings keep as true in soul,
> As doth that orbed continent the fire
> That severs day from night.

The wit and ambiguity through which the apparent meanings are laced with further indications of love, yearning, sexuality, unrest, and happiness glance continuously through the dialogue, especially perhaps in *Much Ado About Nothing.* Often the disguise of identity or sex or else misunderstandings or intentional counterfeiting serve to accentuate the varying levels of consciousness expressed; moments of near unmasking, or of recognition, hold the attention firmly. Witty debates between lovers were not unfamiliar; they had held the stage more than 10 years earlier in John Lyly's fantastic comedies that had been played at court by very young child actors. But never before had sharpness of wit been so matched by gentleness or fineness of sentiment.

Perhaps the most extraordinary achievement of these comedies—which change in mood so rapidly, which are so funny and yet sometimes dangerous and sad, which deal both with fantasy and eloquence—is that the recurrent moments of lifelike feeling are so expressed in words or action that an audience shares in the very moment of discovery. Sometimes this is a second thought, as in Viola's

> I am all the daughters of my father's house,
> And all the brothers too . . . and yet I know not . . .

Sometimes it is a single phrase—Sir Andrew Aguecheek's "I was adored once too"—and on several climactic occasions it is held in a song, whether of happiness, sorrow or peace, or of the "good life":

> O mistress mine, where are you roaming?
> O, stay and hear, your true-love's coming,
>   That can sing both high and low.

The "fancy" of art and sentiment in the great comedies

The structure of these comedies can be explained, their stage devices and language analyzed; but essentially they remain on the wing, alive with the "fancy" of art and sentiment. Puck's epilogue to the *Dream* is often quoted as their characteristic note:

> If we shadows have offended,
> Think but this, and all is mended,
> That you have but slumbered here
> While these visions did appear . . .

But Rosalind's epilogue to *As You Like It* is also apposite:

> . . . I am not furnished like a beggar; therefore to
> beg will not become me: my way is to conjure you . . .

The great tragedies. It is a usual and reasonable opinion that Shakespeare's greatness is nowhere more visible than in the series of tragedies—*Hamlet, Othello, King Lear,* and *Macbeth. Julius Caesar,* which was written before these, and *Antony and Cleopatra* and *Coriolanus,* which were written after, have many links with the four. But, because of their rather strict relationship with the historical materials, they are best dealt with in a group by themselves. *Timon of Athens,* probably written after the above-named seven plays, shows signs of having been unfinished or abandoned by Shakespeare. It has its own splendours but has rarely been considered equal in achievement to the other tragedies of Shakespeare's maturity.

Hamlet. Judged by its reception by the civilized world, *Hamlet* must be regarded as Shakespeare's most successful play. It has unceasing theatrical vitality, and the character of Hamlet himself has become a figure of literary mythology. Yet *King Lear* became for a time the fashionable play in 20th-century criticism, with many critics arguing that in *Hamlet* Shakespeare did not make a psychologically consistent play out of a plot that retained much of the crudity of an earlier kind of "revenge" drama—that he was trying to transform a barbaric "revenge" hero into a subtle Renaissance prince but did not succeed.

Even if this opinion has become unacceptable, it nevertheless taught critics to look for elements other than psychological consistency. In particular, it is worth concentrating not on Elizabethan attitudes toward revenge but on Shakespeare's artistic balance in presenting the play's moral problems. It is likely that an artist will make his work more interesting if he leaves a dilemma morally ambiguous rather than explicit. The revenge situation in *Hamlet,* moreover, is one charged with emotional excitement as well as moral interest. Simply put, the good man (Hamlet) is weak, and the bad man (Claudius) is strong. The good man has suffered a deep injury from the bad man, and he cannot obtain justice because justice is in the hand of the strong bad man. Therefore the weak good man must go around and around in order to achieve a kind of natural justice; and the audience watches in suspense while the weak good man by subtlety attacks and gets his own back upon the strong bad man and the strong bad man spends his time evading the weak good man. Hamlet is given a formidable opponent: Claudius is a hypocrite, but he is a successful one. He achieves his desired effect on everybody. His hypocrisy is that of a skilled politician. He is not dramatically shown as being in any way unworthy of his station—he upholds his part with dignity. He is a "smiling villain" and is not exposed until the final catastrophe. The jealous Hamlet heaps abuse upon him, but Shakespeare makes Claudius the murderer self-controlled. Thus, theatrically, the situation is much more exciting.

Against this powerful opponent is pitted Hamlet, the witty intellectual. He shares his wit with the audience (and a few favoured characters such as Horatio), who thus share his superiority over most other personages in the play. His first words are a punning aside to the audience, and his first reply to the King is a cryptic retort. His sardonic witticisms are unforgettable ("The funeral baked meats/ Did coldly furnish forth the marriage tables"; and "More honoured in the breach than the observance"). Hamlet is an actor in many parts of the play. The range of language in the roles he affects shows that his mimetic powers are considerable. He is skillful in putting on "an antic disposition" and gives a very funny performance in talking to Polonius. He condescends to talk the silly bawdry of Rosencrantz and Guildenstern. He can mimic Osric's style to perfection. He quarrels with Laertes beside Ophelia's grave in a display of verbosity that exceeds the modesty of nature in much the same way as does that of Laertes.

Hamlet himself as an actor

Besides Claudius, set off against Hamlet is Polonius. He is wrong in his judgments, one after another, and this leads to the audience's rejection of his political and human values. In all circumstances he seems slightly ridiculous—a foil for Claudius as he is for Hamlet. His astuteness suffers by comparison with that of the King. His philosophical view of life is hollow compared with Hamlet's. Hamlet has as many general maxims as Polonius; but his seem to be the product of a far more refined sensibility and of an ability to respond truthfully to experience. It is these qualities in the somewhat enigmatic characterization of Hamlet that have won him the fascinated admiration of the world.

Othello. Trusting to false appearances and allowing one's reason to be guided by one's passions had been a theme of many of Shakespeare's comedies. In *Othello* he showed that the consequences of so doing can be tragic rather than comic. Shakespeare adapted the story from an Italian model. His principal innovation consisted in developing the character of Iago, the villain, whose motives are represented as complex and ambiguous. Clearly Shakespeare was keenly interested in a villain who could successfully preserve an appearance of honesty; the bad as well as the good can be "the lords and owners of their faces" (Sonnet 94). Iago is made a plausible villain by being so interesting. He is an actor who enjoys playing his role of "honesty." Shakespeare makes him take the audience into his confidence at every stage of his plotting, and, as a consequence, they have a kind of non-moral participation in his villainy.

The pure and deep love between Desdemona and Othello is stressed from the beginning. Again and again the moral and intellectual stature of Othello is elevated by Shakespeare. He quells tumults in the streets with a few words; he bears himself with dignity before the Venetian council, defending himself compellingly from bitter accusations by Brabantio and accepting his military burden with quiet confidence. Even Iago, in the opening scene of the play, grudgingly admits the dependence of the Venetians on his valour. After his terrible murder of Desdemona, Othello's contrition is agonizing enough to swing the sympathies of the audience back to him.

King Lear. For Shakespeare's contemporaries, Lear, king of Britain, was thought to have been a historical monarch. For Shakespeare, however, although he gave the play something of a chronicle structure, the interest lay not in political events but in the personal character of the King. The main theme of the play is put into the mouth of the evil Regan, speaking to the pitiful Gloucester:

> O sir, to wilful men
> The injuries that they themselves procure
> Must be their schoolmasters.
>
> (Act II, scene 4, 301)

The various stages of Lear's spiritual progress (a kind of "conversion") are carefully marked. He learns the value of patience and the worth of "unaccommodated man." He begins to realize his own faults as a king and almost understands his failure as a father. He begins to feel for the "poor naked wretches" and confesses, "O I have ta'en too little care of this." His initial instability of mind, almost a predisposition to madness, is shown from the beginning. His terrible rages and curses, first upon Cordelia and later upon Goneril and Regan, and his ranting and tyrannical language all foreshadow his breakdown. His faithful counsellor is plain with him: "Be Kent unmannerly/When Lear is mad"; and his daughters shrewdly judge him: "he hath ever but slenderly known himself." He is painfully conscious of approaching madness, but gradually the bombast of his sanity gives place to a remarkable kind of eloquence, flowing easily and never incoherent. His "ravings" are intelligible to the audience, however perturbing they may seem to the other characters on the stage. They express a point of view that, had he understood it earlier, would have saved him from many errors of judgment. The mode of speech of the mad King contrasts strongly with the

Lear's predisposition to madness

congenital inconsequentiality of his fool and the assumed madness of Edgar as "Poor Tom."

King Lear has a distinct underplot, a separable story of the fortunes of Gloucester—another father suffering from "filial ingratitude" and from his false judgment of the characters of his children. This underplot is introduced in the opening scene, in some detail, as if it were of as much importance as the main plot. The stages by which Gloucester similarly learns by suffering are clearly indicated. He begins by being the cheerful sinner, but gradually his sense of pity and duty become stronger, and he reveals himself to Edmund: "If I die for it (as no less is threatened me), the King my master must be relieved." This revelation of his good intentions to the treacherous Edmund leads directly to his downfall and to his being blinded.

Two of the "good" characters, Edgar and Albany, also grow in moral stature and strength in the course of the play. At first, Edgar seems rather ineffectual, quite unable to cope with the villainy of his half brother Edmund; but eventually he emerges as a strong character, confirmed by suffering and by compassion; able to fight and overcome Edmund in the ordeal at arms; and eventually, as one of the survivors, he is entrusted, along with Albany, with the future of the kingdom. Albany, too, is gradually built up, from being the weak husband of Goneril to being the spokesman of virtue and justice, with an authority able to cope with the force of Edmund's malignant energy.

Yet the representatives of goodness and of hope in *King Lear* do not emerge dynamically, and it has been difficult for champions of Shakespeare's moral and religious orthodoxy to combat the pessimism and nihilism that most readers experience when reading the play—precisely the qualities that have made it a favourite in the 20th century.

The historical basis of *Macbeth*

Macbeth. Macbeth is the only play of Shakespeare's that seems, to a large extent, to be related to the contemporary historical situation. It was intended to interest the new monarch, James VI of Scotland, who became James I of England. This was a matter of professional importance; for Shakespeare's company of actors had been taken over by the King on his accession, entitled to wear the royal livery as his retainers (as they did when they walked in the procession at his coronation). But more important than the flattering of the King was the way *Macbeth* satisfied public interest. For its subject was regicide, commonly regarded as the supreme crime. And the public had been profoundly moved by an attempted regicide in November 1605—the famous "Gunpowder Plot"—which the English people, even after three and a half centuries, have still not forgotten. The reign of Macbeth, king of Scotland, belonged, for Shakespeare and his audiences, to Scottish history of many centuries past. But the play of *Macbeth,* both in its treatment of the events of the story and in its details, was devised by Shakespeare with a very clear consciousness of the mood of his own times.

It is the first task of criticism to interpret the success of *Macbeth,* to explain how Shakespeare transformed a crude and horrible story of murderous ambition into a satisfying imaginative vision of good and evil. There are two principal artistic methods by which he effected this transformation. First, he made his play highly poetical; it is audacious in style, relying upon concentrated, brilliant brevity of phrase. So great is the imaginative verbal vigour that some critics, sensitive to poetry but unsympathetic to the theatre, have almost forgotten that *Macbeth* is a play and have encouraged readers to treat it rather as a poem. Second, Shakespeare has consistently humanized the two murderers, so that they almost become sympathetic—and, by making them husband and wife, their human relationship is as interesting as their motives for evil actions. This humanizing process is the key to Shakespeare's success in the play. His control of the reactions of an audience is an achievement of theatrical art, not of intellectual or moral subtlety.

Timon of Athens. Timon of Athens is yet another of Shakespeare's experiments—the exploration of a new kind of tragic form. Certain usual elements of Shakespearean tragedy are reduced in importance or eliminated from the structure of the play: the story, or "plot," is simple and lacks development. There is no maturing of characters—the only change is the single one of Timon, who moves from a fixed character of universal generosity to one of universal hatred. In the first half of the play, there is a consistent effort to build up a world around Timon in terms of which his behaviour can be judged. As he perceives characters and situations unrealistically and responds to them disproportionately, it becomes clear that his is a dream world. Into that world—as the audience watches, with some pain—reality intrudes. The second half of the play, however, is simply a series of interviews between Timon and his visitors, seemingly arranged solely to bring them under his curses; Timon's frenzied vituperation of his fellowmen becomes almost unbroken monotone. Eventually Timon rages himself to death, leaving Alcibiades to lament his death and punish his enemies.

Of the various explanations put forward for the uneven quality of the writing in *Timon of Athens* (collaboration; incomplete revision; completion by an inferior dramatist), much the most probable is that this is Shakespeare's rough draft of a play. It certainly has close analogies with the great plays of the few years to which its composition belongs (iterative words and iterative imagery, ironic preparation and anticipation, "chorus" statements by disinterested observers, plot and subplot parallel, both complementary and contrasted). If it is a rough draft, then it presents a unique opportunity of getting close to Shakespeare's method of writing. It would prove that he put structure before composition; that he went straight ahead drafting the structure of a play, unifying it by means of theme, imagery, and ironic preparation, and paying less attention to prose-verse form and to the characterization of minor personages. It would indicate that his pen wrote speeches quickly, not wasting much time at first about verse form, putting down the gist of what the character had to say, sometimes with imagery that came to him on the spur of the moment, incorporating lines or half lines of blank verse and even occasionally rhymed couplets—all to be "worked up" later. Nor, by judging the play "unfinished," are its worth and importance diminished: certain parts may be roughly written, but the imaginative conception has a wholeness that imperfect composition does not obscure.

Timon of Athens and Shakespeare's working method

The "dark" comedies. Before the death of Queen Elizabeth I in 1603 the country was ill at ease: the House of Commons became more outspoken about monopolies and royal prerogative; uncertainty about the succession to the throne made the future unsettled. In 1603 the Plague again struck London, closing the theatres. In 1604 Shakespeare's patron, the Earl of Southampton, was arrested on charges of treason; he was subsequently released, but such scares did not betoken confidence in the new reign. About Shakespeare's private reaction to these events there can be only speculation, but three of the five plays usually assigned to these years have become known as "dark" comedies for their distempered vision of the world.

Troilus and Cressida. Troilus and Cressida may never have been performed in Shakespeare's lifetime, and it fits no single category. Based on Homer (as translated by George Chapman) and on 15th-century accounts of the Trojan War by John Lydgate and William Caxton, it explores the causes of strife between and within the Greek and Trojan armies and might well have been a history play of a newly questioning and ironic kind. But Shakespeare was also influenced by Geoffrey Chaucer's love poem, *Troilus and Criseyde,* and for the first time portrayed sexual encounters outside the expectation of marriage. Cressida desires Troilus (and later Diomedes) with physical as well as idealistic longing; she considers love frankly as a chase and acts on the principle

That she was never yet that ever knew
Love got so sweet as when desire did sue.

As in Shakespeare's earlier comedies, these lovers are contrasted with others—but with a difference, for they are set beside the shallow and jaded routine of Helen's life with Paris and the assembled Trojans, and beside the jealous humours of Achilles, lounging in his tent with Patroclus. Into Cressida's last scene, when she begs Diomedes to come to her tent, Shakespeare has further introduced three lookers-on who comment directly: the enslaved Troilus watching with the politically wily Ulysses and, at another corner of the stage, Thersites. The audience cannot identify simply with any one of the five characters but, instead, must take note of each varying discord.

Thersites' revulsion from the pursuit of honour and love

In his ubiquitous commentary, Thersites, perhaps the most notable single invention in the play, expresses revulsion against the pursuit of both honour and love. When the lovers have left the stage he has the last word:

> Would I could meet that rogue Diomed! I would croak
> like a raven; I would bode, I would bode. . . . Lechery,
> lechery! Still wars and lechery! Nothing else holds
> fashion. A burning devil take them!

He proclaims Agamemnon "both ass and ox"; Ajax he would scratch from head to foot, to make the "loathsomest scab in Greece"; and as Menelaus and Paris fight in the last battle, he sees them as "bull" and "dog," cuckold and cuckold maker—and then he saves his own life by cowardice. Yet he is not the only commentator in the play. Priam sits on his throne, between his disputing sons; Ulysses arouses Achilles by praising Ajax; and Cassandra cries out from a supernatural certainty of doom. Pandarus describes the handsome warriors as food for love and hurries forward to take prurient pleasure in Cressida's excitement.

Pandarus also speaks the epilogue, and something of the play's irony may be gauged by contrasting his weak and broken appearance as he speaks the last words with that of Prologue, who, dressed in armour, had announced the scene of Troy, the "princes orgulous," and all the brave and massy consequences of war. No history play by Shakespeare had run such a gamut from the heroic to the petty and familiar. Although the earlier comedies sometimes start with tyrannies and loss, they all conclude with a dance or, at their least hopeful, with a procession offstage until a "golden time convents." Nor could this play be called a tragedy, for, despite the death of Hector and the loss of all Troilus' hopes beyond those of hatred and revenge, there is no scope in the last disordered and inglorious battle for the intensity of tragedy. The stature of every character has been progressively diminished.

All's Well That Ends Well; Measure for Measure. The other two comedies of 1601–04, are less completely original. Both *All's Well That Ends Well* and *Measure for Measure* centre on stories of love that lead to marriage, and both end with complicated denouements superintended by a benevolent king or duke. The fantasy of the earlier comedies, however, is largely missing, and, although there are clowns, they haunt brothels or a prison, or they "love as an old man loves money, with no stomach." Subplots are about cowardice in battle or fornication. Only *Measure for Measure* has a song, and that "pleases" the woe of Mariana deserted by her unworthy Angelo.

The intensity of the "dark" comedies

Perhaps the most important element in both comedies is a new intensity. In *All's Well,* Helena's soliloquies of frustrated love and, sometimes, the brevity of her speech reveal her inner struggles; and a few words from her graceless husband indicate his new loss of assurance. In *Measure for Measure,* Isabella pleading for her brother's life and defending her chastity or the Duke disguised as a friar persuading Claudio to disdain life both carry their arguments fiercely and finely; and Angelo is shown twice in soliloquy, tempted to what he knows is evil by Isabella, who he knows is good:

> When I would pray and think, I think and pray
> To several subjects. . . .
>
> (Act II, scene 4, 1 ff.)

In this comedy the intensity of much of the dialogue, the overt religious and legal concerns, and the variety of plot and subplot in Shakespeare's earlier manner all combine to make a searching, unsettling, and, in the opinion of most judges, precarious play, a comedy that reaches its general conclusion only with difficulty, with adroitness, compromise, or dramatic necessity.

Although such issues are present in *All's Well,* they are more in the background, and Shakespeare ostensibly considers other themes—inherited virtue opposed to virtue achieved by oneself, the wisdom of age over and against the impressionability of youth, and the need for each man to make his own choice, even if wrongly.

Only during the 20th century have the three "dark" comedies been frequently performed in anything like Shakespeare's texts, an indication that their questioning, satiric, intense, and shifting comedy could not please earlier audiences.

The late plays. *Pericles, Cymbeline, The Winter's Tale, The Tempest,* and *Henry VIII,* written between 1608 and 1612, are commonly known as Shakespeare's "late plays," or his "last plays," and sometimes, with reference to their tragicomic form, they are called his "romances." Works written by an author in his 40s hardly deserve to be classified as "late" in any critical sense, yet these plays are often discussed as if they had been written by a venerable old author, tottering on the edge of a well-earned grave. On the contrary, Shakespeare must have believed that plenty of writing years lay before him, and indeed the theatrical effectiveness and experimental nature of *Cymbeline, The Winter's Tale,* and *The Tempest* in particular make them very unlike the fatigued work of a writer about to break his staff and drown his book.

One of the common characteristics of these plays is that, although they portray a wide range of tragic or pathetic emotions, events move toward a resolution of difficulties in which reconciliations and reunions are prominent. They differ from earlier comedies in their structural emphasis on a renewal of hope that comes from penitence and forgiveness, together with a faith in the younger generation, who by love will heal or obliterate the wounds inflicted in the past.

There is also an extravagance of story and an unreality of motivation, both prompting the use of the label "romances." From *Coriolanus,* the most austere of his tragedies, Shakespeare turned immediately to *Pericles,* a fantastically episodic play set in a vaguely pre-Christian world. He no longer saw antiquity through the eyes of a historian like Plutarch but through the bright fictions of the imitators of late Greek romances, Heliodorus and Achilles Tatius. In *The Winter's Tale,* for instance, where events are determined by the god Apollo from his oracle on the island of Delphos, the kingdoms of Bohemia and Sicilia nevertheless contain Warwickshire country festivals and conycatchers; and Queen Hermione boldy asserts: "The Emperor of Russia was my father."

Some critics have attributed this change in theatrical manner to Shakespeare's boredom with everything except poetry; others have pointed to a revival of interest in romantic tragicomedy (in *Phylaster,* by Beaumont and Fletcher, many characteristics of Shakespeare's latest plays are discernible, but the precise date of this play is questioned, and thus it is difficult to decide whether the fashion was set by Beaumont and Fletcher with their play or by Shakespeare with *Pericles*). It is at least likely, however, that Shakespeare himself was the pioneer and originator of a new style. The King's Men took over the Blackfriars Theatre in 1608. It was a more expensive theatre, and it has reasonably been conjectured that its facilities influenced Shakespeare to produce a new kind of play for a sophisticated audience more responsive to imaginative experiment in drama.

The stylistic maturity of the second half of *Pericles*

Pericles. The first scenes of *Pericles* are often feeble in expression, frequently unsyntactical, and sometimes scarcely intelligible. The second half is splendidly written, in Shakespeare's mature style. It is now generally supposed that the inadequate parts of the play are due to its being a reconstruction of the text from the actors' imperfect memories. For the second half of the play, either the printer had a manuscript of good quality or the actors' memories were more accurate. Ben Jonson called it "a mouldy tale." Certainly it was a very old one. There was a Greek prose fiction on the story, which survives in a Latin translation. Versions of it are found in many European languages. John Gower, Chaucer's friend and contemporary, had included the story in his poem *Confessio amantis,* and there were two separate prose renderings in Shakespeare's time.

Cymbeline. The main theme of *Cymbeline*—Posthumus' wager on the chastity of his wife, Imogen—Shakespeare derived from a story in Boccaccio's *Decameron.* But he put this Italianate intrigue into a setting that, for his audience, was authentic. Cymbeline, king of Britain, and his two sons, Guiderius and Arviragus, who succeeded him, were, for Shakespeare's audience, historical monarchs. The play is carefully set in the pre-Christian Roman world. The Romans who invade Britain are recognizably derived from the same kind of exploration of the antique world that had produced *Julius Caesar, Antony and Cleopatra,* and *Coriolanus.*

Shakespeare shows great dramatic skill in weaving together many different elements of plot, period, and place. The open-air scenes in Wales, Iachimo's concealment in a trunk in Imogen's bedchamber, the supposed deaths of Posthumus and of Imogen disguised as Fidele, the battle between the Britons and Romans, the vision of the eagle-borne Jupiter—all these in preparation for the amazing complications of the final scene of the play, where (it has been calculated) there are 24 distinct "revelations" in 455 lines.

The Winter's Tale. In *The Winter's Tale* Shakespeare's audacity had increased. He introduced a similar combination of heterogeneous civilizations. But he also abandoned a unity of development in the story: he made a break of 16 years in the middle of the play, introducing the figure of Time to persuade the audience to accept the rapid shift of dramatic time. Shakespeare similarly challenged their confidence by the pathos of the deaths of the little boy Mamillius and, seemingly, of his mother Hermione, followed by the repentance of her jealous husband Leontes. He also allows Antigonus, a most attractive character, to be eaten by a bear. Shakespeare clearly intended the audience to become involved only at certain moments of the play, by intensifying particular episodes without allowing emotional commitment to the whole plot. Three times his characters use the phrase "like an old tale," as if they were themselves commenting on its incredibility. What might have been the most moving scene or series of scenes in this group of plays—the revelation of Perdita's true birth—is related only at second hand by a number of anonymous gentlemen. Presumably, Shakespeare did not wish to anticipate or reduce the theatrical effect of the final scene, in which the "statue" of Hermione comes to life. Or perhaps, having already shown a similar scene of recognition of father and long-lost daughter in *Pericles,* he did not wish to repeat himself.

Ingredients in the plot of *The Tempest*

The Tempest. *The Tempest* shows even greater excellence in the variety of its ingredients. There is story enough, including two assassination plots. There is a group of quickly differentiated characters; there is elaborate dancing and singing; there is an inset entertainment, a marriage masque performed by the goddesses Iris, Ceres, and Juno; there is a theatrical "quaint device," the introduction and vanishing of a banquet; there is a tender love affair; there are marvellous comic turns. Yet a mood of seriousness is felt throughout the play; questions about freedom, about the instinct for revenge, and the conflicting claims of generosity are being asked; there is a sense of subtle seriousness when Prospero speaks his cryptic epilogue as an actor appealing for the good opinion of his audience.

Most of the action of the plot has taken place in the past; only the climax of reconciliation and the events immediately leading up to it are represented. But, although the "unity of time" is almost preserved, the amount of mental progress, the number of mental events, is large. The newcomers to Prospero's island grope their way toward repentance; the Prince finds his way to true love, after his previous tentative explorations; and Miranda awakens to womanhood. Caliban, the subhuman creature, rebels and then comes to learn the error of his ways; Ariel, the supernatural sprite, finds the means of regaining his freedom. Prospero resumes his political authority as duke and pardons all.

The play has a most interesting double focus, geographically speaking. Openly it is a story of Naples and Milan, a world of usurpations, tributes, homages, and political marriages that is familiar in Jacobean tragedy. At the same time the contemporary excitement of the New World permeates the play—a world of Indians and the plantations of the colonies, of the wonders and terrors and credulities of a newly discovered land. A lesser dramatist would surely have set his play far away in the west of the Atlantic to take advantage of this contemporary excitement. Perhaps with a surer theatrical instinct, Shakespeare offered his audience a familiar Italianate fictional world, which then became shot through with glimpses of the New World, too exciting to be fictional.

Henry VIII. *Henry VIII* is a play that has offered many difficulties to criticism. It has had a long and interesting stage history, but, from the mid-19th century, judgment has been confused by doubts about Shakespeare's sole authorship of the play, for many scenes and splendid speeches are written in a style very close to that of John Fletcher (see below *Collaborative and attributed plays*). The best of recent criticism, however, is inclined to restrict itself to consideration of the play as it stands. Although a story of English history, it differs from the "histories" that Shakespeare had written earlier, during the reign of Queen Elizabeth I. It is more episodic—more of a pageant and a series of loosely connected crises—than a skillfully plotted drama. It has a different sort of unity: three tragic episodes involving the deaths of Buckingham, Wolsey, and Queen Katharine led to the prophecy of a new age. For Anne Boleyn's infant, whose christening closes the play, inspires Thomas Cranmer, archbishop of Canterbury, to a marvelous speech about the glories of her future reign as the first Elizabeth. Thus *Henry VIII,* in spite of many differences, resembles Shakespeare's other late plays in its emphasis on the way in which past tragic events lead to reconciliation and to hope for the new generation.

The search for "other hands" in Shakespeare

Collaborative and attributed plays. The busy competition of Elizabethan theatre encouraged collaboration between authors and was almost the rule in some companies. Naturally, therefore, scholars have sought "other hands" in Shakespeare's plays, but, following the magisterial arguments of the distinguished scholar Sir Edmund Chambers, critical opinion has held that each play in the Folio edition of 1623 is substantially Shakespeare's. The possibility remains that parts of the earliest plays, especially *1 Henry VI* and *The Taming of the Shrew,* may derive from earlier plays by other hands. Stronger than this, however, are arguments that parts of *Henry VIII* were written by John Fletcher, who from 1608 onward wrote frequently for the King's Men. Fletcher's name has been linked with Shakespeare's elsewhere, most notably in the play *The Two Noble Kinsmen,* published as their joint work in 1634. Internal evidence of style and structure suggests that Shakespeare planned the whole work and wrote Act I; that then a need for haste arose and Fletcher took over responsiblity for Acts II, III and IV, leaving Shakespeare to complete Act V. Some such story could account for a curious lack of cohesion in the writing and also for the closeness of the theme to interests apparent in Shakespeare's latest plays.

In his own lifetime and in later ages, several plays were attributed to Shakespeare (often with no justification at all). In his own handwriting, however, are 147 lines of a scene in *The Booke of Sir Thomas More,* a play written in about 1595 by a group of five authors (probably including Thomas Dekker, Anthony Munday, and Henry Chettle) and then suppressed by the censor. The attempts of Sir Thomas to quell the anti-alien riots are linked in sentiment to Shylock episodes from *The Merchant of Venice* and echo the plays of the 1590s in imagery and versification. Among the printed texts that have been ascribed to Shakespeare is the anonymous *Edward III* (1596); the stylistic evidence adduced for his authorship of an episode in which the King woos the Countess of Salisbury might, however, be the work of an imitator. Several plays were added to his works in the third edition of the Folio (1664): *Locrine* (1595), *Sir John Oldcastle* (1600), *Thomas Lord Cromwell* (1602), *The London Prodigal* (1605), *The Puritan* (1607), and *A Yorkshire Tragedy* (1608). None is now considered to be of his authorship. Other plays, too, were printed as Shakespeare's, indicating the extent of his prestige—a brand name that could sell even rotten fish. No other dramatist of his time was so misused.

SHAKESPEARE'S READING

With a few exceptions, Shakespeare did not invent the plots of his plays. Sometimes he used old stories (*Ham-*

Shakespeare's plots

let, Pericles). Sometimes he worked from the stories of comparatively recent Italian writers, such as Boccaccio—using both well-known stories (*Romeo and Juliet, Much Ado About Nothing*) and little-known ones (*Othello*). He used the popular prose fictions of his contemporaries in *As You Like It* and *The Winter's Tale.* In writing his historical plays, he drew largely from Plutarch's *Lives of the Noble Grecians and Romans* for the Roman plays and the chronicles of Edward Hall and Ralph Holinshed for the plays based upon English history. Some plays deal with rather remote and legendary history (*King Lear, Cymbeline, Macbeth*)—though it seemed more genuinely historical to Shakespeare's contemporaries than it does today. Earlier dramatists had occasionally used the same material (there were, for example, earlier plays called *The Famous Victories of Henry the fifth* and *King Leir*). But, because many plays of Shakespeare's time have been lost, it is impossible to be sure of the relation between an earlier, lost play and Shakespeare's surviving one: in the case of *Hamlet* it has been plausibly argued that an "old play," known to have existed, was merely an early version of Shakespeare's own.

Availability of literary contacts and accessibility of books

Shakespeare was probably too busy for prolonged study. He had to read what books he could, when he needed them. His enormous vocabulary could only be derived from a mind of great celerity, responding to the literary as well as the spoken language. It is not known what libraries were available to him. The Huguenot family of Mountjoys, with whom he lodged in London, presumably possessed French books. There was, moreover, a very interesting connection between Shakespeare and the book trade. For there survives the record of apprenticeship of one Richard Field, who published Shakespeare's two poems *Venus and Adonis* and *Lucrece,* describing him as the "son of Henry Field of Stratford-upon-Avon in the County of Warwick, tanner." When Henry Field the tanner died in 1592, John Shakespeare the glover was one of the three appointed to value his goods and chattels. Field's son, bound apprentice in 1579, was probably almost exactly the same age as Shakespeare. From 1587 he steadily established himself as a printer of serious literature—notably of Sir Thomas North's translation of Plutarch (1595, reprinted in 1603 and 1610). There is no direct evidence of any close friendship between Field and Shakespeare. But it cannot escape notice that one of the important printer-publishers in London at the time was an exact contemporary of Shakespeare at Stratford, that he can hardly have been other than a schoolfellow, that he was the son of a close associate of John Shakespeare, and that he published Shakespeare's first poems. Clearly, a considerable number of literary contacts were available to Shakespeare, and many books were accessible.

That Shakespeare's plays had "sources" was already apparent in his own time. An interesting contemporary description of a performance is to be found in the diary of a young lawyer of the Middle Temple, John Manningham, who kept a record of his experiences in 1602 and 1603. On February 2, 1602, he wrote:

> At our feast we had a play called *Twelfth Night, or What You Will,* much like *The Comedy of Errors,* or *Menaechmi* in Plautus, but most like and near to that in Italian called *Inganni.* . . .

The first collection of information about sources of Elizabethan plays was published in the 17th century—Gerard Langbaine's *Account of the English Dramatick Poets* (1691) briefly indicated where Shakespeare found materials for some plays. But, during the course of the 17th century, it came to be felt that Shakespeare was an outstandingly "natural" writer, whose intellectual background was of comparatively little significance: "he was naturally learn'd; he needed not the spectacles of books to read nature," said John Dryden in 1668. It was nevertheless obvious that the intellectual quality of Shakespeare's writings was high and revealed a remarkably perceptive mind. The Roman plays, in particular, gave evidence of careful reconstruction of the ancient world.

The first collection of source materials, arranged so that they could be read and closely compared with Shakespeare's plays, was made by Mrs. Charlotte Lennox in the 18th century. More complete collections appeared later, notably those of John Payne Collier (*Shakespeare's Library,* 1843; revised by W. Carew Hazlitt, 1875). These earlier collections have been superseded by one edited by Geoffrey Bullough as *Narrative and Dramatic Sources of Shakespeare* (7 vol., 1957–72).

It has become steadily more possible to see what was original in Shakespeare's dramatic art. He achieved compression and economy by the exclusion of undramatic material. He developed characters from brief suggestions in his source (Mercutio, Touchstone, Falstaff, Pandarus), and he developed entirely new characters (the Dromio brothers, Beatrice and Benedick, Sir Toby Belch, Malvolio, Paulina, Roderigo, Lear's fool). He rearranged the plot with a view to more effective contrasts of character, climaxes, and conclusions (*Macbeth, Othello, The Winter's Tale, As You Like It*). A wider philosophical outlook was introduced (*Hamlet, Coriolanus, All's Well That Ends Well, Troilus and Cressida*). And everywhere an intensification of the dialogue and an altogether higher level of imaginative writing together transformed the older work.

Acquaintance with the literary achievements of other writers

But, quite apart from evidence of the sources of his plays, it is not difficult to get a fair impression of Shakespeare as a reader, feeding his own imagination by a moderate acquaintance with the literary achievements of other men and of other ages. He quotes his contemporary Christopher Marlowe in *As You Like It.* He casually refers to the *Aethiopica* ("Ethiopian History") of Heliodorus (which had been translated by Thomas Underdown in 1569) in *Twelfth Night.* He read the translation of Ovid's *Metamorphoses* by Arthur Golding, which went through seven editions between 1567 and 1612. Chapman's vigorous translation of Homer's *Iliad* impressed him, though he used some of the material rather sardonically in *Troilus and Cressida.* He derived the ironical account of an ideal republic in *The Tempest* from one of Montaigne's essays. He read (in part, at least) Samuel Harsnett's *Declaration of egregious popish impostors* and remembered lively passages from it when he was writing *King Lear.* The beginning lines of one sonnet (106) indicate that he had read Edmund Spenser's poem *The Faerie Queene* or comparable romantic literature.

He was acutely aware of the varieties of poetic style that characterized the work of other authors. A brilliant little poem he composed for Prince Hamlet (Act II, scene 2, line 115) shows how ironically he perceived the qualities of poetry in the last years of the 16th century, when poets such as John Donne were writing love poems uniting astronomical and cosmogenic imagery with skepticism and moral paradoxes. The eight-syllable lines in an archaic mode written for the 14th-century poet John Gower in *Pericles* show his reading of that poet's *Confessio amantis.* The influence of the great figure of Sir Philip Sidney, whose *Arcadia* was first printed in 1590 and was widely read for generations, is frequently felt in Shakespeare's writings. Finally, the importance of the Bible for Shakespeare's style and range of allusion is not to be underestimated. His works show a pervasive familiarity with the passages appointed to be read in church on each Sunday throughout the year, and a large number of allusions to passages in Ecclesiasticus (Wisdom of Jesus the Son of Sirach) indicates a personal interest in one of the uncanonical books.

Understanding Shakespeare

SYMPATHETIC EXPLORATION OF THE TEXTS

On opening the works of Shakespeare, a reader can be held by a few lines of verse or a sentence or one complex, glittering, or telling word. Indeed, Shakespeare's supreme mastery of words and images, of sound, rhythm, metre, and texture, as well as the point, neatness, and lyricism of his lines, has enslaved countless people.

The next step in understanding, for most readers, is an appreciation of individual characters. Many of the early books on Shakespeare were about his "characters," and controversy about them still continues. Appreciation of the argument of the plays usually comes on insensibly, for Shakespeare is not a didactic playwright. But most persistent readers gain an increasing sense of a unity of inspiration, of an alert moral judgment and idealistic vision, both in the individual plays and in the works as a whole.

When the plays are seen in performance, they are further revealed in a new, three-dimensional, flesh and blood reality, which can grow in the minds of individual playgoers and readers as they become more experienced in response to the plays' many suggestions.

Shakespeare's direct appeal to readers

But, while various skills and learned guidance are needed for a developed understanding of Shakespeare, the directness of his appeal remains—the editors of the First Folio commended the plays to everyone "how odd soever your brains be, or your wisdoms." Perhaps most essentially, the plays will continually yield their secrets only to imaginative exploration.

CAUSES OF DIFFICULTY

Questions of authorship. The idea that Shakespeare's plays and poems were not actually written by William Shakespeare of Stratford has been the subject of many books and is widely regarded as at least an interesting possibility. The source of all doubts about the authorship of the plays lies in the disparity between the greatness of Shakespeare's literary achievement and his comparatively humble origin, the supposed inadequacy of his education, and the obscurity of his life. In Shakespeare's writings, people have claimed to discover a familiarity with languages and literature, with such subjects as law, history, politics, and geography, and with the manners and speech of courts, which they regard as inconceivable in a common player, the son of a provincial tradesman. This range of knowledge, it is said, is to be expected at that period only in a man of extensive education, one who was familiar with such royal and noble personages as figure largely in Shakespeare's plays. And the dearth of contemporary records has been regarded as incompatible with Shakespeare's eminence and as therefore suggestive of mystery. That none of his manuscripts has survived has been taken as evidence that they were destroyed to conceal the identity of their author.

The claims put forward for Bacon. The first suggestion that the author of Shakespeare's plays might be Francis Bacon, Viscount St. Albans, seems to have been made in the middle of the 19th century, inquiry at first centring on textual comparison between Bacon's known writings and the plays. A discovery was made that references to the Bible, the law, and the classics were given similar treatment in both canons. In the later 19th century a search was made for ciphered messages embedded in the dramatic texts. In *Love's Labour's Lost*, for example, it was found that the Latin word "honorificabilitudinitatibus" is an anagram of *Hi ludi F. Baconis nati tuiti orbi* ("These plays, the offspring of F. Bacon, are preserved for the world."). Professional cryptographers of the 20th century, however, examining all the Baconian ciphers, have rejected them as invalid, and interest in the Shakespeare–Bacon controversy has diminished.

Edward de Vere, earl of Oxford

Other candidates. A theory that the author of the plays was Edward de Vere, 17th earl of Oxford, receives some circumstantial support from the coincidence that Oxford's known poems apparently ceased just before Shakespeare's work began to appear. It is argued that Oxford assumed a pseudonym in order to protect his family from the social stigma then attached to the stage and also because extravagance had brought him into disrepute at court. Another candidate is William Stanley, 6th earl of Derby, who was keenly interested in the theatre and was patron of his own company of actors. Several poems, written in the 1580s and exhibiting signs of an immature Shakespearean style, cannot well have been written by Shakespeare himself. One of these is in Derby's handwriting, and three of them are signed "W.S." These initials are thought by some to have been a concealment for Derby's identity (for some such motives as were attributed to Oxford) and to have been later expanded into "William Shakespeare."

Christopher Marlowe

Shakespeare has also been identified with Christopher Marlowe, one theory even going so far as to assert that Marlowe was not killed in a tavern brawl in 1593 (the corpse of another being represented as his own) but was smuggled to France and thence to Italy where he continued to write in exile—his plays being fathered on Shakespeare, who was paid to keep silent.

The case for Shakespeare. In spite of recorded allusions to Shakespeare as the author of many plays in the canon, made by about 50 men during his lifetime, it is arguable that his greatness was not as clearly recognized in his own day as one might expect. But on the other hand, the difficulties are not so great as many disbelievers have held, and their proposals have all too often raised larger problems than they have resolved. Shakespeare's contemporaries, after all, wrote of him unequivocally as the author of the plays. Ben Jonson, who knew him well, contributed verses to the First Folio of 1623, where (as elsewhere) he criticizes and praises Shakespeare as the author. John Heminge and Henry Condell, fellow actors and theatre owners with Shakespeare, signed the dedication and a foreword to the First Folio and described their methods as editors. In his own day, therefore, he was accepted as the author of the plays. Throughout his lifetime, and for long after, no person is known to have questioned his authorship. In an age that loved gossip and mystery as much as any, it seems hardly conceivable that Jonson and Shakespeare's theatrical associates shared the secret of a gigantic literary hoax without a single leak or that they could have been imposed upon without suspicion. Unsupported assertions that the author of the plays was a man of great learning and that Shakespeare of Stratford was an illiterate rustic no longer carry weight, and only when a believer in Bacon or Oxford or Marlowe produces sound evidence will scholars pay close attention to it and to him.

Linguistic and historical problems. Since the days of Shakespeare, the English language has changed, and so have audiences, theatres, actors, and customary patterns of thought and feeling. Time has placed an ever-increasing cloud before the mirror he held up to life, and it is here that scholarship can help.

Problems are most obvious in single words. In the 20th century, "presently," for instance, does not mean "immediately," as it usually did for Shakespeare, or "will" mean "lust" or "rage" mean "folly" or "silly" denote "innocence" and "purity." In Shakespeare's day, words sounded different, too, so that "ably" could rhyme with "eye" or "tomb" with "dumb." Syntax was often different, and, far more difficult to define, so was response to metre and phrase. What sounds formal and stiff to a modern hearer might have sounded fresh and gay to an Elizabethan.

Ideas have changed, too, most obviously political ones. Shakespeare's contemporaries almost unanimously believed in authoritarian monarchy and recognized divine intervention in history. Most of them would have agreed that a man should be burned for ultimate religious heresies. It is the office of linguistic and historical scholarship to aid the understanding of the multitude of factors that have significantly affected the impressions made by Shakespeare's plays.

Textual and editorial problems. None of Shakespeare's plays has survived in his handwritten manuscript, and, in the printed texts of some plays, notably *King Lear* and *Richard III*, there are passages manifestly corrupt, with no clue to the words Shakespeare once wrote. Even if the printer received a good manuscript, small errors could still be introduced. Compositors were less than perfect; they often "regularized" the readings of their copy, altered punctuation because they lacked the necessary pieces of type, or made mistakes because they had to work too hurriedly. Even the correction of proof sheets in the printing house could further corrupt the text, since such correction was usually effected without reference to the author or to the manuscript copy; when both corrected and uncorrected states are still available, it is often the uncorrected version that is preferable. Correctors are undoubtedly responsible for many errors now impossible to right.

OVERCOMING SOME DIFFICULTIES

The contribution of textual criticism. The early editors of Shakespeare saw their task chiefly as one of correction and regularization of the faulty printing and imperfect texts of the original editions or their reprints. Many changes in the text of the quartos and folios that are now accepted derive from Nicholas Rowe (1709) and Alexander Pope (1723–25), but these editors also introduced many thousands of small changes that have since been rejected. Later in the 18th century, editors compiled collations of alternative and rejected readings. Samuel Johnson (1765), Edward Capell (1767–68), and Edmond Malone (1790) were notable pioneers. Their work reached its most comprehensive form in the Cambridge edition in nine volumes by W.G. Clark, J. Glover, and W.A. Wright, published in 1863–66. A famous one-volume Globe edition of 1864 was based on this Cambridge text.

The "New Variorum Edition"

Each major editor had added to the great number of annotations on textual problems and on linguistic and historical difficulties, and in 1871 was published the first of a series of large volumes, one or two for each play, called "A New Variorum Edition," which aimed at bringing all previous textual scholarship together. The series remains incomplete, but A.W. Pollard published his *Shakespeare Folios and Quartos* in 1909 and, together with R.B. McKerrow, Sir Walter Greg, and Charles Sisson, began a concerted study of the manuscript plays surviving from Elizabethan theatres and the practice of Elizabethan printers. Their work was summed up in Greg's study *The Shakespeare First Folio* (1955) and Fredson Bowers' *On Editing Shakespeare and the Elizabethan Dramatists* (1955). By this time a new phase of bibliographic and textual inquiry had begun, and Bowers, with Alice Walker, Charlton Hinman, and others, began studying the minutiae of each substantive early edition. Computers will analyze the huge mass of their detailed information and make available the full results of their investigations. Also, individual compositors who set in type the early editions are being identified, and already something of their work habits and habitual errors is known. Processes of correction in the Elizabethan printing houses are also being studied more intensively. Before the end of the 20th century, a new edition of Shakespeare may appear in which a multitude of small errors have been removed. But even then, nothing will compensate for the loss of Shakespeare's manuscripts, and numerous important readings and details of presentation will still have to be supplied by educated guesses.

Historical, linguistic, and dramatic studies. Since the end of the 19th century, many other problems that hinder the understanding of Shakespeare's texts have been at least partly overcome thanks to extensive investigation into, for example, his syntax, vocabulary, and word usage—especially with regard to other Elizabethan and Jacobean literature. Such technical works, as well as studies of current theories of composition (the use of rhetoric, metaphor, and simile), have provided material for critical analysis, as have works that examine the ideas and life-style of Shakespeare's contemporaries. An increasingly large proportion of research is devoted to an investigation of contemporary literary and dramatic conventions.

LITERARY CRITICISM

Literary critics and the theatre. Shakespeare criticism must take into account the certainty that Shakespeare intended his plays to be acted, that he was the professional playwright of a repertory company, that the success of a play in performance by his company was what determined his income, and that during his lifetime he apparently made no effort (or perhaps was too busy) to gain a "literary" reputation from his plays. Yet, his contemporaries had no doubt about his literary eminence. Heminge and Condell, his fellow players, and Ben Jonson, his fellow playwright, commended the great Folio of 1623 to "the great variety of readers."

The tension between the critics and the theatre

The situation has been complicated by the fact that the history of Shakespeare criticism and the stage history of his plays have run parallel but separate courses. It is fair to say that, from about the mid-18th century onward, there has been a constant tension between the critics and the theatres regarding treatment of Shakespeare. Although Dr. Johnson was the contemporary of David Garrick, Coleridge and Hazlitt the contemporaries of Edmund Kean, Dowden of Sir Henry Irving, and Bradley of Beerbohm Tree, the links between these critics and actors were not notably strong. Theatregoers were usually impressed by character impersonations rendered by virtuoso actors, while readers became more and more impressed by an awareness and admiration of the special artistic form of the plays. Rather than by stage performances, Shakespeare criticism was influenced by the dominant literary forms of each age: by the self-revelatory poem of the Romantic period, the psychological and ethical novel of the Victorians, the fragmentary revelations of the human condition in 20th-century poetry. It is a platitude that each age finds what it wants to find in Shakespeare. It will only see what it can. It can only see what it must. But Shakespeare critics, if criticism is to be in a healthy condition, must pay more than lip service to that fact. However deeply embedded in its contemporary situation, good criticism, like all intellectual feats, is a leaping out of the situation. The history of Shakespeare criticism is a subject of more than scholarly interest. It is a cautionary tale. Sometimes it is an awful warning.

The progress of Shakespeare criticism. As a basis for the criticism of an author's work, it is reasonable to begin by inquiring how his contemporaries assessed his achievement. But contemporary literary criticism was surprisingly silent about the plays actually being written (though critics were reasonably articulate about the plays they thought ought to be written). Collections of references made to Shakespeare later, during the 17th century, show that many important writers paid little attention to him. Ben Jonson's reputation was, for a variety of reasons, probably superior for the first half of the century. He was, moreover, the most vocal literary critic of the early 17th century, and he thought of Shakespeare as a naturally gifted writer, who failed to discipline himself. From his criticism derived the distinction between "nature" and "art" that for long proved to be a pertinacious and unproductive theme of Shakespeare criticism. It was further encouraged by John Milton, when he contrasted Jonson's "learned sock" with Shakespeare's "native wood-notes wild" (which refers to the comedies but came to be treated as a general statement, especially the epithet "wild"). A good deal of the spirit of Ben Jonson's cavillings, rather than his magnificent praise in the poem prefixed to the Folio of 1623, was continued by later 17th-century and 18th-century critics, censuring Shakespeare's carelessness, his artistic "faults."

John Dryden (died 1700) was the first great critic of Shakespeare. Much concerned with his own art as a dramatist, he judged Shakespeare in a practical spirit. For 100 years after him, the best literary criticism of Shakespeare was an elaboration and clarification of his opinions. Dryden on some occasions praised Shakespeare in the highest terms and boldly defended the English tradition of the theatre, maintaining that, if it was contrary to the revered classical precepts of Aristotle, it was only because Aristotle had not seen English plays; had he seen them, his precepts would have been different. Dryden nevertheless at times attributes artistic "faults" to Shakespeare, judging him according to Neoclassical principles of taste that derived from France and soon prevailed throughout Europe. Shakespeare's dramatic art was so different from that of the admired tragedy of the times (that of Corneille and Racine) that it was difficult for a critic to defend or interpret it in reasonable terms. The gravest charge was the absence of "poetical justice" in Shakespeare's plays. Although some of the better 18th-century critics, such as Joseph Addison and Dr. Johnson, saw the limitations of "poetical justice" as an artistic theory, they nevertheless generally felt that the ending of *King Lear,* especially, was intolerable, offending all sense of natural justice by the death of Cordelia. The play was thus given a "happy ending"—one congruent with "poetical justice"—by the poet and playwright Nahum Tate: Lear was restored to his authority, and Cordelia and Edgar were to be married and could look forward to a prosperous reign over a united Britain. This change was approved by Dr. Johnson, and the revised play held the stage for generations and was the only form of *King Lear* performed on the stage until the mid-19th century.

In the early 18th century the cumbrous folio editions were replaced by more convenient editions, prepared for the

Pope's theory that Shakespeare's natural genius was hampered by writing for the theatre

reader. Nicholas Rowe, in a six-volume edition of 1709, tidied up the text of the plays, adding scene divisions, lists of dramatis personae, indications of locality, and so on. Rowe was himself a practicing and successful dramatist, and on the whole he gave a good lead to the dramatic criticism of the plays. The preface to Alexander Pope's edition of 1725, however, had an unhappy influence on criticism. Shakespeare's natural genius, he felt, was hampered by his association with a working theatre. Pope fully accepted the artistic form of Shakespeare's writings as due to their being stage plays. But he regarded this as a grave disadvantage and the source of their artistic defectiveness. Lewis Theobald took the opposing view to Pope's, claiming that it was an advantage of Shakespeare that he belonged to the theatrical profession. Dr. Johnson similarly realized that methods of producing the plays on the stage influenced the kind of illusion created. In the splendid preface to his edition (1765), Johnson dismissed, once and for all in English criticism, the Neoclassical theories of "decorum," the "unities," and the mutual exclusiveness of tragedy and comedy—theories now seen as irrelevant to Shakespeare's art and as having confused the discussion of it. In many ways Johnson was the source of the notion of Shakespeare as a realistic dramatist, whose mingling of tragic and comic scenes was justified as "exhibiting the real state of sublunary nature," in which laughter and tears, rejoicing and misery, are found side by side.

Johnson censured Shakespeare as a dramatic artist for his lack of morality. Shakespeare, he wrote, "sacrifices virtue to convenience, and is so much more careful to please than to instruct, that he seems to write without any moral purpose." Later critics have felt compelled to controvert or to circumvent such a judgment. But Johnson's preface, if it did not give the discussion of Shakespeare's artistry a fresh start, cleared away some of the dead or irrelevant doctrines.

During the rest of the 18th century, after Johnson, it was scholarship rather than criticism that advanced in England. The best of literary criticism of the time was in the discovery of new subtleties in Shakespeare, especially in characterization; indeed, the acceptance of Shakespeare as a dramatic artist largely came about through his evident powers of characterization (to which probably the growth of the realistic novel in the 18th century had made readers more sensitive), and inadequate attention was paid to other aspects of his dramatic craftsmanship. There was nothing in England comparable to the brilliant Shakespeare criticism of Lessing, Goethe, and August von Schlegel in Germany. The latter's essay on *Romeo and Juliet* of 1797 demonstrated that, apart from a few witticisms, nothing could be taken away from the play, nothing added, nothing rearranged without mutilating the work of art and confusing the author's intentions. Here modern Shakespeare criticism begins. Indeed, from the time of Lessing, in the 18th century, to the mid-19th century, German critics and scholars made substantial and original contributions to the interpretation of Shakespeare, indicating Shakespeare's superlative artistry, at a time when in England he was admired more as a great poet and a brilliant observer of mankind than as a disciplined artist.

Coleridge's critical opinions

Samuel Taylor Coleridge, the greatest English critic of Shakespeare, vigorously denied that he had been affected by Schlegel (though he acknowledged the influence of Lessing), but there are many similarities between the two. His criticism—which has to be put together from reports of his lectures, from his notebooks, and from memories of his conversation, for he never succeeded in writing an organized book on Shakespeare—at first censured Shakespeare's lack of artistry, but in his lectures, given from 1810 onward, and in his *Biographia Literaria* (1817), he demonstrated that Shakespeare's "irregularities" were in fact the manifestations of a subtle intelligence. It was the purpose of criticism to reveal the reasons why the plays are as they are. Shakespeare was a penetrating psychologist and a profound philosopher; but Coleridge claimed that he was an even greater artist, and his artistry was seen to be "unconscious" or "organic," not contrived. Thus the dominant literary forms of Coleridge's age, which were those of self-revelatory poetry, influenced the criticism of the age: Hamlet was felt to speak with the voice and feeling of Shakespeare, and, as for the sonnets, William Wordsworth, whose greatest achievement was writing a long poem on the growth of his own mind, explained that Shakespeare "unlocked his heart" in his sonnets.

These critical opinions were inclined to degenerate in inferior hands in the course of the 19th century: the belief in Shakespeare's all-pervading artistry led to oversubtle interpretations; the enthusiasm for character analysis led to excessive biography writing outside the strictly dramatic framework; and the acknowledged assessment of Shakespeare's keen intelligence led to his being associated with almost every school of thought in religion, politics, morals, psychology, and metaphysics. Nevertheless, it was the great achievement of the Romantics to have freed criticism from preoccupation with the "beauties" and "faults" in Shakespeare and to have devoted themselves instead to interpreting the delight that people had always felt in the plays, whether as readers or theatregoers. Shakespeare's "faults" now became "problems," and it was regarded an achievement in literary criticism to have found an explanation for some hitherto difficult or irreconcilable detail in a play. Many of the most brilliant writers of Europe were critics of Shakespeare; and their utterances (whether or not they may be regarded as having correctly interpreted Shakespeare) are notable as recording the impressions he made upon great minds.

SHAKESPEARE'S INFLUENCE

Today Shakespeare's plays are performed throughout the world, and all kinds of new, experimental work finds inspiration in them: " . . . in the second half of the twentieth century in England," wrote the innovative theatre director Peter Brook, "we are faced with the infuriating fact that Shakespeare is still our model."

Shakespeare's influence on English theatre was evident from the start. John Webster, Philip Massinger, and John Ford are among the better known dramatists who borrowed openly from his plays. His influence is evident on Restoration dramatists, especially Thomas Otway, John Dryden, and William Congreve. John Osborne, Harold Pinter, Samuel Beckett, and George Bernard Shaw are among 20th-century writers in whose works Shakespearean echoes are to be found. Many writers have taken over Shakespeare's plots and characters: Shaw rewrote the last act of *Cymbeline,* Tom Stoppard invented characters to set against parts of *Hamlet* in *Rosencrantz and Guildenstern Are Dead* (1968), and Edward Bond used *King Lear* as the starting point for his own *Lear* (1971).

Influence on European theatre

Shakespeare has also influenced dramatists and theatre directors outside his own country. In Germany, English acting troupes were welcomed early in the 17th century, and the German version of *Hamlet, Der bestrafte Brudermord* ("Fratricide Punished"), testifes to the immediate influence of that play. His influence on later European dramatists ranges from a running allusion to Hamlet in Anton Chekhov's play *The Seagull* to imitation and parody of *Richard III* in Bertolt Brecht's *Arturo Ui,* adaptation of *King John* by Max Frisch, and André Gide's translation and simplification of *Hamlet.*

Shakespeare's influence on actors since his own day has been almost as widespread. Many European and American actors have had their greatest successes in Shakespearean roles. In England very few actors or actresses reach preeminence without acting in his plays. Each player has the opportunity to make a part his own. This is not because Shakespeare has created only outlines for others to fill but because he left so many and varied invitations for the actor to call upon his deepest, most personal resources.

Theatre directors and designers after Shakespeare's time, with every technical stage resource at their command, have returned repeatedly to his plays, which give opportunity for spectacle and finesse, ritual and realism, music and controlled quietness. Their intrinsic theatricality, too, has led to adaptations into very different media: into opera (as Verdi's *Otello*) and ballet (as versions of *Romeo and Juliet* from several nations); into sound recordings, television programs, and films. Musicals have been made of the comedies (as *Kiss Me Kate* from *The Taming of the Shrew*); even a tragedy, *Othello,* was the inspiration of a "rock" musical in 1971 called *Catch My Soul,* while *Macbeth* has yielded a political-satire show called *Macbird!* (1967).

Shakespeare has Hamlet say that the aim of theatre performance is to "hold the mirror up to nature," and this is what the history of his plays, from their first production to the latest, shows that he has, preeminently, achieved.

MAJOR WORKS

PLAYS: (probable dates of performance given): *2 Henry VI, 3 Henry VI* (1589–91); *1 Henry VI* (1591–92); *Richard III, The Comedy of Errors* (1592–93); *Titus Andronicus, The Taming of the Shrew* (1593–94); *The Two Gentlemen of Verona, Love's Labour's Lost, Romeo and Juliet* (1594–95); *Richard II, A Midsummer Night's Dream* (1595–96); *King John, The Merchant of Venice* (1596–97); *1 Henry IV, 2 Henry IV* (1597–98); *Much Ado About Nothing, Henry V* (1598–99); *Julius Caesar, As You Like It* (1599–1600); *Hamlet, The Merry Wives of Windsor* (1600–01); *Twelfth Night, Troilus and Cressida* (1601–02); *All's Well That Ends Well* (1602–03); *Measure for Measure, Othello* (1604–05); *King Lear, Macbeth* (1605–06); *Antony and Cleopatra* (1606–07); *Coriolanus, Timon of Athens* (1607–08); *Pericles* (1608–09); *Cymbeline* (1609–10); *The Winter's Tale* (1610–11); *The Tempest* (1611–12); *Henry VIII* (1612–13).

POEMS: (dates of publication given): *Venus and Adonis* (1593); *The Rape of Lucrece* (1594); "The Phoenix and the Turtle" (1601); *Sonnets* with "A Lovers complaint" (1609).

BIBLIOGRAPHY

Modern editions: Among one-volume editions of Shakespeare's *Works* are those by GEORGE L. KITTREDGE (1936; rev. by IRVING RIBNER, 1971); PETER ALEXANDER (1951); HARDIN CRAIG (1951; rev. by CRAIG and DAVID BEVINGTON, 1973); and CHARLES J. SISSON (1954). Editions giving one volume for each play include ALFRED HARBAGE (gen. ed.), *The Pelican Shakespeare,* rev. ed. (1969); SYLVAN BARNET (gen. ed.), *The Signet Classic Shakespeare,* 27 vol. (1963–66); the revision of *The Arden Shakespeare,* ed. by UNA ELLIS-FERMOR *et al.* (1951–); and T.J.B. SPENCER (gen. ed.), *The New Penguin Shakespeare.* The *New Cambridge Edition* (gen. eds., SIR A.T. QUILLER-COUCH, J.D. WILSON, *et al.*) is also one volume for each play, but the early publications of the comedies and some histories are now considered eccentric. The *New Variorum Edition,* started by HORACE HOWARD FURNESS in 1871, is still in progress, and the completeness of its textual apparatus and notes ensures continual usefulness.

Textual studies: WALTER EBISCH and LEVIN L. SCHÜCKING compiled a comprehensive *Shakespeare Bibliography* (1931, reissued 1973; supplemented in 1937 for the years 1930–35). The account is updated by GORDON R. SMITH, *A Classified Shakespeare Bibliography, 1936–1958* (1963). JAMES G. MCMANAWAY, *A Selective Bibliography of Shakespeare: Editions, Textual Studies, Commentaries* (1975), covers more than 4,500 items published between 1930 and 1970, mainly in English. The *Shakespeare Quarterly* publishes an annual classified bibliography, while the *Shakespeare Survey* publishes annual accounts of "Contributions to Shakespearian Study," as well as retrospective articles on work done on particular aspects. JOHN BARTLETT, *A New and Complete Concordance . . . to Shakespeare* (1894); and C.T. ONIONS, *A Shakespeare Glossary,* 2nd ed. rev. (1919), are both still most useful reference works and are kept in print. On textual criticism as applied to Shakespeare, the following works represent modern thinking and practice: FREDSON BOWERS, *Bibliography and Textual Criticism* (1964), and *On Editing Shakespeare* (1966); WALTER W. GREG, *The Editorial Problem in Shakespeare,* 3rd ed. (1954), and *The Shakespeare First Folio* (1955); CHARLTON HINMAN, *The Printing and Proof-Reading of the First Folio of Shakespeare,* 2 vol. (1963); E.A.J. HONIGMANN, *The Stability of Shakespeare's Text* (1965); RONALD B. MCKERROW, *An Introduction to Bibliography for Literary Students* (1927, reissued 1967); ALFRED W. POLLARD, *Shakespeare Folios and Quartos: A Study in the Bibliography of Shakespeare's Plays, 1594–1685* (1909, reprinted 1970); ALICE WALKER, *Textual Problems of the First Folio* (1953, reprinted 1978); and FRANK P. WILSON, *Shakespeare and the New Bibliography,* rev. ed. by HELEN GARDNER (1970). Facsimile editions of quartos and folio, which are necessary for any close work on textual problems, have been edited by WALTER W. GREG *et al., Shakespeare Quarto Facsimilies* (1939–66); and by CHARLTON HINMAN, *The First Folio of Shakespeare* (1968).

Biographies: A lively account of the various efforts that have been made to write a biography of Shakespeare from the available materials, with the help of imaginative interpretation, is SAMUEL SCHOENBAUM, *Shakespeare's Lives* (1970). See also Schoenbaum's collections of documents, paintings, and other records: *William Shakespeare: A Documentary Life* (1975); *Shakespeare: The Globe and the World* (1979); and *William Shakespeare: Records and Images* (1981). The first detailed life of Shakespeare was written by NICHOLAS ROWE, *Some Account of the Life of Mr. William Shakespear* (1709; photographic facsimile, 1967). EDMOND MALONE from 1780 onward added much new material; his final account is in the 1821 Variorum Edition. The best of the 19th-century biographies were J.O. HALLIWELL-PHILLIPPS, *Outlines of the Life of Shakespeare* (1881; last revision, 1887); and SIDNEY LEE, *A Life of William Shakespeare* (1898; 14th ed., 1931). A standard work is E.K. CHAMBERS, *William Shakespeare: A Study of Facts and Problems,* 2 vol. (1930), of which there is a useful abridgement by CHARLES WILLIAMS, *A Short Life of Shakespeare with Sources* (1933). More recent works are EDGAR I. FRIPP, *Shakespeare, Man and Artist,* 2 vol. (1938, reissued 1964); PETER ALEXANDER, *Shakespeare's Life and Art* (1939, reprinted 1979); HAZELTON SPENCER, *The Art and Life of William Shakespeare* (1940, reprinted 1970); M.M. REESE, *Shakespeare: His World and His Work,* rev. ed. (1980); GERALD E. BENTLEY, *Shakespeare: A Biographical Handbook* (1961), supplemented by his *Profession of Dramatist in Shakespeare's Time, 1590–1642* (1971); MURIEL C. BRADBROOK, *Shakespeare: The Poet in His World* (1978), with emphasis on the plays; EMRYS JONES, *The Origins of Shakespeare* (1977), a study of mid-Tudor influences on Shakespeare; PETER QUENNELL, *Shakespeare* (1963); and A.L. ROWSE, *William Shakespeare* (1963). Studies of special aspects of the life and environment include JOHN LESLIE HOTSON, *Shakespeare Versus Shallow* (1931, reprinted 1970); THOMAS W. BALDWIN, *William Shakspere's Petty School* (1943, reissued 1973), and *William Shakspere's Small Latine and Lesse Greeke,* 2 vol. (1944, reprinted 1966); and MARK ECCLES, *Shakespeare in Warwickshire* (1961). A valuable collection of background material is contained in *Shakespeare's England: An Account of the Life and Manners of His Age,* ed. by SIDNEY LEE and C.T. ONIONS, 2 vol. (1916). Also helpful are E.M.W. TILLYARD, *The Elizabethan World Picture* (1943, reprinted 1973); and *Shakespeare in His Own Age,* ed. by ALLARDYCE NICOLL (1964, reissued 1976). Standard works on the theatre of Shakespeare's professional life are E.K. CHAMBERS, *The Elizabethan Stage,* 4 vol. (1923, reprinted 1974); and GLYNNE WICKHAM, *Early English Stages 1300–1660* (vol. 2 on 1576–1600, published in two parts, 1963–72; 2nd ed., 1980–). Other helpful studies are ALFRED HARBAGE, *Shakespeare's Audience* (1941, reissued 1969); ALOIS M. NAGLER, *Shakespeare's Stage* (1958, reissued 1981); and BERNARD BECKERMAN, *Shakespeare at the Globe, 1599–1609* (1962). Special aspects of the environment that throw light on the mentality of the dramatist are RICHMOND NOBLE, *Shakespeare's Biblical Knowledge and Use of the Book of Common Prayer* (1935, reissued 1977); R. MUSHAT FRYE, *Shakespeare and Christian Doctrine* (1963); PAUL A. JORGENSEN, *Shakespeare's Military World* (1956, reissued 1973); GEORGE W. KEETON, *Shakespeare's Legal and Political Background* (1967); and KATHARINE M. BRIGGS, *The Anatomy of Puck* (1959, reprinted 1977), and *Pale Hecate's Team* (1962, reprinted 1977), on belief in fairies and witchcraft. A survey of the theories that someone other than William Shakespeare of Stratford-upon-Avon was the author of the plays published under his name has been written by H.N. GIBSON, *The Shakespeare Claimants: A Critical Survey of the Four Principal Theories Concerning the Authorship of the Shakespearean Plays* (1962, reprinted 1971).

See also REGINALD C. CHURCHILL, *Shakespeare and His Betters* (1958); and FRANK W. WADSWORTH, *The Poacher from Stratford* (1958). The possibility of cryptic messages in the plays was conclusively investigated in WILLIAM F. and E.S. FRIEDMAN, *The Shakespearean Ciphers Examined* (1957). The following are some of the principal exponents of the theories: (*Bacon*): EDWIN DURNING-LAWRENCE, *Bacon Is Shake-speare* (1910, reprinted 1971); and A.B. CORNWALL, *Francis the First, Unacknowledged King of Great Britain and Ireland . . .* (1936); see also J.M. ROBERTSON, *The Baconian Heresy* (1913, reprinted 1970). (*Edward de Vere, 17th Earl of Oxford, 1550–1604*): J. THOMAS LOONEY, *"Shakespeare" Identified* (1920; 3rd ed. 1974); PERCY ALLEN, *The Case for Edward de Vere . . . as William Shakespeare* (1930); HILDA AMPHLETT, *Who Was Shakespeare?* (1955, reprinted 1970); and GILBERT SLATER, *Seven Shakespeares* (1931, reprinted 1978), which proposes Oxford and a group of his collaborators. (*6th Earl of Derby*): ABEL LEFRANC, *Sous le masque de "William Shakespeare,"* 2 vol. (1918–19), and *À la découverte de Shakespeare,* 2 vol. (1945–50). (*Marlowe*): CALVIN HOFFMAN, *The Man Who Was Shakespeare* (1955).

Critical studies: Opinion about Shakespeare up to 1700 is collected in *The Shakespeare Allusion-Book,* ed. by JOHN MUNRO, 2 vol. (1909), rev. by EDMUND K. CHAMBERS (1932, reprinted 1970); and in GERALD E. BENTLEY, *Shakespeare and Jonson: Their Reputations in the Seventeenth Century Compared* (1945, reissued 1965). Eighteenth-century criticism is surveyed in DAVID NICHOL SMITH, *Shakespeare in the Eighteenth Century* (1928, reprinted 1978), and in his collection of *Eighteenth Century Essays on Shakespeare* (1903; 2nd ed., 1963); his anthology *Shakespeare Criticism: A Selection* (1916), includes work up to about 1825. BRIAN VICKERS, *Shakespeare: The Critical Heritage,* 6 vol. (1974–81), is an anthology covering the years 1623 to 1801. Dr. Johnson's criticism has been conveniently excerpted in *Johnson on Shakespeare,* ed. by WALTER RALEIGH (1908), and *Samuel Johnson on Shakespeare,* ed. by W.K. WIMSATT (1960). THOMAS M. RAYSOR prepared the standard edition of *Coleridge's Shakespearean Criticism* (1930; 2nd ed., 2 vol., 1960); but a useful compilation is *Coleridge on Shakespeare,* ed. by TERENCE HAWKES (1969). These can be supplemented by *Coleridge on Shakespeare: The Text of the Lectures of 1811–12,* ed. by R.A. FOAKES (1971). General surveys of the development of criticism are given by LOUIS MARDER, *His Exits and His Entrances* (1963); ALFRED HARBAGE, *Conceptions of Shakespeare* (1966); ARTHUR M. EASTMAN, *A Short History of Shakespearean Criticism* (1968, reissued 1974); and RAYMOND POWELL, *Shakespeare and the Critics' Debate* (1980). The best brief accounts are by T.S. ELIOT and J. ISAACS, "Shakespearian Criticism," in *A Companion to Shakespeare Studies,* ed. by HARLEY GRANVILLE-BARKER and G.B. HARRISON (1934, reprinted 1960); and by M.A. SHAABER, "Shakespeare Criticism: Dryden to Bradley," and STANLEY WELLS, "Shakespeare Criticism Since Bradley," in *A New Companion to Shakespeare Studies,* ed. by KENNETH MUIR and SAMUEL SCHOENBAUM (1971). The situation outside Britain may be studied in *Shakespeare in Europe,* a collection edited by OSWALD LE WINTER (1963). PATRICK MURRAY, *The Shakespearian Scene: Some Twentieth-Century Perspectives* (1969), gives a shrewd survey of modern criticism. NORMAN RABKIN, *Shakespeare and the Problem of Meaning* (1981), discusses the inevitably differing interpretations of the plays. Several 20th-century works may be regarded as having had a substantial effect on the criticism of Shakespeare. HARLEY GRANVILLE-BARKER, *Prefaces to Shakespeare* (5 series, 1929–47; 4 vol., 1963), set the plays firmly in their theatrical environment. Enquiries into the artistic conventions of Shakespeare's time were the basis of the criticism of LEVEN L. SCHÜCKING, *Character Problems in Shakespeare's Plays* (1922, reissued 1959; originally published in German, 1919); ELMER E. STOLL, *Art and Artifice in Shakespeare* (1933, reissued 1968); OSCAR J. CAMPBELL, *Shakespeare's Satire* (1943, reprinted 1971); and WILLIAM W. LAWRENCE, *Shakespeare's Problem Comedies,* 2nd ed. (1969). G. WILSON KNIGHT, *The Wheel of Fire* (1930; 4th rev. ed., 1949, reissued 1977), *The Imperial Theme* (1931; 3rd ed., 1968, reprinted 1972), *The Shakespearian Tempest* (1932; 3rd ed., 1953, reissued 1971), and later books; and L.C. KNIGHTS, *How Many Children Had Lady Macbeth?* (1933, reprinted 1973), showed how Shakespeare achieved poetical and symbolic effects. The study of Shakespeare's imagery is well represented by the pioneer work of CAROLINE F.E. SPURGEON, *Shakespeare's Imagery and What It Tells Us* (1935, reissued 1977); WOLFGANG CLEMEN, *The Development of Shakespeare's Imagery* (1951; 2nd ed., 1977); and ROBERT B. HEILMAN, *This Great Stage: Image and Structure in King Lear* (1948), and *Magic in the Web: Action and Language in Othello* (1956, reprinted 1977). S. VISWANATHAN, *The Shakespearean Play As Poem: A Critical Tradition in Perspective* (1980), concentrates on the work of Knight, Knights, and Spurgeon. An understanding of the intellectual and social background of the plays was advanced by WILLARD FARNHAM, *The Medieval Origin of Elizabethan Tragedy* (1936, reissued 1970), and *Shakespeare's Tragic Frontier* (1950, reprinted 1973); HARDIN CRAIG, *The Enchanted Glass: The Elizabethan Mind in Literature* (1936, reissued 1975); C.L. BARBER, *Shakespeare's Festive Comedy* (1959, reissued 1972); THEODORE SPENCER, *Shakespeare and the Nature of Man,* 2nd ed. (1949, reissued 1974); L.B. CAMPBELL, *Shakespeare's "Histories": Mirrors of Elizabethan Policy* (1947, reissued 1978); and SUKANTA CHAUDHURI, *Infirm Glory: Shakespeare and the Renaissance Image of Man* (1981).

Some of the notable explorations of the artistry of the plays are MADELEINE DORAN, *Endeavors of Art* (1954); BERTRAND EVANS, *Shakespeare's Comedies* (1960); and ANNE RIGHTER, *Shakespeare and the Idea of the Play* (1962). JAN KOTT, *Shakespeare, Our Contemporary* (1964; originally published in Polish, 1962), has been very widely read and influential on theatrical productions.

(J.R.Br./T.Sp./Ed.)

Shanghai

Shanghai (Wade-Giles romanization Shang-hai, Pinyin Shanghai), whose name literally means "on the sea," is one of the world's largest seaports and a major industrial and commercial centre of the People's Republic of China. It is located on the coast of the East China Sea between the mouth of the Yangtze River to the north and the bays of Hangchow and Yü-p'an to the south. The municipality covers a total area of 2,383 square miles (6,185 square kilometres), which includes the city itself, surrounding suburbs, and an agricultural hinterland; it is also China's most populous urban area.

Shanghai was the first Chinese port to be opened to Western trade, and it long dominated the nation's commerce. Since the Communist victory in 1949, however, it has become an industrial giant whose products supply China's growing domestic demands. The city has also undergone extensive physical changes with the establishment of industrial suburbs and housing complexes, the improvement of public works, and the provision of parks and other recreational facilities. Shanghai has attempted to eradicate the economic and psychological legacies of its exploited past through physical and social transformation to support its major role in the modernization of China.

The article is divided into the following sections:

Physical and human geography

THE LANDSCAPE

Site. The province-level municipality (*shih*) of Shanghai is bordered by Kiangsu Province on the north and west and Chekiang Province on the south. It includes the city of Shanghai; the nine mainland counties (*hsien*) of Pao-shan, Chia-ting, Ch'ing-p'u, Sung-chiang, Chin-shan (Chu-ching), Shang-hai (Hsin-chuang), Feng-hsien (Nan-ch'iao), Nan-hui, and Ch'uan-sha; and approximately 30 islands in the mouth of the Yangtze and offshore to the southeast in the East China Sea. The largest island, Ch'ung-ming, has an area of 270 square miles (700 square kilometres), extends more than 40 miles (65 kilometres) upstream from the mouth of the Yangtze, and is the 10th county in the greater Shanghai Municipality.

The city's composition

The mainland portion of the city lies on an almost level deltaic plain with an average elevation of 10 to 16 feet (three to five metres) above sea level. It is crisscrossed by an intricate network of canals and waterways that connect the municipality with the T'ai Hu region to the west.

Climate. The city's maritime location fosters a mild climate characterized by minimal seasonal contrast. The average annual temperature is about 58° F (14° C); the July maximum averages about 80° F (27° C), and the average January minimum is about 37° F (3° C). About 45 inches (1,140 millimetres) of precipitation fall annually, with the heaviest rainfall in June and the lightest in December.

Layout. As China's main industrial centre, Shanghai has serious air, water, and noise pollution. Industrial relocation and construction in the suburbs since the 1950s initially helped alleviate central city air pollution, although high population density and mixed industrial-residential land use continued to cause problems. The Wu-sung (Suchou) Chiang (River) and Huang-p'u Chiang, which flow through the city, are severely polluted from industrial discharges, domestic sewage, and ships' wastes; nonetheless, the Huang-p'u is Shanghai's main water source. Environmental protection and urban cleanliness is enhanced by industrial and solid waste resource recovery operations run by a municipal corporation. More than 1,000 different materials are recycled, including plastic, chemical fibre, and residues, machine components, oil and grease, rags, human hair, and animal bones.

The municipality radiates toward the north, west, and south from the confluence of the Wu-sung, and the Huang-p'u, a tributary of the Yangtze. Surrounding the central core is a transitional zone on both banks of the Huang-p'u, which encompasses a partially rural area of about 160 square miles. The suburban industrial complexes of Wu-sung to the north and Minhsing to the south were annexed in 1980. The banks of the Wu-sung, an important inland waterway connection to the interior hinterland, are occupied by a westward arterial extension of the transitional zone. To the south, however, the transitional zone terminates abruptly a few miles south of the central Shanghai urban core, at the Huang-p'u. P'u-tung, directly east across the Huang-p'u from the central business district, was founded in 1870 as one of the earliest industrial areas; it was also notorious as the city's most extensive and appalling slum. Several of the post-1949 industrial workers' residential complexes are now located there, and it is part of the Huang-p'u district.

Downtown Shanghai. The physical perspective of downtown Shanghai is much the same as in the pre-Communist period. Because of the policy of developing integrated residential and industrial complexes in suburban areas, central city development and renewal has been given low priority. Many of the pre-World War II buildings, which housed foreign commercial concerns and diplomatic missions, still dominate the area.

Extending southward and westward from the confluence of the Wu-sung and Huang-p'u rivers, central Shanghai has a gridded street pattern and includes the area originally contained within the British concession. The area is bounded on the east along the Huang-p'u by Chung-shan Tung Lu (Chung-shan Tung Road); on the west by Hsi-tsang Chung Lu; and on the south by Yen-an Tung Lu, which was built on the former Yang-ching-p'ang Canal that separated the British from the French concessions. Chung-shan Tung Lu has several hotels, the central administrative offices of Shanghai, and a residence for foreign seamen. The main commercial artery, Nan-ching Tung Lu, runs westward from the eastern road. On this street are located Shanghai's largest retail establishment—the Shanghai Number One Department Store—as well as restaurants, hotels, and the central communications and telegraph building.

The Hung-k'ou district lies to the north and east of the Wu-sung Chiang. It was originally developed by American and Japanese concessionaires and in 1863 was combined with the British concession to the south to create the International Settlement. It is an important industrial area, with shipyards and factories spread out along the bank of the Huang-p'u in the eastern section of the district.

The Bund, a wide boulevard of buildings along the Huang-p'u Chiang in Shanghai.

Peter Carmichael—Aspect Picture Library

Its best known building, the Shang-hai Ta-hsia (Shanghai Mansions Hotel), overlooks the Huang-p'u.

The old Chinese city, which is now part of central Shanghai, is characterized by a random and labyrinthine street pattern. Until the early 20th century the area was surrounded by a three-mile wall. It is now circumscribed by the two streets of Jen-min Lu and Chung-hua Lu, which follow the course of the original wall; and it is bisected by the main north–south artery, Ho-nan Nan Lu (South Ho-nan Road).

Western Shanghai is primarily residential in character and is the site of the Industrial Exhibition Hall. To the southwest, the district of Hsü-hui, formerly Ziccawei, became a centre of Christian missionary activity in China in the 17th century. During the late 1800s, Jesuit priests established a major library, a printing establishment, an orphanage, and a meteorological observatory in the area.

Land-use patterns in metropolitan Shanghai mirror pre-1949 real-estate market conditions. Much of the high-value land given over to industrial plants, warehouses, and transport facilities lies close to the Huang-p'u and Wu-sung rivers. South of the Wu-sung, which is traversed by about 20 bridges within the city, residential areas extend south from the industrial strip to the Huang-p'u. North of the Wu-sung, residential areas are less clearly demarcated, and there is a more gradual merging of city and country in the transitional zone. Continuous urban settlement is bounded on the north by the two major east–west arteries of Chung-shan Pei Lu and Ssu-ping Lu.

Retail trade is concentrated in the old central business district, although the volume of trade conducted there has diminished with the establishment of the industrial satellite towns and villages on the periphery of Shanghai.

Housing. Shanghai has made considerable progress since 1949 in providing housing for its growing population. Construction of integrated, self-sufficient residential complexes in conjunction with industrial, agricultural, and commercial development throughout metropolitan and suburban Shanghai has helped disperse population from the overcrowded central city and has led to dramatic changes in the urban and suburban landscape. Thousands of families in urban districts, however, remain inadequately housed, and shanties persist in some areas.

The concept of state-supported housing was introduced in 1951 with the development of Ts'ao-yang Hsin Ts'un (Ts'ao-yang New Village) in an existing industrial zone on Shanghai's western periphery. Following the construction of the Ts'ao-yang Hsin Ts'un, many other residential complexes have been built. Some of them were constructed with the partial support of government bureaus or industrial enterprises to satisfy the needs of their employees. Two of the earliest complexes in this category were the Railroad Village and the Post and Telegraph Village.

Five major housing developments were built in the former slum area of Yang-shu-p'u. These include the villages of An-shan, K'ung-chiang, Ch'ang-pai, Feng-ch'eng, and the Feng-nan Erh Ts'un. Other complexes are those at P'eng-p'u, Chen-ju, I-ch'uan, Jih-hui, and Chiang-wan. Some of these are in relatively remote suburban locations in the transitional and hinterland zones near older rural marketing centres. The P'eng-p'u workers' housing project is typical. Those who work in nearby factories live in a garden-apartment complex that includes apartment buildings, administrative offices, workshops, clinics, and a nursery. The adjacent fields supply wheat, clover, beans, cabbage, melons, and rapeseed (for cooking oil) for consumption by the inhabitants of the complex.

THE PEOPLE

The greater municipality can be divided into three distinct population zones—the densely populated central city, the transitional zones, and the rural hinterland, which is one of the world's most densely settled agricultural areas.

Within metropolitan Shanghai, there are few, if any, concentrations of ethnic minority groups. The majority of the population is of Han Chinese origin.

THE ECONOMY

China's leading industrial centre

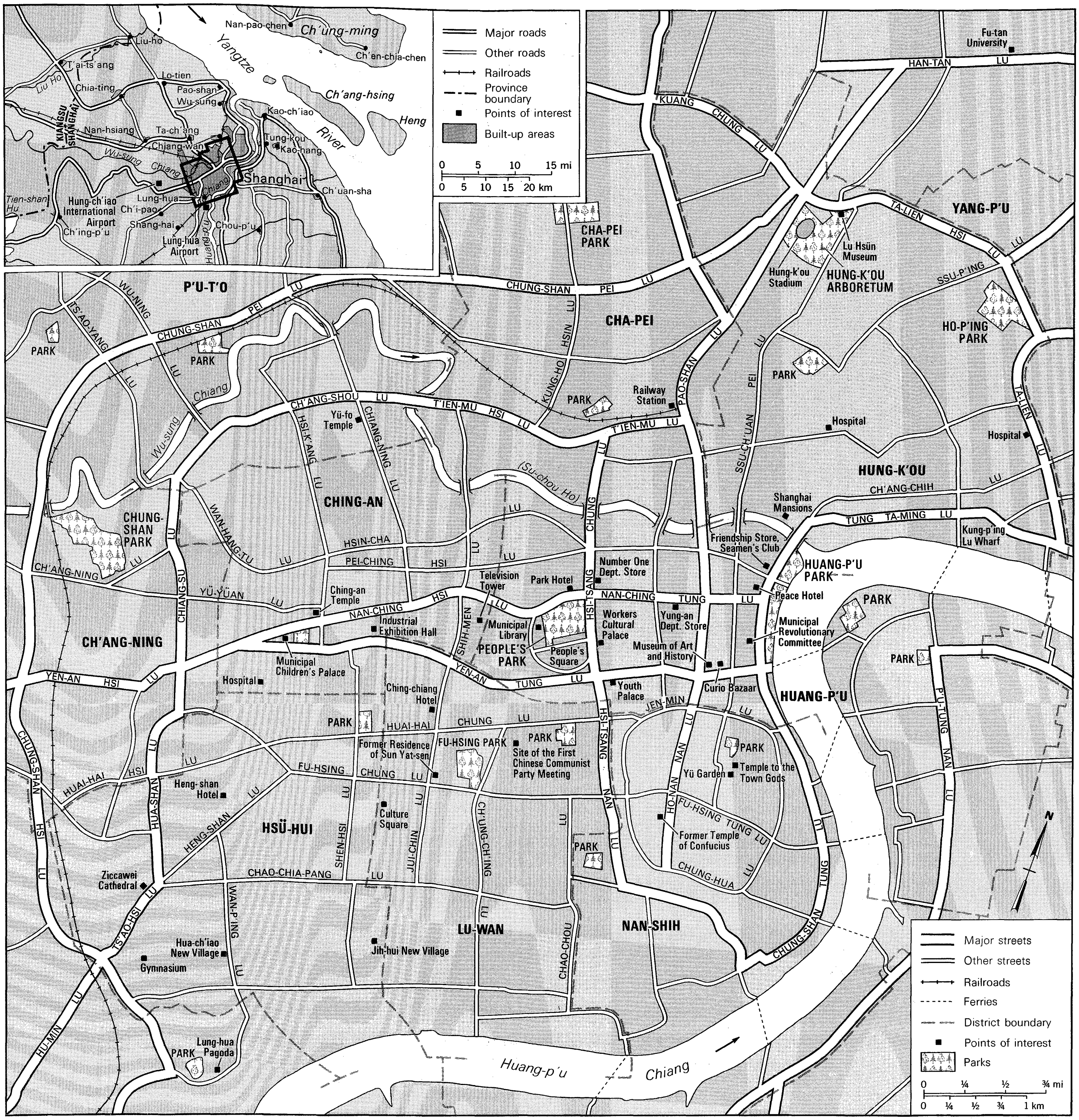
Industry. Shanghai has become the nation's leading industrial and manufacturing centre because of a distinctive combination of factors. These include the availability of a large, highly skilled, and technologically innovative work force; a well-grounded and broadly based scientific-research establishment supportive of industry; a tradition of cooperation among producers; and excellent internal and external communication and supply facilities.

The iron and steel industry was one of the earliest to be established in China. In the 1950s the blast-furnace capacity of the industry was enlarged, and attempts were made to integrate the operations of the iron and steel industry more closely with the machine-manufacturing industry.

Shanghai's machine and machine tool industry is especially important in China's modernization plans. Among the varieties of industrial equipment produced are multiple-use lathes, wire-drawing dies, and manufacturing equipment for assembling computers and other electronic devices, precision instruments, and polymer synthetics.

The chemical and petrochemical industries are almost fully integrated, and there is increasing cooperation among individual plants in the production and supply of chemical raw materials for plastics, synthetic fibres, dyes, paint, pharmaceuticals, agricultural pesticides, chemical fertilizers, synthetic detergents, and refined petroleum products. Heavy industry (especially metallurgical and chemical) predominated until the late 1970s. Light industry is now favoured in an effort to reduce pollution, alleviate transport congestion, and compensate for energy and raw material shortages associated with heavy industry.

The textile industry has been reorganized to assure efficient utilization of the mills' productive capacity at all stages of the manufacturing process. The textile mills cooperate in their use of raw materials and have estab-



Central Shanghai and (inset) its metropolitan area.

lished cooperative relationships with plants that manufacture rubber shoes, tires, zippers, industrial abrasives, and conveyor belts.

Shanghai is also a primary source of a wide variety of consumer goods such as watches, cameras, radios, fountain pens, glassware, stationery products, leather goods, and hardware. Factories producing such goods have made a special effort to meet consumer demands and to produce durable and attractive products.

Commerce. The retail trade in manufactured consumer goods is managed by the First Commercial Bureau. A number of commercial corporations under the Bureau are responsible, in turn, for the wholesaling, distribution, and warehousing of specific commodity groups. A separate corporation manages the larger retail stores, while the smaller retail establishments and some specialized wholesaling organizations are controlled by local commerce bureaus in the various districts of the city.

Banking operations

Finance and trade. Shanghai's two major banks—the People's Construction Bank and the Bank of China—function as administrative organs of the Ministry of Finance. They are responsible for the disbursement and management of capital investment funds for state enterprises. Two British banks, the Hong Kong and Shanghai Banking Corporation and the Chartered Bank, along with other foreign banks, maintain Shanghai branch offices that underwrite foreign trade transactions and exchange foreign currency in connection with trading operations. Remittances from Chinese living abroad (mainly in Hong Kong and in a number of Southeast Asian countries) are managed and collected by several overseas Chinese banks.

Industrial products are exported from Shanghai to all parts of China. Imports are mainly unprocessed food grains, petroleum and coal, construction materials, and such industrial raw materials as pig iron, salt, raw cotton, tobacco, and oils. In domestic trade, Shanghai still imports more than it exports. In foreign trade, however, the value of exported commodities exceeds that of imported goods, and the proportion of manufactured exports is steadily increasing.

The port of Shanghai

Transportation. Shanghai is China's major transport centre. The central city is both a sea and river port, with the Huang-p'u Chiang serving as an excellent harbour; at high tide, oceangoing vessels can sail up the river to the city.

In the early 1950s, the harbour was divided into a number of specialized sections. P'u-tung, on the east bank of the Huang-p'u and in the Huang-p'u district, is used for the storage of bulk commodities and for transportation maintenance and repair facilities, while P'u-hsi, in the Nan-shih District on the west bank, and Fu-hsing Tao (Fu-hsing Island) are the sites of general cargo wharves. Kao-yang Lu Wharf and I-hui contain general and bulk cargo wharves, and the Wu-sung Chiang is lined with riverine and small-craft terminals and cargo-handling facilities. Ocean terminals were constructed after 1952 at Jih-hui Chiang, south of the city, and at Chang-hua-peng, to the north at Wu-sung; a third passenger-and-freight terminal for Yangtze River and coastal traffic was opened in 1982.

Heavily used inland-waterway connections, via the Wu-sung Chiang, and an extensive canal network are maintained with Su-chou (Soochow), Wu-hsi, and Yang-chou in Kiangsu Province, and with Hang-chou (Hangchow), in Chekiang Province.

Railway facilities

The railway network reflects the efforts that have been made since 1949 to reorient the city's industrial economy to balance export and domestic development needs. Before World War II, Shanghai was the terminus of two major rail lines south of the Yangtze—the Hu-ning line, from Nanking to Shanghai, and the Hu-han-jung line, from Shanghai to the port of Ning-po in Chekiang Province. A short spur line also ran from Shanghai to Wu-sung. Additional spur lines, built since 1949, connect the industrial districts to the main trunk routes. These spurs include the Wen-tsao-pin and Min-hang lines and several short spurs emanating from Nan-hsiang (just west of the central city) to Ho-chia-wan, Peng-p'u, T'ao-p'u, and the Shanghai General Petrochemical Plant at Jinshanwei to the south.

Shanghai is served by two airports. The older Lung-hua Airport, a few miles south of the city, is used mainly for domestic flights; the Hung-ch'iao International Airport, southwest of Shanghai, is one of China's busiest. Intraurban transport by electric trolleybus, trolley, and motorbus has been substantially improved since 1949.

ADMINISTRATIVE AND SOCIAL CONDITIONS

Government. As a first-order, province-level administrative unit, Shanghai Municipality is, in theory, directly controlled by the central government in Peking. It is difficult, however, to gauge the precise nature of this relationship. Since the Cultural Revolution of the late 1960s, China's administrative apparatus at all levels of the hierarchy has been in a process of readjustment so as to bring governmental organization in line with political reality. In 1967, at the beginning of the Cultural Revolution, the Shanghai Municipal Revolutionary Committee was established as the top governing body in the municipality after a chaotic period in which a number of popular-based revolutionary organizations seized control of the city for brief periods. The committee at that time was composed of representatives of the army, the mass revolutionary organizations, and some former Communist Party officials. By the mid-1970s, this was replaced by a municipal government made up of commissions, offices, and bureaus responsible to the Shanghai People's Congress, an elected body. These units serve both policy advisory and administrative functions and function as administrative links to both the national government in Peking as well as the local governing bodies.

Public utilities. Modern public works improvements include the installation and improvement of drainage and sewage treatment facilities, public water supply systems, street lights, and public refuse bins. Roads have been widened and repaired, flood walls constructed in low-lying areas subject to tidal inundation, and housing built. The sea walls surrounding Shanghai have also been strengthened and enlarged; two long sea walls extend east of the Huang-p'u for a total of more than 13 miles (21 kilometres).

Shanghai is also one of China's major electric power generating centres. Electricity can be generated by coal-fired thermal plants, and the Shanghai area is linked via a major transmission network with Nanking to the northwest and with Hangchow and Hsin-an-chiang (the site of a hydroelectric generating facility) in Chekiang Province to the southwest. China's largest gas works is located at Lung-hua. Increased energy demands for industry and domestic use in the early 1980s led to a decision by the national authorities to construct one of China's first two nuclear power plants in Shanghai.

Health. Shanghai's health-care facilities range from thousands of small clinics associated with factories, schools, retail establishments, and government offices to numerous major research and teaching hospitals. Most hospitals have facilities for practicing and teaching both traditional Chinese and Western medicine. Medical schools have concentrated on the training of "barefoot" doctors, practitioners with sufficient medical skills to supply basic care to people in rural areas.

Education. Shanghai is China's leading centre of higher education and scientific research. There are numerous universities and other institutions of higher learning—including Fu-tan, Chiao-t'ung, T'ung-chi, and the Hua-tung Shih-fan Ta-hsueh—as well as technical and higher education institutes. Many factories have affiliated work-study colleges to equip workers for more highly skilled jobs. In 1960 the Shanghai Municipal Part-Work Part-Study Industrial University was established through the cooperation of more than 1,000 industrial establishments. A large segment of the city's total work force is enrolled in one of these schools.

Work and college programs

The Shanghai Branch of the Chinese Academy of Sciences, China's leading scientific research and development body, is located in Shanghai. During the Cultural Revolution, practical applications of scientific work in agriculture and industry were encouraged. Since the late 1970s, extensive research investments have been made in such high technology areas as nuclear energy, computers, semiconductors, laser and infrared technology, and satellites.

CULTURAL LIFE

Shanghai's cultural attractions include museums, historical sites, and scenic gardens. The Shanghai Museum of Art and History houses an extensive collection of bronzes, ceramics, and other artifacts dating over several thousand years. The Shanghai Revolutionary History Memorial Hall displays photographs and objects that trace the city's evolution. The Ta Shih-chieh ("Great World"), founded in the 1920s, is Shanghai's leading theatrical centre and offers folk operas, dance performances, plays, story readings, and specialized entertainment forms typical of China's national minority groups. The city also has many workers' and children's recreational clubs and several large motion picture theatres, including the Kuang-ming Theatre.

The old Chinese city houses the 16th-century Yü-Yuan Garden (Garden of the Mandarin Yü), an outstanding example of late Ming garden architecture, and the Former Temple of Confucius. Other points of attraction are the Ch'ing dynasty Lung-hua Pagoda, the Industrial Exhibition Hall, and the tomb and former residence of Lu Hsün, a 20th-century revolutionary writer.

The major publishing houses of Shanghai are a branch of the People's Literature Publishing House (at Peking) and the People's Educational Publishing House. In addition to the large branch of the library of the Chinese Academy of Sciences, Shanghai has numerous other libraries. Shanghai's art and music schools include a branch of the Central Conservatory (Tientsin), the Shanghai Conservatory, and the Shanghai Institute of Drama. There is also a variety of professional performing arts troupes, including ballet and opera companies, symphonies, and puppet troupes.

Parks, open spaces, and playing fields were notably expanded after 1949. Two of the earliest to be opened for public use were the People's Park in central Shanghai and the Huang-p'u Park on the shore of the Huang-p'u Chiang. Every section of the city has large parks and playing fields. Among the largest are the Hung-k'ou Arboretum and Stadium in the north; the Peace Park and Ho-p'ing Park and playing field in the northeast; the P'u-tung Park in eastern Shanghai, the Hu-nan and Fu-hsing parks in the south, and the Chung-shan Park on the western periphery of the central city.

History

EVOLUTION OF THE CITY

As late as the 5th to 7th century AD the Shanghai area, then known as Shen or Hu Tu, was sparsely populated and undeveloped. Despite the steady southward progression of Chinese settlement, the exposed deltaic position of the area retarded its economic growth.

The city's origins

During the Sung dynasty (960–1126) Shanghai emerged from its somnolent state as a small, isolated fishing village. The area to the west around T'ai Hu (T'ai Lake) had developed a self-sustaining agricultural economy on protected reclaimed land and was stimulated by an increase in population resulting from the southward migration of Chinese fleeing the invading Mongols in the north. The natural advantages of Shanghai as a deepwater port and shipping centre were recognized as coastal and inland shipping expanded rapidly. By the beginning of the 11th century, a customs office was established; and by the end of the 13th century, Shanghai was designated as a county seat and placed under the jurisdiction of Kiangsu Province.

During the Ming dynasty (1368–1644), roughly 70 percent of the cultivated acreage around Shanghai was given to the production of cotton to feed the city's cotton- and silk-spinning industry. By the middle of the 18th century there were more than 20,000 persons employed as cotton spinners.

The impact of Western imperialism

After the 1850s, the predominantly agricultural focus of the economy was quickly transformed. At this time the city became the major Chinese base for commercial imperialism by nations of the West. Following their humiliating defeat by Great Britain in 1842, the Chinese surrendered Shanghai and signed the Treaty of Nanking, which opened the city to unrestricted foreign trade. The British, French, and Americans took possession of designated areas in the city within which they were granted special rights and privileges, and the Japanese received a concession in 1895 under the terms of the Treaty of Shimonoseki.

The opening of Shanghai to foreign business immediately led to the establishment of major European banks and multipurpose commercial houses. The city's prospects as a leading centre of foreign trade were further enhanced when Canton, a rival port in the southeastern coastal province of Kwangtung, was cut off from its hinterland by the Taiping Rebellion (1850–64). Impelled by this potential threat to the uninterrupted expansion of their commercial operations in China, the British obtained rights of navigation on the Yangtze in 1857. As the natural outlet for the vast hinterland of the Lower Yangtze, Shanghai rapidly grew to become China's leading port and by 1860 accounted for about 25 percent of the total shipping tonnage entering and departing the country.

Shanghai did not, however, show promise of becoming a major industrial centre until the 1890s. Except for the Chiang-nan Arsenal organized by the Ch'ing dynasty (1644–1911) in the early 1860s, most industrial enterprises were small-scale offshoots of the larger foreign trading houses. As the flow of foreign capital steadily increased after the Sino-Japanese War of 1894–95, light industries were established within the foreign concessions, which took advantage of Shanghai's ample and cheap labour supply, local raw materials, and inexpensive power.

THE 20TH CENTURY

By contrast, local Chinese investment in Shanghai's industry was minimal until World War I diverted foreign capital from China. From 1914 through the early 1920s, Chinese investors were able to gain a tenuous foothold in the scramble to develop the industrial economy. This initial involvement was short-lived, however, as the post-World War I resurgence of Western and Japanese economic imperialism—followed closely by the Depression of the 1930s—overwhelmed many of the newly established Chinese industries. Competition became difficult, as cheaper foreign goods were dumped on the Shanghai market, and labour was attracted to relatively higher paying jobs in foreign-owned factories. Prior to the Sino-Japanese War of 1937–45 the Japanese had gained control over about half of the city's yarn-spinning and textile-weaving capacity.

Founding of the Chinese Communist Party

The 1920s was also a period of growing political awareness in Shanghai. Members of the working class, students, and intellectuals became increasingly politicized as foreign domination of the city's economic and political life became ever more oppressive. When the agreements signed by the United Kingdom, the United States, and Japan at the Washington Conference of 1922 failed to satisfy Chinese demands, boycotts of foreign goods were instituted. The Chinese Communist Party was founded in Shanghai in 1921, and four years later the Communist Party led the "May 30" uprising of students and workers. This massive political demonstration was directed against feudalism, capitalism, and official connivance in foreign imperialistic ventures. The student–worker coalition actively supported the Nationalist armies under Chiang Kai-shek, but the coalition and the Communist Party were violently suppressed by the Nationalists in 1927.

Shanghai was occupied by the Japanese during the Sino-Japanese War of 1937–45, and the city's industrial plants suffered extensive war damage. In the brief interim before the fall of Shanghai to the People's Liberation Army (PLA) in 1949, the city's economy suffered even greater dislocation through the haphazard proliferation of small, inefficient shop industries, rampant inflation, and the absence of any overall plan for industrial reconstruction.

After 1949 Shanghai's development was temporarily slowed because of the emphasis on internal regional development, especially during the period up to 1960 when close cooperation was maintained with the Soviet Union. With the cooling of relations after 1960, Shanghai has resumed its key position as China's leading scientific and technological research centre, with the nation's most highly skilled labour force.

BIBLIOGRAPHY. RHOADS MURPHEY, *Shanghai: Key to Modern China* (1953), is an authoritative study of Shanghai's pre-World War II political and economic organization. For the post-1949 period, NEALE HUNTER, *Shanghai Journal* (1969), recounts the author's experiences as an English teacher in the Shanghai Foreign Language Institute during the Cultural Revolution. Scholarly accounts of the organization and management of Shanghai's industry, trade, and financial institutions through the 1960s may be found in AUDREY G. DONNITHORNE, *China's Economic System* (1967); and BARRY M. RICHMAN, *Industrial Society in Communist China* (1969). Economic and political developments are treated in CHRISTOPHER HOWE, "The Level and Structure of Employment and the Sources of Labor Supply in Shanghai, 1949–1957," and LYNN T. WHITE III, "Shanghai's Polity in Cultural Revolution," in J.W. LEWIS (ed.), *The City in Communist China* (1971). A carefully documented collection of papers on Shanghai's political life, economic development, cultural and ideological milieu, and spatial development is brought together by CHRISTOPHER HOWE (ed.), *Shanghai: Revolution and Development in an Asian Metropolis* (1981). LYNN T. WHITE III, *Careers in Shanghai: The Social Guidance of Personal Energies in a Developing Chinese City, 1949–1966* (1978), examines political and social influences on career choices in relation to national goals. (B.Bo.)

Shintō

Shintō is the name given to indigenous religious beliefs and practices of Japan. The word Shintō literally means "the way of *kami*" (*kami* means "mystical," "superior," or "divine," generally sacred or divine power, specifically the various gods or deities); it came into use in order to distinguish indigenous Japanese beliefs from Buddhism, which had been introduced into Japan in the 6th century AD. Shintō has no founder, no official sacred scriptures, and no dogma, but it has preserved its ethos throughout the ages.

This article is divided into the following sections:

NATURE AND VARIETIES

Shintō consists of the traditional Japanese religious practices as well as the beliefs and life attitudes that are in accord with these practices. Shintō is more readily observed in the social life of the Japanese people and in their personal motivations than in a pattern of formal belief or philosophy. It remains closely connected with the Japanese value system and the Japanese people's ways of thinking and acting.

Three major types of Shintō

Shintō can be roughly classified into the following three major types: Shrine Shintō, Sect Shintō, and Folk Shintō. Shrine Shintō (Jinja Shintō), which has been in existence from the beginning of Japanese history to the present day, constitutes a main current of Shintō tradition. Shrine Shintō includes within its structure the now defunct State Shintō (Kokka Shintō)—based on the total identity of religion and state—and has close relations with the Japanese Imperial family. Sect Shintō (Kyōha Shintō) is a relatively new movement consisting of 13 major sects that originated in Japan around the 19th century. Each sect was organized into a religious body by either a founder or a systematizer. Folk Shintō (Minzoku Shintō) is an aspect of Japanese folk belief that is closely connected with the other types of Shintō. It has no formal organizational structure nor doctrinal formulation but is centred in the veneration of small roadside images and in the agricultural rites of rural families. These three types of Shintō are interrelated: Folk Shintō exists as the substructure of Shintō faith, and a Sect Shintō follower is usually also a parishioner (*ujiko*) of a particular Shintō shrine.

HISTORY TO 1900

Much remains unknown about religion in Japan during the Paleolithic and Neolithic ages. It is unlikely, however, that the religion of these ages has any direct connection with Shintō. Yayoi culture, which originated in the northern area of the island of Kyushu in about the 3rd or 2nd century BC, is directly related to later Japanese culture and hence to Shintō. Among the primary Yayoi religious phenomena were agricultural rites and shamanism.

Early clan religion and ceremonies. In ancient times small states were gradually formed at various places. By the middle of the 4th century AD a nation with an ancestor of the present Imperial Household as its head had probably been established. The constituent unit of society at that time was the *uji* (clan or family), and the head of each *uji* was in charge of worshipping the clan's *ujigami*—its particular tutelary or guardian deity. The prayer for good harvest in spring and the harvest ceremony in autumn were two major festivals honouring the *ujigami.* Divination, water purification, and lustration (ceremonial purification), which are all mentioned in the Japanese classics, became popular, and people started to build shrines for their *kami.*

Kami

Ancient Shintō was polytheistic. People found *kami* in nature, which ruled seas or mountains, as well as in outstanding men. They also believed in *kami* of ideas such as growth, creation, and judgment. Though each clan made the tutelary *kami* the core of its unity, such *kami* were not necessarily the ancestral deities of the clan. Sometimes *kami* of nature and *kami* of ideas were regarded as their tutelary *kami.*

Two different views of the world were present in ancient Shintō. One was the three-dimensional view in which the Plain of High Heaven (Takama no Hara, the *kami*'s world), Middle Land (Nakatsukuni, the present world), and the Hades (Yomi no Kuni, the world after death) were arranged in vertical order. The other view was a two-dimensional one in which this world and the Perpetual Country (Tokoyo, a utopian place far beyond the sea) existed in horizontal order. Though the three-dimensional view of the world (which is also characteristic of North Siberian and Mongolian shamanistic culture) became the representative view observed in Japanese myths, the two-dimensional view of the world (which is also present in Southeast Asian culture) was dominant among the populace.

Early Chinese influences on Shintō. Confucianism is believed to have reached Japan in the 5th century AD, and by the 7th century it had spread among the people,

By courtesy of the Japan National Tourist Organization

Procession at Tōshō-gū (Tōshō Shrine) Festival at Nikkō, Japan, May 17–18.

together with Chinese Taoism and Yin-Yang (harmony of two basic forces of nature) philosophy. All of these stimulated the development of Shintō ethical teachings. With the gradual centralization of political power, Shintō began to develop as a national cult as well. Myths of various clans were combined and reorganized into a pan-Japanese mythology with the Imperial Household as its centre. The *kami* of the Imperial Household and the tutelary *kami* of powerful clans became the *kami* of the whole nation and people, and offerings were made by the state every year. Such practices were systematized supposedly around the start of the Taika-era reforms in 645. By the beginning of the 10th century about 3,000 shrines throughout Japan were receiving state offerings. As the power of the central government declined, however, the system ceased to be effective, and after the 13th century only a limited number of important shrines continued to receive the Imperial offerings. Later, after the Meiji Restoration in 1868, the old system was revived.

The encounter with Buddhism. Buddhism was first introduced into Japan in AD 538 and developed gradually. In the 8th century there emerged tendencies to interpret Shintō from a Buddhist viewpoint. Shintō *kami* were viewed as protectors of Buddhism; hence shrines for tutelary *kami* were built within the precincts of Buddhist temples. *Kami* were made equivalent to *deva* (the Buddhist Sanskrit term for "gods") who rank highest in the Realm of Ignorance, according to Buddhist notions. Thus *kami,* like other creatures, were said to be suffering because they were unable to escape the endless cycle of transmigration; help was therefore offered to *kami* in the form of Buddhist discipline. Buddhist temples were even built within Shintō shrine precincts, and Buddhist *sūtras* (scriptures) were read in front of *kami.* By the late 8th century *kami* were thought to be avatars, or incarnations, of Buddhas and *bodhisattvas. Bodhisattva* names were given to *kami,* and Buddhist statues were placed even in the inner sanctuaries of Shintō shrines. In some cases Buddhist priests were in charge of the management of Shintō shrines.

Shintō–Buddhist syncretism

From the beginning of the Kamakura period (1192–1333), theories of Shintō–Buddhist amalgamation were formulated. The most important of the syncretic schools to emerge were Ryōbu (Dual Aspect) Shintō and Sannō Ichijitsu (One Truth of Sannō) Shintō. According to Ryōbu Shintō—also called Shingon Shintō—the two realms of the universe in Shingon Buddhist teachings corresponded to the *kami* Amaterasu Ōmikami and Toyuke (Toyouke) Ōkami enshrined at the Ise-daijingū (Grand Shrine of Ise, commonly called Ise-jingū, or Ise Shrine) in Mie Prefecture. The theorists of Sannō Ichijitsu Shintō—also called Tendai Shintō—interpreted the Tendai belief in the central, or absolute, truth of the universe (*i.e.,* the fundamental Buddha-nature) as being equivalent to the Shintō concept that the sun goddess Amaterasu was the source of the universe. These two sects brought various Buddhist rituals into Shintō. Buddhistic Shintō was popular for several centuries and was highly influential until its extinction at the Meiji Restoration.

Shintō reaction against Buddhism. Ise, or Watarai, Shintō was the first theoretical school of anti-Buddhistic Shintō in that it attempted to exclude Buddhist accretions and also tried to formulate a pure Japanese version. Watarai Shintō appeared in Ise during the 13th century as a reaction against the Shintō–Buddhist amalgamation. Konton (chaos), or Kizen (non-being), was the basic *kami* of the universe for Watarai Shintō and was regarded as the basis of all beings, including the Buddhas and *bodhisattvas.* Purification, which had been practiced since the time of ancient Shintō, was given much deeper spiritual meanings. *Shōjiki* (defined as uprightness or righteousness) and prayers were emphasized as the means by which to be united with *kami.*

Yoshida Shintō, a school in Kyōto that also emerged during the 13th century, inherited various aspects of Watarai Shintō and also showed some Taoist influence. The school's doctrines were largely the work of Yoshida Kanetomo (1435–1511). Its fundamental *kami* (the source of all things and beings in the universe) was Taigen Sonjin (the Great Exalted One). According to its teaching, if one is truly purified, his heart can be the *kami*'s abode. The ideal of inner purification was a mysterious state of mind in which one worshipped the *kami* that lived in one's own heart. Although the Watarai and Yoshida schools were thus free of Buddhistic theories, the influence of Chinese thought was still very much present.

Neo-Confucian Shintō. In 1603 the Tokugawa shogunate was founded in Edo (Tokyo), and contact between Shintō and Confucianism was resumed. Scholars tried to interpret Shintō from the standpoint of Neo-Confucianism, emphasizing the unity of Shintō and Confucian teachings. Schools emerged based on the teachings of the Chinese philosophers Chu Hsi and Wang Yang-ming, and Neo-Confucianism became an official subject of study for warriors. Yoshikawa Koretaru (1616–94) and Yamazaki Ansai (1619–82) were two representative scholars of Confucian Shintō. They added Neo-Confucian interpretations to the traditional theories of Watarai Shintō, and each established a new school. The T'ai Chi (Supreme Ultimate) concept of Neo-Confucianism was regarded as identical with the first *kami* of the *Nihon shoki,* or *Nihon-gi* ("Chronicles of Japan"). One of the characteristics of Yoshikawa's theories was his emphasis on political philosophy. Imperial virtues (wisdom, benevolence, and courage), symbolized by the Sanshu no Jingi (Three Sacred Treasures), and national ethics, such as loyalty and filial piety, constituted the way to rule the state. Yamazaki Ansai further developed this tendency and advocated both mystic pietism and ardent emperor worship.

Yoshikawa Koretaru and Yamazaki Ansai

Fukko Shintō. Fukko (Restoration, or Revival) Shintō is one of the Kokugaku (National Learning) movements that started toward the end of the 17th century. Advocates of this school maintained that the norms of Shintō should not be sought in Buddhist or Confucian interpretations but in the beliefs and life-attitudes of their ancestors as clarified by philological study of the Japanese classics. Motoori Norinaga (1730–1801) represented this school. His emphasis was on the belief in *musubi* (the mystical power of becoming or of creation), which had been popular in ancient Shintō, and on a this-worldly view of life, which anticipated the eternal progress of the world in ever-changing mutations. These beliefs, together with the inculcation of respect for the Imperial line and the teaching of absolute faith—according to which all problems beyond human capability were turned over to *kami*—exercised great influence on modern Shintō doctrines.

The most important successor of Motoori in the field of Shintō was Hirata Atsutane (1776–1843), who showed the influence of Roman Catholic teachings in some respects—derived from the writings of Jesuits in China—by advancing the idea of a creator god and retribution for ethical and religious failings in another world. These doctrines, however, were not accepted into the main current of Shintō. Hirata developed the philological studies started by Motoori and trained many capable disciples. He also wrote prayers, worked out formulas for family cults of tutelary *kami* and ancestors, and promoted Shintō practices. His spirituality, reverence for the emperor, and desire to restore the Imperial regime based on the spirit of ancient Shintō enlisted many supporters and served as one of the factors in bringing about the Meiji Restoration in 1868.

Formation of Sect Shintō. At the fall of the Tokugawa shogunate in 1867, new religious movements emerged out of the social confusion and unrest of the people. What these new movements taught differed widely: some were based on mountain-worship groups, which were half Buddhist and half Shintō; some placed emphasis on purification and ascetic practices; and some combined Confucian and Shintō teachings. New religious movements—such as Kurozumi-kyō (in this sense *kyō* means "religion," or "religious body"), founded by Kurozumi Munetada (1780–1850); Konkō-kyō (Konkō is the religious name of the founder of this group and means, literally, "golden light") by Kawate Bunjirō (1814–83); and Tenri-kyō (*tenri* means "divine reason or wisdom") by Nakayama Miki (1798–1887)—were based mostly on individual religious experiences and aimed at healing diseases or spiritual salvation. These sectarian Shintō groups, numbering 13 during the

Meiji period (1868–1912), were stimulated and influenced by Restoration Shintō. They can be classified as follows:

1. Revival Shintō sects: Izumo-ōyashiro-kyō (or Taisha-kyō), Shintō-taikyō, Shinri-kyō
2. Confucian sects: Shintō Shūsei-ha, Shintō Taisei-ha
3. Purification sects: Shinshū-kyō, Misogi-kyō
4. Mountain worship sects: Jikkō-kyō, Fusō-kyō, Ontake-kyō (or Mitake-kyō)
5. "Faith-healing" sects: Kurozumi-kyō, Konkō-kyō, Tenri-kyō

SHINTŌ LITERATURE AND MYTHOLOGY

Kojiki and *Nihon-gi* (*Nihon shoki*)

Shintō has no founder. When the Japanese people and Japanese culture became aware of themselves, Shintō was already there. Nor has it any official scripture that can be compared to the Bible in Christianity or to the Qur'ān in Islām. The *Kojiki* ("Records of Ancient Matters") and the *Nihon-gi,* or *Nihon shoki* ("Chronicles of Japan") are regarded in a sense as sacred books of Shintō. They were written in AD 712 and 720, respectively, and are compilations of the oral traditions of ancient Shintō. But they are also books about the history, topography, and literature of ancient Japan. It is possible to construct Shintō doctrines from them by interpreting the myths and religious practices they describe.

Stories partially similar to those found in Japanese mythology can be found in the myths of Southeast Asia; and in the style of description in Japanese myths some Chinese influence is detectable. The core of the mythology, however, consists of tales about the sun goddess Amaterasu Ōmikami, the ancestress of the Imperial Household, and tales of how her direct descendants unified the Japanese people under their authority. In the beginning, according to Japanese mythology, a certain number of *kami* simply emerged, and a pair of *kami,* Izanagi and Izanami, gave birth to the Japanese islands, as well as to the *kami* who became ancestors of the various clans. Amaterasu, the ruler of Takama no Hara; the moon god Tsukiyomi no Kami; and Susanoo (Susanowo) no Mikoto, the brother of Amaterasu, were the most important among them. A descendant of Amaterasu, Jimmu, is said to have become the first emperor of Japan. Japanese mythology says that the Three Sacred Treasures (the mirror, the sword, and the jewels), which are still the most revered symbols of the Imperial Household, were first given by Amaterasu to her grandson. The Inner Shrine (Naikū) of the Ise-jingū is dedicated to this ancestral goddess and is the most venerated shrine in Shintō.

The Japanese classics also contain myths and legends concerning the "800 myriads of *kami*" (*yao-yorozu no kami;* literally, *yao* equals 800 and *yorozu* 10,000). Some of them are the tutelary deities of clans and later became the tutelary *kami* of their respective local communities. Many others, however, are not enshrined in sanctuaries and have no direct connections with the actual Shintō faith. Little attention was paid to Shintō mythology until the development of Restoration Shintō, and even after that it has remained well understood only by Shintō priests and scholars.

DOCTRINES

Kami and *makoto*

Concept of the sacred. At the core of Shintō are beliefs in the mysterious creating and harmonizing power (*musubi*) of *kami* and in the truthful way or will (*makoto*) of *kami.* The nature of *kami* cannot be fully explained in words because *kami* transcends the cognitive faculty of man. Devoted followers, however, are able to understand *kami* through faith and usually recognize various *kami* in polytheistic form.

Parishioners of a shrine believe in their tutelary *kami* as the source of human life and existence. Each *kami* has a divine personality and responds to truthful prayers. The *kami* also reveals *makoto* to people and guides them to live in accordance with it. In traditional Japanese thought, truth manifests itself in empirical existence and undergoes transformation in infinite varieties in time and space. *Makoto* is not an abstract ideology. It can be recognized every moment in every individual thing in the encounter between man and *kami.*

In Shintō all the deities are said to cooperate with one another, and life lived in accordance with a *kami's* will is believed to produce a mystical power that gains the protection, cooperation, and approval of all the particular *kami.*

Precepts of truthfulness and purification. As the basic attitude toward life, Shintō emphasizes *makoto no kokoro* ("heart of truth"), or *magokoro* ("true heart"), which is usually translated as "sincerity, pure heart, uprightness." This attitude follows from the revelation of the truthfulness of *kami* in man. It is, generally, the sincere attitude of a person in doing his best in the work he has chosen or in his relationship with others, and the ultimate source of such a life-attitude lies in man's awareness of the divine.

Although Shintō ethics do not ignore individual moral virtues such as loyalty, filial piety, love, and faithfulness, it is generally considered more important to seek for *magokoro,* which constitutes the dynamic life-attitude that brings forth these virtues. In ancient scriptures *magokoro* was interpreted as "bright and pure mind" or "bright, pure, upright, and sincere mind." Purification, physical and spiritual, is stressed even in contemporary Shintō to produce such a state of mind. It is a necessary means to make communion between *kami* and man possible and to enable individuals to accept the blessings of *kami.*

Nature of man and other beliefs. In Shintō it is commonly said that "Man is *kami's* child." First, this means that a person was given his life by *kami* and that his nature is therefore sacred. Second, it means that daily life is made possible by *kami,* and, accordingly, the personality and life of people are worthy of respect. An individual must revere the basic human rights of everyone (regardless of race, nationality, and other distinctions) as well as his own. The concept of original sin is not found in Shintō. On the contrary, man is considered to have a primarily divine nature. In actuality, however, this sacred nature is seldom revealed in man. Purification is considered symbolically to remove the dust and impurities that cover one's inner mind.

Role of the individual in the history of his family

Shintō is described as a religion of *tsunagari* ("continuity or communion"). The Japanese, while recognizing each man as an individual personality, do not take him as a solitary being separated from others. On the contrary, he is regarded as the bearer of a long, continuous history that comes down from his ancestors and continues in his descendants. He is also considered as a responsible constituent of various social groups.

Motoori Norinaga stated that the human world keeps growing and developing while continuously changing. Similarly Japanese mythology speaks of an eternity of history in the divine edict of Amaterasu. In its view of history, Shintō adheres to the cyclical approach, according to which there is a constant recurrence of historical patterns. Shintō does not have the concept of the "last day": there is no end of the world or of history. One of the divine edicts of Amaterasu says:

> This Reed-plain-1,500-autumns-fair-rice-ear Land is the region which my descendants shall be lords of. Do thou, my August Grandchild, proceed thither and govern it. Go! and may prosperity attend thy dynasty, and may it, like Heaven and Earth, endure forever.

Modern Shintōists interpret this edict as revealing the eternal development of history as well as the eternity of the dynasty. From the viewpoint of finite individuals, Shintōists also stress *naka-ima* ("middle present"), which repeatedly appears in the Imperial edicts of the 8th century. According to this point of view, the present moment is the most valuable of all conceivable times. In order to participate directly in the eternal development of the world, it is required of Shintōists to live fully each moment of life, making it as worthy as possible.

Historically, the *ujigami* of each local community played an important role in combining and harmonizing different elements and powers. The Imperial system, which has been supported by the Shintō political philosophy, is an example of unity and harmony assuming the highest cultural and social position in the nation. After the Meiji Restoration (1868), Shintō was used as a means of spiritually unifying the people during repeated wars. Since the end of World War II the age-old desire for peace has been restressed. *The General Principles of Shintō Life* proclaimed by the Association of Shintō Shrines in 1956 has the following article: "In accordance with the Emperor's will, let us be harmonious and peaceful, and pray for the nation's development as well as the world's co-prosperity."

RITUAL PRACTICES AND INSTITUTIONS

Matsuri

Shintō does not have a weekly religious service. People visit shrines at their convenience. Some may go to the shrines on the first and 15th of each month and on the occasions of rites or festivals (*matsuri*), which take place several times a year. Devotees, however, may pay respect to the shrine every morning.

Rites of passage. Various Shintō rites of passage are observed in Japan. The first visit of a newborn baby to the tutelary *kami,* which occurs 30 to 100 days after birth, is to initiate the baby as a new adherent. The Shichi-go-san (Seven-Five-Three) festival on November 15 is the occasion for boys of five years and girls of three and seven years of age to visit the shrine to give thanks for *kami*'s protection and to pray for their healthy growth. January 15 is Adults' Day. Youth in the village used to join the local young men's association on this day. At present it is the commemoration day for those Japanese who have attained their 20th year. The Japanese usually have their wedding ceremonies in Shintō style and pronounce their wedding vows to *kami.* Shintō funeral ceremonies, however, are not popular. The majority of the Japanese are Buddhist and Shintōist at the same time and have their funerals in Buddhist style. A traditional Japanese house has two family altars: one, Shintō, for their tutelary *kami* and the goddess Amaterasu Ōmikami, and another, Buddhist, for the family ancestors. Pure Shintō families, however, will have all ceremonies and services in Shintō style. There are other Shintō *matsuri* concerning occupations or daily life, such as a ceremony of purifying a building site or for setting up the framework for a new building, a firing or purifying ceremony for the boilers in a new factory, a completion ceremony for a construction works, or a launching ceremony for a new ship.

Varieties of festival, worship, and prayer. Each Shintō shrine has several major festivals each year, including the Spring Festival (Haru Matsuri, or Toshigoi Matsuri; Prayer for Good Harvest Festival), Autumn Festival (Aki Matsuri, or Niiname-sai; Harvest Festival), an Annual Festival (Rei-sai) and the Divine Procession (Shinkō-sai). The Divine Procession usually takes place on the day of the Annual Festival, and miniature shrines (*mikoshi*) carried on the shoulders are transported through the district. The order of rituals at a grand festival is usually as follows: (1) Purification rites (*harae*)—commonly held at a corner of the shrine precincts before participants come into the shrine but sometimes held within the shrine before beginning a ceremony. (2) Adoration—the chief priest and all the congregation bow to the altar. (3) Opening of the door of the inner sanctuary (by the chief priest). (4) Presentation of food offerings—rice, sake wine, rice cakes, fish, seaweed, vegetables, salt, water, etc., are offered but animal meat is not because of the taboo on shedding blood in the sacred area. In the past cooked food was usually offered to *kami,* but nowadays uncooked food is more often used. In accordance with this change, the idea of entertaining *kami* changed to that of thanksgiving. (5) Prayer—the chief priest recites prayers (*norito*) modelled on ancient Shintō prayers. These prayers were compiled in the early 10th century and were based on the old belief that spoken words had spiritual potency. (6) Sacred music and dance. (7) General offering—participants in the festival make symbolic offerings using little branches of the evergreen sacred tree to which strips of white paper are tied. (8) Taking offerings away. (9) Shutting the door of the inner sanctuary. (10) Final adoration. (11) Feast (*naorai*).

Ritual at festivals

By courtesy of the Izumo-taisha

Inner sanctuary (centre background) of the Grand Shrine of Izumo (Izumo-taisha), dedicated to Ōkuninushi no Kami.

In the olden days *naorai,* a symbolic action in which participants held communion with *kami* by having the same food offered to the deity, came in the middle of the festival ceremony. The custom is still observed sometimes at the Imperial Household and at some old shrines, but it is more common to have communion with *kami* by drinking the offered sake after the festival. Since World War II it has become popular to have a brief sermon or speech before the feast.

Most Shintō festivals are observed generally in accordance with the above-mentioned order. On such occasions as the Annual Festival, various special rites may be held; *e.g.,* special water purification (*misogi*) and confinement in shrines for devotional purposes (*o-komori*), the procession of a sacred palanquin (*o-miyuki*) or of boats (*funa matsuri*), a ceremonial feast (*tōya matsuri*), sumo wrestling, horseback riding (*kurabe-uma*), archery (*matoi*), a lion dance (*shishi mai*), and a rice planting festival (*o-taue matsuri*).

Shintai and *mitama-shiro*

Types of shrines. A simple torii (gateway) stands at the entrance of the shrine precincts (see Tōshō-gū illustration). After proceeding on the main approach, a visitor will come to an ablution basin where the hands are washed and the mouth is rinsed. Usually he will make a small offering at the oratory (*haiden*) and pray. Sometimes a visitor may ask the priest to conduct rites of passage or to offer special prayers. The most important shrine building is the main, or inner, sanctuary (*honden*), in which a sacred symbol called *shintai* ("*kami* body") or *mitama-shiro* ("divine spirit's symbol") is enshrined. The usual symbol is a mirror, but sometimes it is a wooden image, a sword, or some other object. In any case, it is carefully wrapped and placed in a container. It is forbidden to see it: only the chief priest is allowed inside the inner sanctuary.

In the beginning Shintō had no shrine buildings. At each festival people placed a tree symbol at a sacred site, or they built a temporary shrine to invite *kami.* Later they began to construct permanent shrines where *kami* were said to stay permanently. The *honden* of the Inner Shrine at Ise and of Izumo-taisha (Grand Shrine of Izumo, in Shimane Prefecture) illustrate two representative archetypes of shrine construction. The style of the former probably developed from that of a storehouse for crops, especially for rice, and the style of the latter from ancient house construction. In the course of time, variations of shrine architecture were adopted and additional buildings were attached in front of the *honden.* The *honden* and *haiden* are in many cases connected by a hall of offering (*heiden*) where prayers are usually recited. Large shrines also have a hall for liturgical dancing (*kaguraden*).

Other practices and institutions. *Ujigami* belief is the most popular form of Shintō in Japan. Originally referring to the *kami* of an ancient clan, after the 13th century *ujigami* was used in the sense of the tutelary *kami* of a local community, and all the members in the community were that *kami*'s adherents (*ujiko*). Even today a *ujiko* group consists of the majority of the residents in a given community. A Shintōist, however, can believe at the same time in shrines other than his own local shrine. It was only after World War II that some large shrines also started to organize believers' groups (*sūkeisha*). The Believers' Association of the Meiji Shrine, for instance, has a substantial membership living in and around Tokyo.

Kokugakuin University in Tokyo and Kōgakkan University at Ise are the primary training centres for Shintō priests. Though anyone who goes through certain training processes may be a priest (or a priestess), many priests are, in fact, from the families of hereditary Shintō priests.

SHINTŌ RELIGIOUS ARTS

The Japanese from ancient times have valued emotional and aesthetic intuitions in expressing and appreciating their religious experiences. They found symbols of *kami* in natural beauty and the forces of nature, and they developed explicitly religious poetry, architecture, and visual arts. Shrine precincts are covered with green trees and are places of a serene and solemn atmosphere, which is effective in calming worshippers' minds. In the larger shrines, surrounded by expansive woods with mountains as their background, a harmony of nature and architecture may be achieved. Ise-jingū and Izumo-taisha still retain the ancient architectural styles. After the 9th century an intricate form of shrine construction was developed adopting both Buddhist and Chinese architectural styles and techniques. The curving roof style is one example. Unpainted timbers are most frequently used, but wherever Buddhistic Shintō was popular, Chinese vermilion-lacquered shrines were also built.

Torii: the sacred gates

A torii always stands in front of a shrine. Various kinds of torii can be seen in Japan, but their function is always the same: to divide the sacred precincts from the secular area. A pair of sacred stone animals called *komainu* ("Korean dogs") or *karajishi* ("Chinese lions") are placed in front of a shrine. Originally they served to protect the sacred buildings from evil and defilements. After the 9th century they were used for ornamental purposes on ceremonial occasions at the Imperial Court and later came to be used at various shrines generally. Some of the stone lanterns (*ishidōrō*) used at the shrines are works of art. The dedicator's name and the year are inscribed on the lanterns to inform viewers of the long tradition of faith and to urge them to maintain it.

Compared with Buddhist statuary, visual representations of *kami* are not outstanding either in their quality or quantity. Images of *kami* were, in fact, not used in ancient Shintō until after the introduction of Buddhism into Japan. These are placed in the innermost part of the *honden* and are not the objects of direct worship by the people. *Kami* icons are not worshipped at shrines.

The history of the shrine, its construction arrangements, and ritual processions are recorded in picture scrolls (*emakimono*), and at the older shrines there are many votive pictures (*ema*)—small wooden picture plaques—that have been dedicated over the years by worshippers. Other articles, such as specimens of calligraphy, sculpture, swords, and arms, dedicated by the Imperial families, nobles, or feudal lords, are also kept at shrines. Several hundred such items and shrine constructions have been designated by the Japanese government as national treasures and important cultural properties.

Religious music and dance

The traditional religious music and dance of shrines were performed for the purpose of entertaining and appeasing *kami,* rather than to praise them. *Gagaku* (literally, "elegant music") involves both vocal and instrumental music, specifically for wind, percussion, and stringed instruments. *Gagaku* with dance is called *bugaku. Gagaku* was patronized by the Imperial Household as court music and was much appreciated by the upper classes from the 9th to the 11th century. Later some of the more solemn and graceful pieces were used as ritualistic music by shrines and temples. Today *gagaku* is widely performed at larger shrines. The authentic tradition of *gagaku* has been transmitted by the Bureau of Music (Gagaku-ryō, now called Gakubu) of the Imperial Household (established in 701).

Apart from *gagaku* there are also *kagura* (a form of indigenous religious music and dance based on blessing and purification), *ta-asobi* (a New Year's dance-pantomime of the cycle of rice cultivation), and *shishi mai,* which developed originally from magico-religious dances, and are now danced for purification and as prayers. *Matsuri-bayashi* is a gay lively music with flutes and drums to accompany divine processions. Some organizations of both Shrine and Sect Shintō have recently begun to compose solemn religious songs to praise *kami,* making use of Western musical forms. (See also EAST ASIAN ARTS.)

POLITICAL AND SOCIAL ROLES

Until the end of World War II, Shintō was closely related to the state. Offerings to *kami* were made every year by the government and the Imperial Household, and prayers were offered for the safety of the state and people. The *matsuri-goto* (the affairs of worship) offered by the emperor from olden days included not only ceremonies for *kami* but also for ordinary matters of state. "Shintō ceremonies and political affairs are one and the same" was the motto of officials. Administrators were required to have a religious conscience and develop political activities with *magokoro.*

This tradition was maintained as an undercurrent throughout Japanese history. Villagers prayed to the tutelary *kami* of the community for their peace and welfare and promoted unity among themselves with village festivals. After the Meiji Restoration, the government treated Shintō like a state religion and revived the system of national shrines, which dated from the 10th century. In order to propagate Revival Shintō as the foundation of the national structure, they initiated the "great promulgation movement" in which the emperor was respected like *kami.* Although the Japanese constitution enacted in 1889 guaranteed freedom of faith under certain conditions, priority was, in fact, given to Shintō. In elementary schools Shintō was taught to children, and most of the national holidays were related to Shintō festivals. Shintō of this nature was called State Shintō and came under the control of the Bureau of Shrines in the Ministry of Home Affairs.

Relation of State and Sect Shintō

State Shintō was regarded as a state cult and a national ethic and not as "a religion." The free interpretation of its teachings by individual Shintō priests was discouraged. Priests of the national shrines were prohibited from preaching and presiding over Shintō funerals. By 1945 there were 218 national and approximately 110,000 local shrines. The number of Sect Shintō groups was limited to 13 after the organization of Tenri-kyō. Legally these 13 sects were treated as general religious bodies, similar to Buddhism and Christianity, and came under the supervision of the Ministry of Education.

After the end of World War II, the Supreme Commander for the Allied Powers ordered the Japanese government to disestablish State Shintō. All government financial support from public funds and all official affiliation with Shintō and Shintō shrines were also discontinued. State rites performed by the emperor were henceforth to be regarded as the private religious practices of the Imperial family. These rulings were carried into the new Japanese constitution that was enacted in 1947. Presently, Shrine Shintō is faced with two serious problems. The first is determining how the traditional unifying function of Shintō can be promoted in local communities or in the nation without interfering with freedom of faith. The second is the necessity of harmonizing Shintō with rapid modernization, especially in organizing believers and dealing with human problems or the meaning of life.

The number of Shintō shrines has been decreasing since the beginning of the Meiji era, in part because a municipal unification plan in 1889 called for the shrines of tutelary *kami* to be combined with the municipality. The great majority of shrines belong to the Association of Shintō Shrines, established in 1946, and most of the others are independent or belong to small groups.

About 15 percent of 16,251 Sect Shintō churches were damaged during World War II. Although they were not affected by the occupation policies after the war, many sects, in fact, went through difficult years because of unrest among the people and disunion within their own organizations. In the late 20th century, Tenri-kyō in particular was most active among the Sect Shintō groups and was extending its missionary activities overseas to nations in East Asia and in North and South America, and also to other areas.

PLACE OF SHINTŌ IN JAPANESE AND WORLD RELIGION

Shintō together with Buddhism is closely related both culturally and socially to the life of the Japanese people. Its relationships to other religions in Japan are generally cooperative and harmonious. Most Shintōists believe that cooperation between different religions could contribute to world peace, but this is not to imply a facile religious syncretism. Shintōists insist on maintaining their own characteristics and inner depth while working toward the peaceful coexistence of human beings.

BIBLIOGRAPHY. NAOFUSA HIRAI, *Japanese Shinto* (1966), a brief general sketch; STUART D.B. PICKEN, *Shinto: Japan's Spiritual Roots* (1980), a short introduction to the origins and modern forms of Shintō; H. BYRON EARHART, *Japanese Religion: Unity and Diversity,* 3rd ed. (1982), a brief work on the formation, development, and interaction of religions; JOSEPH M. KITAGAWA, *Religion in Japanese History* (1966), a widely used textbook on Japanese religious background; W.G. ASTON (trans.), *Nihongi: Chronicles of Japan from the Earliest Times to A.D. 697,* 2 vol. (1896, reissued 1972), a standard translation into English; DONALD L. PHILIPPI (trans.), *Kojiki* (1968), a translation with introduction using contemporary Japanese philological studies; D.C. HOLTOM, *The National Faith of Japan* (1938, reissued 1965), strong in history and political philosophy; TSUNETSUGU MURAOKA, *Studies in Shinto Thought,* trans. by DELMER M. BROWN and J.T. ARAKI (1964), a dependable description of Shintō thought by an eminent philologist; ROBERT S. ELLWOOD, *The Feast of Kingship* (1973), describing the enthronement ceremonies of Japanese emperors.

(N.H.)

Sikhism

Sikhism is the religion of an Indian group founded in the Punjab (or Pañjāb) in the late 15th century AD by Gurū Nānak. Its members are known as Sikhs. The great majority of Sikhs live in the state of Punjab. Most of the remainder are in Haryāna state and Delhi or are scattered in other parts of India. Some Sikhs have also settled in Malaysia, Singapore, East Africa, the United Kingdom, the United States, and Canada. The word Sikh is derived from the Pāli *sikkha* or Sanskrit *śiṣya,* meaning "disciple." Sikhs are disciples of their Ten Gurūs (religious teachers), beginning with Nānak (1469–1539) and ending with Gobind Singh (1666–1708).

This article is divided into the following sections:

HISTORY AND BACKGROUND

Religious and cultural origins. Sikhism was a historical development of the Hindu Vaiṣṇava Bhakti movement—a devotional movement among followers of the god Vishnu—that began in Tamil country and was introduced to the north by Rāmānuja (traditionally, 1017–1137). In the 14th and 15th centuries, and after prolonged confrontation with Islām, the movement spread across the Indo-Gangetic Plain. The Bhaktas (devotees) maintained that God, though known by many names and beyond comprehension, is the one and the only reality; that all else is illusion (*māyā*); and that the best way to approach God is through repetition of his name (Sanskrit *nāma*), singing hymns of praise (Punjabi *kīrtan*), and meditation under the guidance of a Gurū. Traditional Hindu religion and society were hierarchically structured; the Bhakti movement opposed the Brahmin hegemony over religious ritual and the caste system.

Hindu Vaiṣṇava and Islāmic Ṣūfī origins

Kabīr (1440–1518), a medieval mystic poet and religious synthesist, was the link between Hindu Bhakti and Islāmic Ṣūfism (mysticism), which had gained a large following among Indian Muslims. Ṣūfīs (mystics) also believed in singing hymns and in meditation under guidance of a leader. They welcomed non-Muslims in their hospices. Sikhism drew inspiration from both Bhaktas and Ṣūfīs.

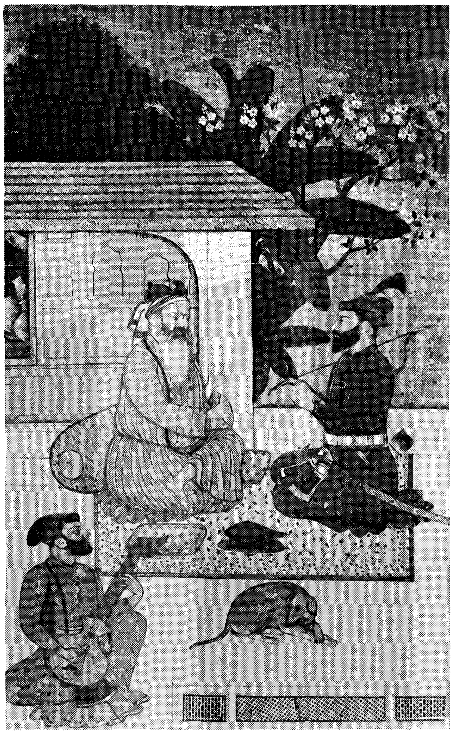
The Ten Gurūs: Nānak and his tradition. Nānak was born in 1469 in the village of Rāi Bhoi dī Talvaṇḍī, 40 miles (65 kilometres) from Lahore (in present-day Pakistan). His father was a revenue collector belonging to the Bedī (conversant with the Vedas—the revealed scriptures of Hinduism) subcaste of Kṣatriyas (Warriors). Nānak received an education in traditional Hindu lore and in the rudiments of Islām. Early in life he began associating with holy men. For a time he worked as the accountant of the Afghān chieftain at Sultānpur. There a Muslim family servant, Mardānā, who was also a rebec player, joined him. Nānak began to compose hymns. Mardānā put them to music and the two organized community hymn singing. They organized a canteen where Muslims, as well as Hindus of different castes, could eat together. At Sultānpur, Nānak had his first vision of God, in which he was ordered to preach to mankind. He disappeared while bathing in a stream. When he reappeared on the third day, he proclaimed: "There is no Hindu, there is no Mussulman."

Sikh tradition relates that Nānak also undertook four long voyages: east as far as Assam; south through the Tamil country to Ceylon; north to Ladakh and Tibet; and west as far as Mecca, Medina, and Baghdad. He spent the last years of his life in Kartārpur (in present-day Pakistan), where he raised the first Sikh temple. He nominated one of his disciples, Aṅgad, as his successor.

Aṅgad (Gurū 1539–52) was followed by another disciple, Amar Dās (Gurū 1552–74), who later nominated his son-in-law, Rām Dās Soḍhī (Gurū 1574–81) as his successor. Thereafter, the office of Gurū remained in the Soḍhī family. Rām Dās was succeeded by his youngest son, Arjun Mal (Gurū 1581–1606), who, before his death by torture in Lahore on May 30, 1606, nominated his son Hargobind (Gurū 1606–44). The seventh Gurū, Har Rāī (Gurū 1644–61), was Hargobind's grandson, who, after his tenure, nominated his young son Hari Krishen (Gurū 1661–64), who died of smallpox at the age of eight. Tegh Bahādur (Gurū 1664–75), who succeeded him, was the son of the sixth Gurū, Hargobind. Before his execution in Delhi on November 11, 1675, Tegh Bahādur passed succession to his son, Gobind Rāī (Gurū 1675–1708).

The founding of the Khālsā. The execution of two Gurūs and persecution by the Mughals compelled the Sikhs to take to arms. This was given religious sanction when, on the Hindu New Year's Day (April 13, 1699), Gobind Rāī baptized five Sikhs into a new fraternity he called the Khālsā, meaning the "Pure" (from the Persian *khāleṣ,* also meaning "pure"), and gave himself and them a common surname, Singh ("Lion"). Kaur ("Lioness") is the corresponding name given to all Sikh women. Gobind Singh's military career was not very successful. He lost most of his followers, including his four sons. He was hounded out of the Punjab and assassinated at Nānded (now in Mahārāshtra) on October 7, 1708. Before his death he declared the succession of Gurūs at an end. The military leadership of the Sikhs devolved upon Bandā Singh Bahādur. For

The surname Singh



The first Sikh Gurū, Nānak, conversing with the 10th and last Gurū, Gobind Singh. The imaginary meeting is expressive of the religion's development from a pacifist to a militant brotherhood. Painting of the Guler school, *c.* 1820. In the collection of Mohan Singh, Punjab, India.

By courtesy of the Victoria and Albert Museum, London

eight years Bandā defied the Mughals and devastated large tracts of eastern central Punjab, until he was captured and, along with 700 of his followers, executed in Delhi in the summer of 1716.

For a few years the Khālsā disappeared into the hills. But when Mughal power was weakened by the incursion in 1738–39 of the Persian Nāder Shāh, they reemerged into the plains. They organized themselves under *misl*s (from Persian *mēsāl,* meaning both "example" and "equal") and began to extract protection money from towns and villages. The series of invasions between 1747 and 1769 that were led by Aḥmad Shāh Durrānī completely disrupted Mughal administration. In the battle of Panipat in 1761, the Afghāns destroyed rising Marāthā power in the north. The power vacuum thus created allowed the Sikhs to establish themselves as the rulers of the Punjab.

The Sikh Empire of Ranjit Singh. In the years of turmoil between the Persian and Afghān invasions, Sikh *misl*s had operated in loosely defined areas. Two main divisions emerged: Cis-Sutlej, the area between the Sutlej and the Yamuna rivers, and the Trans-Sutlej, between the Sutlej and the Indus rivers. In 1761 Sikhs wrested the capital city of Lahore from the Mughal governor.

Ranjit Singh

Ranjit Singh's (1780–1839) *misl,* the Śukerchakīās, was based at Gujrānwāla, north of Lahore. Ranjit took possession of the capital in 1799 and two years later had himself crowned maharaja of the Punjab. The English, who had advanced beyond Delhi, took the Cis-Sutlej states under their protection and compelled Ranjit Singh to accept the Sutlej River as the southeastern limit of his kingdom. There, after Ranjit Singh systematically brought the Trans-Sutlej region under his suzerainty, he took Multan in 1818 and Kashmir in 1819. In the following winter he extended his domain north and west beyond the Indus River into the land of the Pathans.

Ranjit Singh then began modernizing his army by employing European officers to train his troops. This army defeated the Pathans and Afghāns and extended Sikh power to the Khyber Pass.

Relations between the Sikhs and the British. After taking the Cis-Sutlej states under their protection, the British began to make plans for extending their empire up to the Indus River. Even during Ranjit Singh's lifetime they had been interfering in the affairs of Afghanistan, and they had persuaded him to join in an Anglo-Sikh expedition to Kabul. After the death of Ranjit Singh the Sikh kingdom disintegrated rapidly. Ranjit's eldest son and successor, Kharak Singh, was deposed by his own son, Naonihal Singh, and died of excessive use of opium. On the same day, Naonihal Singh was mortally injured when a gateway collapsed on his head. Kharak Singh's widow, Chand Kaur, occupied the throne for a few months until she was deposed and later murdered by Ranjit Singh's second son, Sher Singh. On September 15, 1843, Sher Singh, his son Pratap Singh, and Chief Minister Dhian Singh Dogra were murdered by Chand Kaur's kinsmen, who in their turn were slain by Dhian Singh's son, Hira Singh Dogra. Ranjit Singh's youngest son, Dalip Singh, was proclaimed maharaja with his mother, Jindan Kaur, as regent and Hira Singh Dogra as prime minister. Power passed, however, into the hands of the pañcāyat (elected council) of the Khālsā Army, which compelled the Dogra to flee Lahore and then slew him in flight.

The British began to move their troops to the Sikh frontier and made preparations to cross the Sutlej. On December 11, 1845, the Khālsā Army began crossing the river to intercept a British force led by their commander-in-chief and the governor general. In a series of bitterly contested battles at Mudki (December 18), Fīroz Shāh (Fīrozpur; December 21–22), Alīwāl (January 28, 1846), and Sobrāon (February 10)—often called the First Sikh War—the Khālsā were defeated. The British annexed the territory between the Sutlej and Beās rivers; forced the Sikhs to reduce their army; and, on their failure to pay a large war indemnity, forced them to cede Jammu and Kashmir, which were then sold to Gulab Singh Dogra. A British resident was posted at Lahore to administer the rest of the Sikh kingdom during the minority of Dalip Singh.

Administrative measures taken by the resident aroused resentment among the people. The banishment of Jindan Kaur, the queen mother, on charges of conspiracy brought matters to a head in the winter of 1848 and touched off a general Sikh uprising, also referred to as the Second Sikh War. A bloody but inconclusive battle was fought at Chiliānwāla (January 13, 1849); however, at Gujrāt (February 21, 1849) the Khālsā were totally defeated and laid down their arms. The Sikh kingdom was annexed, and Maharaja Dalip Singh was exiled from the Punjab.

Sikh role in the Indian Mutiny

After many years of chaos, the Punjab was administered efficiently and fairly. Consequently, when the Indian Mutiny broke out in 1857, the province stayed loyal to the British, and the Sikhs took a prominent role in suppressing the Mutiny. For this loyalty and help they were rewarded by grants of land. The proportion of Sikhs in the British Army was increased. A regulation was passed requiring Sikh soldiers to observe Khālsā traditions. With the reclamation of desert lands through an extensive system of canals, unprecedented prosperity came to the Punjab. Sikhs were the most favoured settlers. Sikh loyalty was evidenced in World War I, in which Sikhs formed more than one-fifth of the British Indian Army.

The depression that followed the war led to widespread disturbances, climaxed in the killing, on April 13, 1919, at Amritsar of almost 400 people. Sikhs also clashed with the authorities over the possession of their *gurdwārā*s (temples), which were under the control of hereditary priests. The Sikh masses turned from their British connection to join Gandhi's freedom movement. The progressive introduction of democratic reforms further reduced their earlier privileged status under British rule. Their participation on the British side in World War II was considerably less enthusiastic than it had been in 1914–18.

When the subcontinent was partitioned into India and Pakistan in 1947, the Sikh population was divided equally on both sides of the boundary line. Since the partition had been preceded by savage Sikh–Muslim riots, some 2,500,000 Sikhs were compelled to leave Pakistan.

Sikhism since 1947. The government of free India abolished privileges previously extended by the British to religious minorities, including the Sikhs. Thus, the proportion of Sikhs in defense and civil services declined. The partition also adversely affected the Sikh agricultural classes, who had abandoned rich farmlands in Pakistan and changed places with Muslims of east Punjab whose holdings were much smaller. The decline in their fortunes nurtured a sense of grievance and gave birth to agitation for a Punjabi-speaking province in India in which Sikhs would form a majority of the population. This demand was conceded after the Indo-Pakistan War in 1965.

Increased wheat production during the 1970s brought unprecedented prosperity to Sikh farmers. Material improvement was accompanied by the growth of Sikh fundamentalism under the leadership of Jarnail Singh Bhindranwale. Tension increased between Sikhs and Hindus as the Akālī Religious Party (Shiromanī Akālī Dal; SAD), the predominant Sikh political party, began demanding more political and economic advantages for Sikhs.

By the early 1980s the demands of the SAD had become strongly militant, and there was an escalation in sectarian violence. The Indian government responded by arresting and imprisoning thousands of Sikhs. Armed bands, under the direction of Bhindranwale, spread a reign of terror throughout the Punjab region. Matters came to a head in 1984, when Bhindranwale and his followers entrenched themselves in the compound of the Harimandir (Golden Temple). In June the Indian Army launched an assault on the temple that killed several hundred Sikhs (including Bhindranwale) and resulted in heavy damage to the temple buildings. In October Indian Prime Minister Indira Gandhi was assassinated by two Sikh members of her bodyguard, touching off widespread Hindu violence against Sikhs. These two events caused deep resentment in the Sikh community and fueled the movement demanding the establishment of a separate Sikh state.

SIKH LITERATURE, MYTH, AND LORE

Canonical and noncanonical literature. The earliest source materials on Nānak are the *janam-sākhīs* ("life stories"), written 50 to 80 years after the death of the Gurū. Most Sikh scholars reject them and rely instead on the Gurū's compositions incorporated in the *Ādi Granth* and the *Vārs* (heroic ballads) composed by Bhāī Gurdās (died 1629). Neither Nānak's hymns nor Gurdās' *Vārs* are specific regarding the events of Nānak's life. Other historical writings date from the 18th and the 19th centuries.

The canonical work

There is only one canonical work: the *Ādi Granth* ("First Book") compiled by the fifth Gurū, Arjun, in 1604. There are at least three recensions (versions) of the *Ādi Granth* that differ from each other in minor detail. The version accepted by Sikhs as authentic is said to have been revised by Gobind Singh in 1704. The *Ādi Granth* contains nearly 6,000 hymns composed by the first five Gurūs: Nānak (974), Aṅgad (62), Amar Dās (907), Rām Dās (679), and Arjun (2,218). Gobind incorporated 115 hymns of his father, Tegh Bahādur, in it. Besides these compositions, the *Ādi Granth* contains hymns of the Bhakta saints and Muslim Ṣūfīs, notably Farīd and Kabir, and of a few of the bards attached to the courts of the Gurūs.

The *Dasam Granth* ("Tenth Book") is a compilation of writings ascribed to Gobind Singh. Scholars are not agreed on the authenticity of the contents of this *Granth,* and it is not accorded the same sanctity as the *Ādi Granth.* Traditions of the Khālsā are contained in the *Rahatnāmās* (codes of conduct) by contemporaries of Gobind Singh.

Myths and lore. Although the Gurūs themselves disclaimed miraculous powers, a vast body of *sākhīs* ("stories") recounting such miracles grew up, and with them *gurdwārās* (temples) commemorating the sites where they were performed. It also became an article of belief that the spirit of one Gurū passed to his successor "as one lamp lights another." This notion gained confirmation through the fact that the Gurūs used the same poetic pseudonym, "Nānak," in their compositions.

A composition about which little is known, but which has played an important role in Sikh affairs, is a collection of prophecies, *Sau Sākhī* ("Hundred Stories"), ascribed to Gobind Singh. Various versions are known to have been published prophesying changes of regimes and the advent of a redeemer who will spread Sikhism over the globe.

SIKH DOCTRINES, PRACTICES, AND INSTITUTIONS

Use of Hindu cosmogony

Doctrines. *Views on the nature of man and the universe.* Speculation on the origin of the cosmos is largely derived from Hindu texts. Sikhs accept the cyclic Hindu theory of *saṃsāra*—birth, death, and rebirth—and *karma,* whereby the nature of one's life is determined by his actions in a previous life. Humans are, therefore, equal to all other creatures, except insofar as they are sentient. Human birth is the one opportunity to escape *saṃsāra* and attain salvation.

Concept of the Khālsā. Khālsā is a concept of a "chosen" race of soldier-saints committed to a Spartan code of conduct (consisting of abstinence from liquor, tobacco, and narcotics and devotion to a life of prayer) and a crusade for *dharmayudha*—the battle for righteousness. The number five has always had mystic significance in the Punjab—"land of the five rivers." "Where there are five, there am I," wrote Gobind Singh. The first Khālsā were *pañj piyāres*—the five beloved ones. The ideal goal of all young Sikhs is to take *pahul* ("baptism") and thus become Khālsā. The *sahajdhārī* ("slow-adopter") is assumed to be preparing himself gradually for the initiation.

The notion of the five Ks. The five emblems of the Khālsā, all beginning with the letter *k,* have no scriptural basis but are mentioned in the *Rahatnāmās,* written by Gobind Singh's contemporaries. The most important of the Ks is *keṣa* ("hair"), which the Khālsā must retain unshorn. A Khālsā who cuts off his hair is a *patit* ("renegade"). The sanctity of unshorn hair is older than Gurū Gobind Singh—the founder of the Khālsā—for many of the earlier Gurūs also followed the tradition (common among certain sects of Hindu ascetics as well) of letting their hair and beards grow. The other four Ks are *kaṅghā* ("comb"); *kacch* ("drawers"), worn by soldiers; *kirpān* ("sabre"); and *kārā* ("bracelet") of steel, commonly worn on the right arm. The usually accepted explanation of the *kārā* is that it is the Gurū's charm against evil—a variation of the Hindu *rakhri* tied by sisters on the wrist of their brothers to keep them from harm.

Monotheism. Unity of the Godhead is emphasized in Sikhism. Nānak used the Hindu Vedāntic concept of *om,* the mystic syllable, as a symbol of God. To this he added the qualifications of singleness and creativity and thus constructed the symbol *ik* ("one") *om kār* ("creator"), which was later given figurative representation as ੴ. The opening lines of his morning prayer, *Japjī,* called the Mul Mantra ("Root Belief") of Sikhism, define God as the One, the Truth, the Creator, immortal, and omnipresent. God is also formless (*niraṇkār*) and beyond human comprehension. Sikh scriptures use many names, both Hindu and Muslim, for God. Nānak's favourite names were Sat-Kartār ("True Creator") and Sat-Nām ("True Name"). Later the word Wāh-Gurū ("Hail Gurū") was added and is now the Sikh synonym for God.

Concepts of spiritual authority. The sole repository of spiritual authority is the *Ādi Granth.* In the event of disputes, a conclave is summoned to meet at the Akāl Takht ("Throne of the Timeless"), a building erected by the sixth Gurū, Hargobind, facing the Harimandir temple in Amritsar. Resolutions passed at the Akāl Takht have spiritual sanction. Sikh religion and politics have always been intimately connected, and belief in a Sikh state is an article of faith. "Raj karey Ga Khālsā—"the Khālsā shall rule"—is chanted at the conclusion of every service.

Aniconic orientation and anti-ritualism

Views on idolatry and rituals. Sikhism forbids representation of God in pictures and the worship of idols. Nevertheless the *Ādi Granth* itself has become an object of intense ceremonial reverence and as such is known as *Granth Sahib* ("The *Granth* Personified"). *Granth Sahib* is "roused" in the morning and placed under an awning draped in fineries. Devotees do obeisance and place offerings before it. In the evening it is put to rest for the night. On festival days it is taken in procession through the streets. Most rituals centre on the *Ādi Granth.* The nonstop recitation from cover to cover by a relay of readers (*akhand-path*), which takes two days and nights, has become popular.

Social consequences of beliefs. The main consequence of Sikh belief has been a gradual breaking away from the

Sikh social structure

Hindu social system and the development of Sikh separatism. The singular worship of the *Ādi Granth* excludes worship of all other objects common among Hindus (*i.e.,* the Sun, rivers, trees, etc.) and also puts a stop to the practice of ritual purifications and pilgrimages to the Ganges. Since every Sikh is entitled to read the scripture, Sikhs do not have a priestly caste similar to the Brahmans in Hinduism. Sikh insistence on commensality (eating together) at the *Gurū ka laṅgar* ("kitchen of the Gurū") destroyed the traditional Hindu pattern of caste among them and substituted a far less rigid social structure. Sikhs are grouped into three broad categories based largely on ethnic differences: Jāṭs (agricultural tribes), non-Jāṭs (erstwhile Brahmans, Kṣatriyas, and Vaiśyas—the three highest groups of the traditional Hindu social system), and Mazahabis (untouchables). The Jāṭs, though low in the caste hierarchy, are preeminent; the Mazahabis, though converts from Hindu outcastes (untouchables outside the caste system), and still discriminated against, have a much higher status than untouchables in Hindu society. This three-tiered system is in a state of flux: among the educated urban classes it is breaking up, but in the villages a form of apartheid persists (see also SOCIAL DIFFERENTIATION: *Social stratification: Caste*).

Practices and institutions. *The Gurū and the disciple.* The guidance of the Gurū toward the attainment of *mokṣa*—release—is absolutely essential. The Gurū or the Satgurū—true Gurū—is accorded a status only a shade below that of God. His function is to point the way to the realization of the truth, to explain the nature of reality, and to give the disciple the gift of the divine word (*nām-dān*). Although the line of Gurūs ended with Gobind Singh and Sikhs regard the *Ādi Granth* as their "living" Gurū, the practice of attaching oneself to a *sant* ("saint") and elevating him to a status of a Gurū has persisted and is widely practiced.

Recitation of Nāma. Sikhism is often described as *nām-mārga* ("the way of *nāma*") because it emphasizes the constant repetition (*jap*) of the name of God and the *gurbāni* (the divine hymns of the Gurūs). *Nāma* cleanses the soul of sin and conquers the source of evil, *haumain* ("I am")—the ego. Thus tamed, the ego becomes a weapon with which one overcomes lust, anger, greed, attachment, and pride. *Nāma* stills the wandering mind and induces a super-conscious stillness (*divya dṛṣṭi*), opens the *dasam duār* ("10th gate"—the body has only nine natural orifices) through which enters divine light; and thus a person attains the state of absolute bliss.

Rites of passage and other ceremonies. No specific rites are prescribed for birth, but the practice of chanting the first five verses of Nānak's *Japjī* is observed among some Sikhs when a child is born. A few days later the child is brought to the *gurdwārā*. The *Ādi Granth* is opened and the child given a name beginning with the first letter of the first word on the left page. When a child has learned some Gurmukhi script he is initiated into reading the *Ādi Granth.* The most important ceremony is that of *pahul* ("baptism"), usually administered at puberty. The initiate takes *amrit* ("nectar") and is admitted to the Khālsā fraternity. During a Sikh marriage ceremony (*anand karaj*), the groom and bride are required to go around the *Ādi Granth* four times to the chanting of wedding hymns. On death, there is continual chanting of hymns until the body is prepared for cremation. A final *ardāsā* ("supplicatory") prayer is said before the funeral pyre is lit. Ashes of the dead are usually immersed in the Beās at Kīratpur—or in one of the Hindu sacred rivers, preferably the Ganges.

Pilgrimage places

Sacred times and places. Early hours of the dawn are ambrosial hours (*amritvelā*) most appropriate for prayer and meditation. Though not specifically prescribed as such, *gurdwārā*s with historical associations are, in fact, places of pilgrimage. Preeminent among them is the Harimandir at Amritsar, the holiest shrine of the Sikhs. Nankāna, the birthplace of Nānak (now in Pakistan), comes second. There are also five thrones (Akāl Takhts) that are accorded special sanctity; these are at Amritsar, Anandpur, Patiāla, Patna, and Nānded. The last four are the places associated with Gobind Singh. From all of them, proclamations can be made to all the Khālsā.

The first Sikh place of worship was built by Nānak at Kartārpur and was, like Hindu temples, known as *dharamsālā* ("place of faith"). At a later stage, a Sikh temple was called a *gurdwārā,* meaning "gateway to the Gurū." There are more than 200 historical *gurdwārā*s associated with the Gurūs, which are controlled by the Shiromanī Gurdwārā Prabaṅdhak Committee (SGPC) set up by the Sikh Gurdwārās Act of 1925.

In addition to historical *gurdwārā*s, every place with a sizable Sikh population is likely to have a *gurdwārā* of its own. In well-to-do homes, a room is often set apart for this purpose. The only object of worship is the *Ādi Granth.* Sikhs observe all festivals celebrated by the Hindus of northern India. In addition, they celebrate the birthdays of the first and the last Gurūs and the martyrdom of the fifth (Arjun) and the ninth (Tegh Bahādur). The biggest fair is on the first of Baisākh (mid-April), which is also the birthday of the Khālsā itself.

The Khālsā Saṅgat. The Saṅgat ("Congregation") is usually called the Sādh-saṅgat ("Congregation of Holy Men") and thus is invested with sanctity. The Saṅgat in each *gurdwārā* elects its own governing body, and decisions are taken by vote. As a rule women do not participate in the deliberations. The SGPC at Amritsar is the general governing body of Sikhism.

Minor subgroups and orders

Sectarian differences. The first dissenters from the mainstream of Sikhism—known as the Udāsīs—were followers of Nānak's elder son, Śri Chand. The order inclined toward asceticism and later furnished priests (*mahants*) for *gurdwārā*s. They were ousted from control by the SGPC in 1925. Followers of Rām Rāī, who was passed over by his father, Har Rāī (seventh Gurū), in favour of a younger son, Hari Krishen (eighth Gurū), broke away to become Rām Rāiyās. They have their headquarters in Dehra Dūn, Uttar Pradesh.

Khālsā who do not believe that the line of Gurūs ended with Gobind Singh have continued the tradition of having a living Gurū. Among these, the Bandaī Khālsā (followers of Bandā Bahādur) are now extinct, but the Nāmdhārīs and Niraṅkārīs worship living Gurūs.

Sikh welfare and educational institutions. The SGPC is the chief welfare organization of the Sikhs. The Sikh Educational Conferences, meeting annually since 1908, are the chief educational organizations credited with the establishment of a large number of schools. In 1965 two socioreligious organizations, the Gurū Gobind Singh Foundation and the Gurū Nānak Foundation, endowed many university chairs for the study of Sikhism and the publication of material on Sikh history and religion.

BIBLIOGRAPHY. JOSEPH D. CUNNINGHAM, *History of the Sikhs,* 2nd ed. (1853), was the first scholarly work on the Sikhs up to the first Anglo-Sikh war of 1845–46. His interpretation of Sikhism as an eclectic Hindu-Muslim faith remained unquestioned until McLeod's work in 1968. Cunningham was censured for suggesting that British designs on the Sikh kingdom provoked Sikh aggression. MAX A. MACAULIFFE, *The Sikh Religion,* 6 vol. (1909, reissued 1963) is a compilation of all the legends about the Sikh Gurūs based on *janam-sākhī*s and on saints whose writings are in the Sikh scriptures; full of literal translations of Sikh hymns. SHER SINGH, *Philosophy of Sikhism* (1944), is a scholarly work interpreting Sikhism as an offshoot of Vaiṣṇavite Hinduism. A selection of hymns from the *Ādi Granth* and Gobind Singh's *Dasam Granth* carried out by a panel of Sikh scholars may be found in *Selections from the Sacred Writings of the Sikhs,* trans. by TRILOCHAN SINGH *et al.* (1960). KHUSHWANT SINGH, *A History of the Sikhs,* 2 vol. (1963–66), interprets Sikhism as an aspect of Punjabi nationalism, and his *Ranjit Singh, Maharajah of the Punjab* (1963) is a biography based on Persian, Punjabi, and English sources. W.H. MCLEOD, *Gurū Nānak and the Sikh Religion* (1968, reissued 1976), casts serious doubt on source material on the life of Nānak, rejects the theory of Sikhism as an eclectic faith, and asserts that it is a branch of Hindu Vaiṣṇavism tinged with yogism. His *Evolution of the Sikh Community* (1976) is an excellent history and analysis. W. OWEN COLE and PIARA SINGH SAMBHI, *The Sikhs: Their Religious Beliefs and Practices* (1978), is an introduction, and Cole's *Sikhism and Its Indian Context: 1469–1708* (1984) explores the relations between early Sikhism and other Indian religious beliefs and practices. INDIA, *White Paper on the Punjab Agitation* (1984), deals with the storming of the Golden Temple by the Indian Army.

(K.S.)

Slavery

There is no consensus on what a slave was or on how the institution of slavery should be defined. Nevertheless, there is general agreement among historians, anthropologists, economists, sociologists, and others who study slavery that most of the following characteristics should be present in order to term a person a slave. The slave was a species of property; thus, he belonged to someone else. In some societies slaves were considered movable property, in others immovable property, like real estate. They were objects of the law, not its subjects. Thus, like an ox or an ax, the slave was not ordinarily held responsible for what he did. He was not personally liable for torts or contracts. The slave usually had few rights and always fewer than his owner, but there were not many societies in which he had absolutely none. As there are limits in most societies on the extent to which animals may be abused, so there were limits in most societies on how much a slave could be abused. The slave was removed from lines of natal descent. Legally and often socially he had no kin. No relatives could stand up for his rights or get vengeance for him. As an "outsider," "marginal individual," or "socially dead person" in the society where he was enslaved, his rights to participate in political decision making and other social activities were fewer than those enjoyed by his owner. The product of a slave's labour could be claimed by someone else, who also frequently had the right to control his physical reproduction.

For coverage of related topics in the *Macropædia* and *Micropædia,* see the *Propædia,* section 513.

This article is divided into the following sections:

Slavery was a form of dependent labour performed by a nonfamily member. The slave was deprived of personal liberty and the right to move about geographically as he desired. There were likely to be limits on his capacity to make choices with regard to his occupation and sexual partners as well. Slavery was usually, but not always, involuntary. If not all of these characterizations in their most restrictive forms applied to a slave, the slave regime in that place is likely to be characterized as "mild"; if almost all of them did, then it ordinarily would be characterized as "severe."

Sources for slaves

Slaves were generated in many ways. Probably the most frequent was capture in war, either by design, as a form of incentive to warriors, or as an accidental by-product, as a way of disposing of enemy troops or civilians. Others were kidnapped on slave-raiding or piracy expeditions. Many slaves were the offspring of slaves. Some people were enslaved as a punishment for crime or debt, others were sold into slavery by their parents, other relatives, or even spouses, sometimes to satisfy debts, sometimes to escape starvation. A variant on the selling of children was the exposure, either real or fictitious, of unwanted children, who were then rescued by others and made slaves. Another source of slavery was self-sale, undertaken sometimes to obtain an elite position, sometimes to escape destitution.

Slavery existed in a large number of past societies whose general characteristics are well-known. It was rare among primitive peoples, such as the hunter-gatherer societies, because for slavery to flourish, social differentiation or stratification was essential. Also essential was an economic surplus, for slaves were often consumption goods who themselves had to be maintained rather than productive assets who generated income for their owner. Surplus was also essential in slave systems where the owners expected economic gain from slave ownership.

Ordinarily there had to be a perceived labour shortage, for otherwise it is unlikely that most people would bother to acquire or to keep slaves. Free land, and more generally, open resources, were often a prerequisite for slavery; in most cases where there were no open resources, nonslaves could be found who would fulfill the same social functions at lower cost. Last, some centralized governmental institutions willing to enforce slave laws had to exist, or else the property aspects of slavery were likely to be chimerical. Most of these conditions had to be present in order for slavery to exist in a society; if they all were, until the abolition movement of the 19th century swept throughout most of the world, it was almost certain that slavery would be present. Although slavery existed almost everywhere, it seems to have been especially important in the development of two of the world's major civilizations, Western (including ancient Greece and Rome) and Islāmic.

Two types of slavery

There have been two basic types of slavery throughout recorded history. The most common has been what is called household, patriarchal, or domestic slavery. Although domestic slaves occasionally worked outside the household, for example, in haying or harvesting, their primary function was that of menials who served their owners in their homes or wherever else the owners might be, such as in military service. Slaves often were a consumption-oriented status symbol for their owners, who in many societies spent much of their surplus on slaves. Household slaves sometimes merged in varying degrees with the families of their owners, so that boys became adopted sons or women became concubines or wives who gave birth to heirs. Temple slavery, state slavery, and military slavery were relatively rare and distinct from domestic slavery, but in a very broad outline they can be categorized as the household slaves of a temple or the state.

The other major type of slavery was productive slavery. It was relatively infrequent and occurred primarily in classical Athenian Greece and Rome and in the post-Columbian circum-Caribbean New World. It also was found in 9th-century Iraq, among the Kwakiutl Indians of the American Northwest, and in a few areas of sub-Saharan Africa in the 19th century. Although slaves also were employed in the household, slavery in all of those societies seems to have existed predominantly to produce marketable commodities in mines or on plantations.

A major theoretical issue is the relationship between productive slavery and the status of a society as a slave or a slave-owning society. In a slave society, slaves composed a significant portion (at least 20–30 percent) of the total population, and much of that society's energies were mobilized toward getting and keeping slaves. In addition the institution of slavery had a significant impact on the society's institutions, such as the family, and on its social thought, law, and economy. It seems clear that it was quite possible for a slave society to exist without productive slavery; the known historical examples were concentrated in Africa and Asia. It is also clear that most of the slave societies have been concentrated in Western (including Greece and Rome) and Islāmic civilizations. In a slave-owning society slaves were present, but in smaller numbers, and they were much less the focus of the society's energies.

Distinction between slavery and serfdom

Slavery was a species of dependent labour differentiated from other forms primarily by the fact that in any society it was the most degrading and most severe. Slavery was the prototype of a relationship defined by domination and power. But throughout the centuries man has invented other forms of dependent labour besides slavery, including serfdom, indentured labour, and peonage. The term serfdom is much overused, often where it is not appropriate (always as an appellation of opprobrium). In the past a serf usually was an agriculturalist, whereas, depending upon the society, a slave could be employed in almost any occupation. Canonically, serfdom was the dependent condition of much of the western and central European peasantry from the time of the decline of the Roman Empire until the era of the French Revolution. This included a "second enserfment" that swept over central and some of eastern Europe in the 15th and 16th centuries. Russia did not know the "first enserfment"; serfdom began there gradually in the mid-15th century, was completed by 1649, and lasted until 1906. Whether the term serfdom appropriately describes the condition of the peasantry in other contexts is a matter of vigorous contention. Be that as it may, the serf was also distinguished from the slave by the fact that he was usually the subject of the law—*i.e.,* he had some rights, whereas the slave, the object of the law, had significantly fewer rights. The serf, moreover, was usually bound to the land (the most significant exception was the Russian serf between about 1700 and 1861), whereas the slave was always bound to his owner; *i.e.,* he had to live where his owner told him to, and he often could be sold by his owner at any time. The serf usually owned his means of production (grain, livestock, implements) except the land, whereas the slave owned nothing, often not even the clothes on his back. The serf's right to marry off his lord's estate often was restricted, but the master's interference in his reproductive and family life ordinarily was much less than was the case for the slave. Serfs could be called upon by the state to pay taxes, to perform corvée labour on roads, and to serve in the army, but slaves usually were exempt from all of those obligations.

A person became an indentured servant by borrowing money and then voluntarily agreeing to work off the debt during a specified term. In some societies indentured servants probably differed little from debt slaves (*i.e.,* persons who initially were unable to pay off obligations and thus were forced to work them off at an amount per year specified by law). Debt slaves, however, were regarded as criminals (essentially thieves) and thus liable to harsher treatment. Perhaps as many as half of all the white settlers in North America were indentured servants, who agreed to work for someone (the purchaser of the indenture) upon arrival to pay for their passage. Some indentured servants alleged that they were treated worse than slaves; the economic logic of the situation was that slave owners thought of their slaves as a long-term investment whose value would drop if maltreated, whereas the short-term (typically four years) indentured servants could be abused almost to death because their masters had only a brief interest in them. Practices varied, but indenture contracts sometimes specified that the servants were to be set free with a sum of money, sometimes a plot of land, perhaps even a spouse, whereas for manumitted slaves the terms usually depended more on the generosity of the owner.

Peons

Peons were either persons forced to work off debts or criminals. Peons, who were the Latin-American variant of debt slaves, were forced to work for their creditors to pay off what they owed. They tended to merge with felons because people in both categories were considered criminals, and that was especially true in societies where money fines were the main sanction and form of restitution for crimes. Thus, the felon who could not pay his fine was an insolvent debtor. The debt peon had to work for his creditor, and the labour of the criminal peon was sold by the state to a third party. Peons had even less recourse to the law for bad treatment than did indentured servants, and the terms of manumission for the former typically were less favourable than for the latter.

HISTORICAL SURVEY

The origins of slavery are lost to human memory. It is sometimes hypothesized that at some moment it was decided that persons detained for a crime or as a result of warfare would be more useful if put to work in some way rather than if killed outright and discarded or eaten. But both if and when that first occurred is unknown.

Slavery in China

Slave-owning societies. Slavery is known to have existed as early as the Shang dynasty (18th–12th century BC) in China. It has been studied thoroughly in ancient Han China (206 BC–AD 25), where perhaps 5 percent of the population was enslaved. Slavery continued to be a feature of Chinese society down to the 20th century. For most of that period it appears that slaves were generated in the same ways they were elsewhere, including capture in war, slave raiding, and the sale of insolvent debtors. In addition, the Chinese practiced self-sale into slavery, the sale of women and children (to satisfy debts or because the seller could not feed them), and the sale of the relatives of executed criminals. Finally, kidnapping seems to have produced a regular flow of slaves at some times. The go-between or middleman was an important figure in the sale of local people into slavery; he provided the distance that made such slaves into outsiders, for the purchasers did not know their origins. Chinese family boundaries were relatively permeable, and some owners established kinlike relations with their slaves; male slaves were appointed as heirs when no natural offspring existed. As was also the case in other slave-owning societies, slaves in China were often luxury consumption items who constituted a drain on the economy. The reasons China never developed into a slave society are many and complex, but certainly an abundance of non-slave labour at low prices was one of the major ones.

Korea had a very large slave population, ranging from a third to half of the entire population for most of the millennium between the Silla period and the mid-18th century. Most of the Korean slaves were indigenously generated. In spite of their numbers, slaves seem to have had little impact on other institutions, and thus the society can be categorized as a slave-owning one.

Slavery existed in ancient India, where it is recorded in the Sanskrit Laws of Manu of the 1st century BC. The institution was little documented until the British colonials in the 19th century made it an object of study because of their desire to abolish it. In 1841 there were an estimated 8,000,000 or 9,000,000 slaves in India, many of whom were agrestic or predial slaves, that is, slaves who were attached to the land they worked on but who nevertheless could be alienated from it. Malabar had the largest proportion of slaves, about 15 percent of the total population. The agrestic slaves initially were subjugated communities. The remainder of the slaves was recruited individually by purchase from dealers or parents or by self-sale of the starving, and they can be classified as household slaves. Slavery in Hindu India was complicated by the slave owners' ritual need to know the origins of their slaves, which explains why most of them were of indigenous origin. Although there were exceptions, slaves were owned primarily for prestige.

Slavery was widely practiced in other areas of Asia as well. A quarter to a third of the population of some areas of Thailand and Burma were slaves in the 17th through the 19th centuries and in the late 19th and early 20th centuries, respectively. But not enough is known about them to say that they definitely were slave societies.

Other societies in the Philippines, Nepal, Malaya, Indonesia, and Japan are known to have had slavery from ancient until fairly recent times. The same was true among the various peoples inhabiting the Asian parts of the modern Soviet Union: the peoples of Sogdiana, Khorezm, and other advanced civilizations of Central Asia; the Mongols, the Kalmyks, the Kazakhs; and the numerous Turkic peoples, most of whom converted to Islām.

Slave-owning societies in the New World

In the New World some of the best-documented slave-owning societies were the Klamath and Pawnee and the fishing societies, such as the Yurok, that lived along the coast from what is now Alaska to California. Life was easy in many of those societies, and slaves are known to have sometimes been consumption goods that were simply killed in potlatches.

Other Amerindians, such as the Creek of Georgia, the Comanche of Texas, the Callinago of Dominica, the Tupinambá of Brazil, the Inca of the Andes, and the Tehuelche of Patagonia, also owned slaves. Among the Aztecs of Mexico, slavery generally seems to have been relatively mild. People got into the institution through self-sale and capture and could buy their way out relatively easily. Slaves were often used as porters in the absence of draft animals in Mesoamerica. The fate of other slaves was less pleasant: chattels purchased from the Mayans and others were sacrificed in massive numbers. Some of the sacrifices may have been eaten by the social elite.

In England about 10 percent of the population entered in Domesday Book in 1086 were slaves, with the proportion reaching as much as 20 percent in some places. Slaves were also prominent in Scandinavia during the Viking era, AD 800–1050, when slaves for use at home and for sale in the international slave markets were a major object of raids. Slaves also were present in significant numbers in Scandinavia both before and after the Viking era.

Continental Europe—France, Germany, Poland, Lithuania, and Russia—all knew slavery. Rus was essentially founded as a by-product of slave raiding by the Vikings passing from Scandinavia to Byzantium in the 9th century, and slavery remained a major institution there until the early 1720s, when the state converted the household slaves into house serfs in order to put them on the tax rolls. House serfs were freed from their lords by an edict of Tsar Alexander II in 1861. Many scholars argue that the Soviets reinstituted a form of state slavery in the Gulag camps that flourished until 1956.

Slavery was much in evidence in the Middle East from the beginning of recorded history. It was treated as a prominent institution in the Babylonian Code of Hammurabi of *c.* 1750 BC. Slaves were present in ancient Egypt and are known to have been murdered to accompany their deceased owners into the afterlife. It once was believed that slaves built the great pyramids, but contemporary scholarly opinion is that the pyramids were constructed by peasants when they were not occupied by agriculture. Slaves are also mentioned prominently in the Bible among the Hebrews in Palestine and their neighbours.

Slaves were owned in all Islāmic societies, both sedentary and nomadic, ranging from Arabia in the centre to North Africa in the west and to what is now Pakistan and Indonesia in the east. Some Islāmic states, such as the Ottoman Empire, the Crimean Khanate, and the Sokoto caliphate, must be termed slave societies because slaves there were very important numerically as well as a focus of the polities' energies.

Slavery in black Africa

Slaves have been owned in black Africa throughout recorded history. In many areas there were large-scale slave societies, while in others there were slave-owning societies. Slavery was practiced everywhere even before the rise of Islām, and black slaves exported from Africa were widely traded throughout the Islāmic world. Approximately 18,000,000 Africans were delivered into the Islāmic trans-Saharan and Indian Ocean slave trades between 650 and 1905. In the second half of the 15th century Europeans began to trade along the west coast of Africa, and by 1867 between 7,000,000 and 10,000,000 Africans had been shipped as slaves to the New World. Although some areas of Africa were depleted by slave raiding, on balance the African population grew after the establishment of the transatlantic slave trade because of new food crops introduced from the New World, particularly manioc, corn (maize), and possibly peanuts (groundnuts). The relationship between African and New World slavery was highly complementary. African slave owners demanded primarily women and children for labour and lineage incorporation and tended to kill males because they were troublesome and likely to flee. The transatlantic trade, on the other hand, demanded primarily adult males for labour and thus saved from certain death many adult males who otherwise would have been slaughtered outright by their African captors. After the end of the transatlantic trade, a few African societies at the end of the 19th century put captured males to productive work as slaves, but this usually was not the case before that time.

Athenian slave society

Slave societies. The first known major slave society was that of Athens. In the early Archaic period the elite worked its estates with the labour of fellow citizens in bondage (often for debt). After the lawgiver Solon abolished citizen slavery about 594 BC, wealthy Athenians came to rely on enslaved peoples from outside Attica. The prolonged wars with the Persians and other peoples provided many slaves, but the majority of slaves were acquired through regular trade with non-Greek peoples around the Aegean. At the time of classical Athens (the 5th through the 3rd centuries BC) slaves constituted about a third of the population. A particularly noteworthy locus of slave employment was the Laurium silver mines, where private individuals could pick out a lode and put their slaves to mining it. As in all other slave societies, it was the profitability of slavery that determined its preeminence in Athens. (Also important were political conditions that made the gross exploitation of citizens impossible.) Slaves were responsible for the prosperity of Athens and the leisure of the aristocrats, who had time to create the high culture now considered the beginning of Western civilization. The existence of large-scale slavery was also responsible, it seems logical to believe, for the Athenians' thoughts on freedom that are considered a central part of the Western heritage. Athenian slave society was finally destroyed by Philip II of Macedonia at the battle of Chaeronea (338 BC), when, on the motion of Lycurgus, many (but not all) slaves were freed.

Roman slave society

The next major slave society was Roman Italy between about the 2nd century BC and the 4th century AD. Initially Rome was a polity consisting primarily of small farmers. But the process of creating the empire took them away from their farms for extended periods, and the prolonged wars of conquest in Spain and the eastern Mediterranean during the 3rd and 2nd centuries BC created a great flood of captives. Nothing was more logical than to put the captives to work farming, especially the olives and grapes that created much of the prosperity of the late republic and the principate. Slaves and freedmen were responsible for much of the empire's commodity production, and in the early principate they ran its governmental bureaus as well. The conditions were right to put the captives to work: private ownership of land; developed commodity production and markets; a perceived shortage of internal labour supply; and an appropriate moral, political, and legal climate. Roughly 30 percent of the population was enslaved. Roman slave society ended as the slaves were legally converted into coloni, or serfs, and the lands became populated and the frontiers so remote that finding great numbers of outsider slaves was increasingly difficult.

Some lesser Islāmic slave societies are also of interest. One is the Baghdad caliphate founded in the 7th and lasting through the 10th century. Many tens of thousands of military captives were imported from Sogdiana, Khazaria, and other Central Asian locales. In the 9th and 10th centuries several tens of thousands of black Zanj slaves were imported from Zanzibar to Lower Iraq, where they constituted more than half the total population and were put to work to clear saline lands for irrigation and to cultivate sugar. More long-term was the slavery practiced in the Crimean Khanate between roughly 1475 and its liquidation by the Russian empress Catherine the Great in 1783. The Crimean Tatar society was based on raiding the neighbouring Slavic and Caucasian sedentary societies and selling the captives into the slave markets of Eurasia. Approximately 75 percent of the Crimean population consisted of slaves or freedmen, and much of the free population was highly predatory, engaged either in the gathering of slaves or in the selling of them. It is known that for every slave the Crimeans sold in the market, they killed outright several other people during their raids, and a couple more died on the way to the slave market. The reasons for the transition of the Crimean Khanate from a slave-owning society to a slave society have not been studied in detail. Probable reasons, however, include the combination of high demand for slaves throughout the Islāmic world, the defenselessness of the sedentary agricultural Slavs and others, and the existence of a relatively poor class of Crimean horsemen, who were led by a predatory elite that got rich by slave raiding. Crimean Tatar slave raids into Muscovy were greatly curtailed by the building of a series of walls along the frontier in the years 1636–53 and ultimately by the liquidation of the khanate in 1783.

Slavery in the Ottoman Empire

It is probable that the Ottoman Empire, and especially its centre in Turkey, should be termed a slave society. Slaves from both the white Slavic north and the black African south flowed into Turkish cities for half a millennium after the Turks seized control of much of the Balkans in the 14th century. The proportion of the population that was slave ranged from about one-fifth in Istanbul, the capital, to much less in remoter provincial areas. Perhaps only people such as the slave owners of the circum-Caribbean sugar islands and the American South were as preoccupied with slaves as were the Ottomans.

Slaves in the Ottoman Empire served in various capacities. They were janissary soldiers (see below), and they ran the empire, manned its ships, generated much of its handicraft product, and served as domestic servants and in harems. Contemporaries believed that the absolute power of the ruler was based on his military and administrative slaves. The Tanzimat enlightenment movement of the mid-19th century initiated the abolition of slavery; by the 1890s only a few slaves were being smuggled illegally into the empire, and the slave population was greatly reduced.

Other prominent Islāmic slave societies were on the east coast of Africa in the 19th century. The Arab-Swahili slave systems have been well-studied, and it is known that, depending on the date, 65 to 90 percent of the population of Zanzibar was enslaved. Close to 90 percent of the population on the Kenya coast was also enslaved, and in Madagascar half the population was enslaved. It may be assumed that similar situations prevailed elsewhere in the vicinity and also earlier, but studies to verify the proposition have not been undertaken.

Another notable Islāmic slave society was that of the Sokoto caliphate formed by Hausas in sub-Saharan Africa (northern Nigeria and Cameroon) in the 19th century. At least half the population was enslaved. That was only the most notable of the Fulani jihad states of the western and central Sudan, where between 1750 and 1900 from one- to two-thirds of the entire population consisted of slaves. In Islāmic Ghana, between 1076 and 1600, about a third of the population were slaves. The same was true among other early states of the western Sudan, including Mali (1200–1500), Segou (1720–1861), and Songhai (1464–1720). It should be noted that slavery was prominent in Ghana and Mali, and presumably elsewhere in Africa in areas for which information is not available, long before the beginnings of the transatlantic slave trade. The population of the notorious slave-trading state of the central Sudan, Ouidah (Whydah), was half-slave in the 19th century. It was about a third in Kanem (1600–1800) and perhaps 40 percent in Bornu (1580–1890). Most slaves probably were acquired by raiding neighbouring peoples, but others entered slavery because of criminal convictions or defaulting on debts (often not their own); subsequently, many of those people were sold into the international slave trade. After the limiting and then abolition of the transatlantic slave trade, a number of these African societies put slaves to work in activities such as mining gold and raising peanuts, coconuts (palm oil), sesame, and millet for the market.

Among some of the various Islāmic Berber Tuareg peoples of the Sahara and Sahel, slavery persisted at least until 1975. The proportions of slaves ranged from around 15 percent among the Adrar to perhaps 75 percent among the Gurma. In Senegambia, between 1300 and 1900, about a third of the population consisted of slaves. In Sierra Leone in the 19th century close to half the population was enslaved. In the Vai Paramount chiefdoms in the 19th century as much as three-quarters of the population consisted of slaves. Among the Ashanti and Yoruba a third were enslaved. In the 19th century over half the population consisted of slaves among the Duala of the Cameroon, the Ibo and other peoples of the lower Niger, the Kongo, and the Kasanje kingdom and Chokwe of Angola.

Caribbean slave societies

The best-known slave societies were those of the circum-Caribbean world. Slave imports to the islands of the Caribbean began in the early 16th century. Initially the islands often were settled as well by numerous indentured labourers and other Europeans, but following the triumph after 1645 of the sugar revolution (initially undertaken because superior Virginia tobacco had left the Barbadian planters with nothing to sell) and after the nature of the disease climate became known to Europeans, they came to be inhabited almost exclusively by imported African slaves. In time the estate owners moved to England, and the sugar plantations were managed by sometimes unstable and unsavoury Europeans who, with the aid of black overseers and drivers, controlled masses of slaves. About two-thirds of all slaves shipped across the Atlantic ended up in sugar colonies. By 1680 in Barbados the average plantation had about 60 slaves, and in Jamaica in 1832 about 150. The sugar plantations were among the contemporary world's largest and most profitable enterprises, paying about 10 percent on invested capital and on some occasions, such as in Barbados in the 1650s, as much as 40 to 50 percent. The proportions of slaves on the islands ranged from more than a third in Cuba, which went into the sugar and gang-labour business on a large scale only after the local planters had gained control in 1789, to 90 percent and more on Jamaica in 1730, Antigua in 1775, and Grenada up to 1834.

Slaves were of varying importance in Mesoamerica and on the South American continent. Initially slaves were imported because of a labour shortage, aggravated by the high death rate of the indigenous population after the introduction of European diseases in the early 16th century. They were brought in at first to mine gold, and they were shifted to silver mining or simply let go when gold was exhausted in the mid-16th century. In Brazil, where sugar had been tried even before its planting in the Caribbean, the coffee bush was imported from Arabia or Ethiopia via Indonesia, and it had an impact similar to that of sugar in the Caribbean. Around 1800 about half the population of Brazil consisted of slaves, but that percentage declined to about 33 percent in 1850 and to 15 percent after the shutting off of imports around 1850 combined with free immigration to raise the proportion of Europeans. In some parts of Brazil, such as Pernambuco, some two-thirds of the population consisted of Africans and their offspring.

The final circum-Caribbean slave society was what became the southern United States. Slaves first were brought to Virginia in 1619. Subsequently, Africans were transshipped to North America from the Caribbean in increasing numbers. Initially, however, the English relied for their dependent labour primarily on indentured servants from the mother country. But in the two decades of the 1660s and 1670s the laws of slave ownership were clarified (for example, Africans who converted to Christianity did no longer have to be manumitted), and the price of servants may have increased because of rising wage rates in prospering England; soon thereafter African slaves replaced English indentured labourers. Tobacco initially was the profitable crop that occupied most slaves in the Chesapeake. The invention of the cotton gin by Eli Whitney in 1793 changed the situation, and thereafter cotton culture created a huge demand for slaves, especially after the opening of the New South (Alabama, Mississippi, Louisiana, and Texas). By 1850 nearly two-thirds of the plantation slaves were engaged in the production of cotton. Cotton could be grown profitably on smaller plots than could sugar, with the result that in 1860 the average cotton plantation had only about 35 slaves, not all of

Cotton culture and slavery

whom produced cotton. During the reign of "King Cotton," about 40 percent of the Southern population consisted of black slaves; the percentage of slaves rose as high as 64 percent in South Carolina in 1720 and 55 percent in Mississippi in 1810 and 1860. More than 36 percent of all the New World slaves in 1825 were in the southern United States. Like Rome and the Sokoto caliphate, the South was totally transformed by the presence of slavery. Slavery generated profits comparable to those from other investments and was only ended as a consequence of the War Between the States.

The international slave trade. Organized commerce began in the Neolithic Period, and it may be assumed that slaves were not far behind high-value items such as amber and salt in becoming commodities. Even among relatively simple peoples one can trace the international slave trade. Thus such a trade was going on among the peoples of Siberia before the arrival of the Russians in the 16th and 17th centuries. The slaves so traded were neighbouring people captured in warfare, who were then shipped to distant points where they would be without kin and whence they would be unlikely to flee. Similar commerce in slaves occurred on nearly all continents and provided the bulk of household slaves throughout the world.

Slave raiding

The international slave trades that provided much of the chattel for the slave societies flowed out of the great "population reservoirs." Two such reservoirs were the Slavs and contiguous agriculturalist Iranians from antiquity to the 19th century and the sub-Saharan Africans from around the beginning of the Christian Era to the middle of the 20th century. A third such reservoir probably was the Germanic, Celtic, and Romance peoples who lived north of the Roman Republic and Empire and who half a millennium later became the victims of the Vikings' slave raids. The dynamics of these raids were as follows: A large demand for slave labour prompted neighbouring peoples (typically migratory or nomadic in habit) to prey on the sedentary agriculturalists living in the reservoir. The raiders developed techniques, of which surprise was perhaps the major one, that put the settled peoples at a disadvantage, for they never knew when and where the raiders might strike. Populations in the reservoir could be completely depleted, as happened to the East Slavs living in the steppe south of the Oka and between the Volga and the Dnepr rivers from 1240 to the 1590s, or they could migrate half a continent away to escape the slave raiders, as did the Ndembu in Africa. Ruthenians, frontier Poles, Caucasians, and numerous African peoples were sorely depleted by slave raids. One alternative was to fight back, as did the Muscovite Russians and the Baya of Adamawa (now northern Cameroon in West Africa), and the consequence in both instances was the creation of an authoritarian garrison state.

The international slave trades developed into elaborate networks. For example, in the 9th and 10th centuries Vikings and Russian merchants took East Slavic slaves into the Baltic. They were then gathered in Denmark for further transshipment and sold to Jewish and Arabic slave traders, who took them to Verdun and León. There some of the males were castrated. From those places the slaves were sold to harems throughout Moorish Spain and North Africa. In the 9th century the Baghdad caliphate got slaves from western Europe via Marseille, Venice, and Prague; Slavic and Turkic slaves from eastern Europe and Central Asia via Derbent, Itil, Khorezm, and Samarkand; and African slaves via Mombasa, Zanzibar, the Sudan, and the Sahara. The Mongols in the 13th century brought their slaves first to Karakorum, whence they were sold throughout Asia, and then later to Sarai on the Lower Volga, whence they were retailed throughout much of Eurasia. Following the breakup of the Golden Horde, the Crimean Tatars took their chattel to Kefe (Feodosiya) in the Crimea, whence it was transported across the Black Sea and sold throughout the Ottoman Empire and elsewhere. Arabs developed similar supply networks out of black Africa across the Sahara, across the Red Sea (from Ethiopia and Somalia), and out of East Africa, which supplied the Islāmic world and the Indian Ocean region with human chattel.

The transatlantic slave trade

Beginning around 1500 a similar process occurred along the coast of West Africa to supply the transatlantic slave trade. The Africans were captured by other Africans in raids and then transported to the coast; one may assume that the number of casualties of African slave raiding was nearly as high as that of Crimean Tatar slave raiding. The captives, primarily adult males, were assembled on the coast by African rulers and kept in holding pens until wholesaled to European ship captains who sailed up and down the coast looking for slave cargo. (As stated above, the women and children often were not sent to the coast for export but were kept by the Africans themselves, often for incorporation into their lineages.) African rulers, who did not allow the Europeans to move inland, often conducted their wholesale business on the coast, such as at Ouidah in Dahomey (now Benin). (Because of the disease climate the Europeans also were reluctant, even unable, to move inland until the mid-19th century.) But African rulers did everything they could to encourage the European sea captains to come to their port.

Once a ship was loaded, the trip, known as "the Middle Passage," usually to Brazil or an island in the Caribbean, was a matter of a few weeks to several months. Between 1500 and the end of the 19th century the time of the voyage diminished considerably. That change was important, because death rates, which ranged from around 10 to more than 20 percent on the Middle Passage, were directly proportional to the length of the voyage. The ship captains had every interest in the health of their cargo, for they were paid only for slaves delivered alive. The death rates among the European captains and crew engaged in the slave trade were at least as high as those among their cargo on the Middle Passage. Of the slave-ship crews that embarked from Liverpool in 1787, less than half returned alive.

Arriving in Brazil or the Caribbean islands, the slaves were sold at auction. The slave auctions were elaborate markets in which the prices of the slaves were determined. The auctions told the captains and their superiors what kind of cargo was in demand, usually adult males. Credit almost always was part of the transaction, and inability to collect was one of the major reasons companies went bankrupt. After the auction the slave was delivered to the new owner, who then put him to work. That also began the period of "seasoning" for the slave, the period of about a year or so when he either succumbed to the disease environment of the New World or survived it. Many slaves landed on the North American mainland before the early 18th century had already survived the seasoning process in the Caribbean.

It can be assumed that the other international slave trades were comparable in many respects to the transatlantic one, but they have not been adequately studied.

Ways of ending slavery. Slavery came to an end in numerous ways. Household slavery ended because of an exhaustion of supplies, because slavery evolved into some other system of dependent labour, because it withered away, or because it was formally abolished. Productive slavery came to an end for the additional reasons that it ceased to be profitable or that it was abolished by warfare or the threat of warfare.

Attitudes toward slavery

Throughout history there have been people who in one way or another believed that slavery was not a good or natural condition. Jean Bodin (1530–96), the French founder of antislavery thought, for example, condemned the institution as immoral and counterproductive and advocated that no group of men should be excluded from the body politic. Nevertheless, remarkably few people found the institution of slavery to be unnatural or immoral until the second half of the 18th century. Until that time Christians commonly thought of sin as a kind of slavery rather than slavery itself as a sin. When concern was expressed for slaves, it was for their good care, not for their unfree status.

Frequently, when slavery passed from the scene, it did so with little fanfare. In most societies, such as ancient Babylonia, Israel, Egypt, or Athens, the institution of slavery had little or no connection with the society's rise or demise. In Rome, on the other hand, slavery began to yield to tenancy and the antecedents of serfdom before the fall of the empire, as the diminishing supply of slaves and the rise of their price coincided with the disintegration of the olive oil- and wine-producing plantations of southern Italy and loss of markets in the face of competition from Spain, Gaul, and North Africa. (This standard interpretation has been challenged, however.) In the Eastern Roman Empire (Byzantium) serfdom was the predominant form of dependent labour, and slavery was definitely secondary. Manumitting slaves became much easier, according to the laws, and the Ecloga and the Procheiron Nomos (see below) prescribed that the slaves of persons who died without testament had to be freed. Throughout most of Europe household slavery persisted well into the late Middle Ages and even later and only gradually died out. Slavic slaves were plentiful, for example, in the Italian city-states as late as the 14th century, and African slaves could be found in Spain and Portugal in the 16th century. Serfdom replaced slavery in medieval Germany. By the end of the Middle Ages slavery no longer existed in England, and the famous Cartwright decision of the reign of Elizabeth I (1569) held that "England was too pure an air for slaves to breathe in."

Enserfment in eastern Europe

Slavery persisted longer in eastern Europe. In Poland it was replaced by the second enserfment; the sale and purchase of slaves were forbidden in the 15th century. A similar process occurred in Lithuania, where slavery was formally abolished in 1588. In Russia it came to an end with the first enserfment: agricultural slaves were formally converted into serfs in 1679, and household slaves were converted into house serfs in 1723. In the Caucasus and in Central Asia slavery persisted until the second half of the 19th century. As the Russian Empire grew and its hegemony spread, it adopted the tendency of 19th-century imperialist powers to enforce abolition when embarking upon colonization. Thus the conquest of the Caucasus led to the abolition of slavery by the 1860s and the conquest in Central Asia of the Islāmic khanates of Bukhara, Samarkand, and Khiva by the 1870s.

Abolition movement

Slavery in England was formally abolished by judicial fiat in 1772, when Chief Justice Mansfield held that the captured fugitive slave James Somersett, from Virginia, could not be claimed by his owner and that furthermore any slave by the act of walking on English soil became free. The fate of slavery in most of the rest of the world depended on the British abolition movement, which was initiated by the English Quakers in 1783 when they presented the first important antislavery petition to Parliament. They were following the Pennsylvania Quakers, who had voiced opposition to slavery in 1688. The Vermont constitution of 1777 was the first document in the United States to abolish slavery. Another sign of the spread of antislavery feeling was the declaration in the U.S. Constitution that the importation of slaves could be forbidden after 20 years (in 1808). An act of March 2, 1807, forbade trading in slaves with Africa. Well before the rise of cotton some people hoped that natural processes combined with a prohibition on infusions would put an end to slavery.

In 1807 the British abolished the slave trade with their colonies. In the Caribbean, slavery was abolished by British Parliamentary fiat, effective July 31, 1834, when 776,000 slaves in the British plantation colonies were freed. The British imperial emancipation can be attributed to the growing power of the philanthropic movement and a double switch in the focus of the British Empire, geographically from west (the Caribbean) to east (India) and economically from protectionism to laissez-faire.

The British move in 1807 to abolish the slave trade had an immediate impact on the juntas struggling for independence in Spanish America. The slave trade was declared illegal in Venezuela and Mexico in 1810, in Chile in 1811, and in Argentina in 1812. In 1817 Spain signed a treaty with Britain agreeing to abolish the slave trade in 1820, but the trade continued to the remaining Spanish colonies until 1880. Chile freed its black slaves in 1823; Mexico abolished slavery in 1829, and Peru in 1854.

The American antislavery movement, linked to the "Second Great Awakening," succeeded in arousing immense hostility between the non-slave North, where most states had voluntarily abolished slavery by 1804, and the slaveholding South, where the "peculiar institution" became even further entrenched because of the spread of cotton cultivation. By the 1850s, however, the old abolition movement had flagged. It took political developments and forces (especially the emergence of the Free-Soil movement and the conflict over the expansion of slavery), the South's secession, the Civil War, and Abraham Lincoln's Emancipation Proclamation on Jan. 1, 1863, to put slavery on the road to extinction in the United States. The proclamation was confirmed by the Thirteenth Amendment to the Constitution, which put an end to slavery.

Abolition of slavery in Brazil

Puerto Rico abolished slavery (with provisions for periods of apprenticeship) in 1873 and Cuba in 1880. Brazil was the last Western Hemisphere nation to abolish slavery. The British antislavery movement of the 1810s had almost put an end to the institution, but a thriving world market for coffee revitalized it in the 1820s. In 1850 Britain declared that a squadron would enter Brazilian territorial waters to seize vessels carrying slaves, and later that year Brazil responded by equating the slave trade with piracy. On May 13, 1888, all Brazilian slaves were manumitted. Initially there was some opposition by the coffee growers, but their resistance crumbled immediately.

The European colonization movement of the second half of the 19th century put an end to slavery in many parts of Africa, East Asia, and Southeast Asia. The abolition of slavery in both Hindu and Muslim India by Act V of 1843 meant only that the British courts would not enforce claims to a slave, but the Penal Code of 1861 made holding a slave a crime. Having seen to the abolition of slavery in most of Latin America and South Asia, the British turned their attention back to Africa. They moved onto the continent, took control of those governments that were thriving on slavery, and attempted to abolish the institution. Lagos was annexed in 1861, and all of Nigeria followed. In the 1870s British missionaries moved into Malaŵi, the place of origin of the Indian Ocean Islāmic slave trade, in an attempt to interdict it at its very source. In 1890 Zanzibar was made a British protectorate after the sultan's authority had been destroyed by the struggle over the slave trade. In Dahomey the French abolition of slavery resulted in the cessation of ceremonial human sacrifice.

The Imperial government formally abolished slavery in China in 1906, and the law became effective on Jan. 31, 1910, when all adult slaves were converted into hired labourers and the young were freed upon reaching age 25. Slavery was legally abolished in Korea in the Gap-o reform of 1894 but remained extant in reality until 1930.

Abolition efforts in the 20th century

Some parts of Africa and much of the Islāmic world retained slavery at the end of World War I. For this reason the League of Nations and later the United Nations took the final extinction of slavery to be one of their obligations. The league had considerable success in Africa, with the assistance of the colonial powers, and by the late 1930s slavery was abolished in Liberia and Ethiopia. After World War II the United Nations Universal Declaration of Human Rights and the European Convention of Human Rights proclaimed the immorality and the illegality of slavery. Slavery was abolished in most Islāmic countries, although it persisted in Saudi Arabia into the 1960s. It finally was made illegal in the Arabian Peninsula in 1962. It is probable that slavery no longer exists as a legal phenomenon recognized by a political authority or government any place in the world.

THE LAW OF SLAVERY

Sources of slavery law. By definition slavery must be sanctioned by the society in which it exists, and such approval is most easily expressed in written norms or laws. Thus it is not accidental that even the briefest code of a relatively uncomplicated slave-owning society was likely to contain at least a few articles on slavery.

Both slave-owning and slave societies that were part of the major cultural traditions borrowed some of their laws about slavery from the religious texts of their respective civilizations. Principles regarding slavery that proved to be either unprofitable or unworkable were among the first to be discarded. An obvious example is provided by the Old Testament law that Hebrew slaves were to be manumitted after six years (Exodus 21:2; Deuteronomy 15:12). A similar general recommendation that slaves be freed after six years in bondage was adhered to by many Islāmic slave-owning societies; it helps to account for the ferocity and frequency of their slave raids, for they had a need for constant replenishment of their slave supplies. In Christian slave societies, on the other hand, the principle that the tenure of slavery should be limited was almost completely ignored.

Practically every society that possessed slaves wrote about them in its laws, and thus only a few codes can be mentioned here. The ancient Mesopotamian laws of Eshnunna (*c.* 1900 BC) and the Code of Hammurabi had a number of articles devoted to slavery, as did the Pentateuch. In ancient India the Laws of Manu of the 1st century BC contained numerous laws on slaves.

Roman slave law

Little is known about the Athenian law of slavery, but the Roman law of slavery was extraordinarily elaborate. Roman law was summed up in the great Pandects of Justinian of AD 533, and some of its slave norms later found their way into the Byzantine Ecloga (which incorporated Syrian norms as well) of AD 726 and, more deliberately, into the Procheiron Nomos of AD 867–879. Romano-Byzantine norms also found their way into the Bulgarian Court Law for the People ("Zakon Sudnyi Liudem") of the end of the 9th century and the 13th-century Ethiopian Fetha Nagast.

The European "barbarian" (Germanic) codes, which first appeared in the 5th century AD and remained in effect for about half a millennium, were derived from customary law influenced by Roman law. The slave statutes of the Russian Russkaya Pravda of the 11th–13th century were all clearly of native East Slavic origin. The same was true of the Muscovite court handbooks (Sudebniki) of 1497, 1550, 1589, and 1606. The Muscovite Russians had a special government office to deal with slavery matters, the Slavery Chancellery (1571–1704), and its practice became the basis of chapter 20 of the great Ulozhenie of 1649, which constituted 119 of the 967 articles of the code; other articles dealt with slavery as well.

Islāmic slave law

The Qur'ān was the fundamental starting point for Islāmic law (Sharīʿah), including the law of slavery. It was supplemented by the *ijmāʿ*, the scholarly legal consensus, and the *qiyās*, juristic reasoning by analogy. Islāmic law regulated in detail every part of the institution of slavery, from the jihad (holy war) and the distribution of booty to the treatment of slaves and emancipation. The last Islāmic slave law was promulgated in 1936 by King Ibn Saʿūd of Saudi Arabia, which restated the teachings of the Qur'ān. It also required owners to register slaves with the government and licensed slave traders.

Some sub-Saharan African societies followed Islāmic law; others had their own. The latter ordinarily were not systematized until the European colonization movement, and so their law of slavery was oral common law.

Slavery was a relatively prominent institution in the Chinese T'ang Code of the 7th century AD. Subsequently it was mentioned in every Chinese law down to the 20th century and was also important in the Korean legal system. The slavery norms of the Mongol Great Yassa of Genghis Khan were locally generated, but subsequent Mongol law reveals considerable influence of the T'ang Code.

The circum-Caribbean world had several basic laws of slavery. The slave law of the Spanish-speaking colonies and then independent countries was based on the Siete Partidas of 1263–65 of Alfonso X of Castile and Léon and the Spanish Slave Code of 1789. Another important code in Latin America was Louis XIV's Code Noir of 1685. The Louisiana Slave Code of 1824 was based on the Siete Partidas and the Code Napoléon.

The Danish Virgin Islands had two largely locally generated codes of 1733 and 1755, although they were approved by the colonial administration of Denmark. The English colonies were completely autonomous, for England had no law of slavery from which to borrow. The first code was that of Barbados of 1688, whose origins are unknown. It was imitated by the South Carolina code of 1740. Beginning with Virginia in 1662, each colony in North America worked out its own ex post facto law of slavery before independence, a process that continued after the creation of the United States and until the Civil War. Slavery is mentioned only three times and referred to at most 10 times (and then only indirectly) in the U.S. Constitution, and, except for a handful of measures on fugitives, there was no federal slave law. The basic protection for the institution of slavery was the Tenth Amendment of 1791, the reserved powers clause, which left the issue of slavery and other matters to the states.

Legal definitions of slavery. Some of the definitions of slavery discussed above were legal, but the majority were not. This section focuses exclusively on legal definitions of slavery. Most groups, whether national or religious, forbade the enslavement of their fellows; thus, the Spanish could not enslave Spaniards, Arabs could not enslave Arabs, and Christians and Muslims could not enslave their coreligionists. Legally the slave ordinarily had to be an outsider. In law the slave was usually defined as property, and the question then was whether he was movable property (chattel) or real property. In most societies he was movable property but in some, real property.

Some societies, such as Muscovy in the 16th and 17th centuries, had different legal categories of slaves. There some slaves were inherited, others were purchased forever, others for a limited time could become perpetual slaves, and still others for specific functions such as estate managers. Different varieties or gradations of slaves were found elsewhere as well, as in China and in certain African societies.

Master–slave legal relationships. The master–slave relationship was the cornerstone of the law of slavery, and yet it was an area about which the law often said very little. In many societies the subordination of the slave to his owner was supposed to be complete; in general, the more complete an owner's control over his slave, the less the law was likely to say about it.

The owner's right to kill his slave

A major touchstone of the nature of a slave society was whether or not the owner had the right to kill his slave. In most Neolithic and Bronze Age societies slaves had no such right, for slaves from ancient Egypt and the Eurasian steppes were buried alive or killed to accompany their deceased owners into the next world. Among the Northwest Coast Tlingit, slave owners killed their slaves in potlatches to demonstrate their contempt for property and wealth; they also killed old or unwanted slaves and threw their bodies into the Pacific Ocean. An owner could kill his slave with impunity in Homeric Greece, ancient India, the Roman Republic, Han China, Islāmic countries, Anglo-Saxon England, medieval Russia, and many parts of the American South before 1830.

The slave's right to life

That was not the case in other societies. The Hebrews, the Athenians, and the Romans under the principate restricted the right of slave owners to kill their human chattel. The Code of Justinian changed the definition of the slave from a thing to a person and prescribed the death penalty for an owner who killed his slave by torture, poison, or fire. Spanish law of the 1260s and 1270s denied owners the right to kill their slaves. Lithuanian and Muscovite law forbade the killing, maiming, or starving of a returned fugitive slave. Ch'ing Chinese law punished a master who killed his slave, and that punishment was more severe if the slave had done no wrong. The Aztecs under some circumstances put to death a slave owner who killed his slave. No society, on the other hand, had the slightest sympathy for the slave who killed his owner. Roman law even prescribed that all other slaves living under the same roof were to be put to death along with the slave who had committed the homicide.

Assault and general brutality were other concerns of the law of slavery. In antiquity slaves often had the right to take refuge in a temple to escape cruel owners, but that sometimes afforded little protection. The ancient Franks and the Germans warned owners against cruelty. The Code of Justinian and the Spanish Siete Partidas deprived cruel owners of their slaves, and that tradition went into the Louisiana Black Code of 1806, which made cruel punishment of slaves a crime. In modern societies brutality and sadistic murder of slaves by their owners were rarely condoned on the grounds that such episodes demoralized other slaves and made them rebellious, but few slave owners were actually punished for maltreating their slaves. In the American South 10 codes prescribed forced sale to another owner or emancipation for maltreated slaves. Nevertheless, cases such as *State* v. *Hoover* (North Carolina, 1839) and *State* v. *Jones* (Alabama, 1843) were considered sensational because slave owners were punished for savagely "correcting" their slaves to death.

Sexual relations between masters and slaves

It was not an axiom of the master–slave relationship that the former automatically had sexual access to the latter. That was indeed the case in most societies, ranging from the ancient Middle East, Athens, and Rome to Africa, all Islāmic countries, and the American South. Places such as Muscovy, however, forbade owners to rape their female slaves, while the Chinese and the Lombards forbade the raping of married slave women. More problematic were sexual relations between mistresses and male slaves. Athens and Rome both put the slave to death, and Byzantine law prescribed that the mistress was to be executed and the slave to be burned alive. The Danish Virgin Islands' laws of 1741, 1755, and 1783, in an attempt to protect northern Europeans from African "contamination," prescribed a fine of 2,000 pounds of sugar for a man who raped a black slave, and a white woman who had sexual relations with a black slave was to be fined, imprisoned, and then deported.

The labour and food regimes were central to almost every slave's life. In societies where the owner's control over his slave was total, such as the Roman Empire or the pre-1830 American South, the law said little or nothing about how long he could work him and whether his slave had a right to food and clothing. In South India the slave owner had an absolute right to whatever labour his slave was capable of rendering. In Muscovy, on the other hand, a slave owner was jailed for forcing his slaves to labour on Sunday. In Judea in 200 BC, in Sicily in 135–32 BC, and on the Nile in AD 46 regulations prescribed the food rations a slave could expect. The Lithuanian Statute of 1588 and the Russians in 1603 and 1649 decreed that slaves had a right to be fed. The Danish Virgin Islands in 1755 prescribed adequate food rations. The Alabama Slave Code of 1852 mandated that the owner had to provide slaves of working age a sufficiency of healthy food, clothing, attention during illness, and necessities in old age.

The slave's right of marriage and offspring

A major issue was whether the master had to allow the slave to marry and what rights the owner had over slave offspring. In general, a slave had far fewer rights to his offspring than to his spouse. Babylonian, Hebrew, Tibetan-speaking Nepalese Nyinba, Siamese, and American Southern slave owners thought nothing of breaking up both the conjugal unit and the nuclear family. Unexpectedly the 1755 Danish Virgin Islands Reglement prohibited separating minors from their parents. In Muscovy and China, slave owners could sell or will children apart from their parents, but marriages were inviolable.

The slave's right of property

In North America, India, Rome, Muscovy, most of the Islāmic world, and among the Tuareg a fundamental principle was that the slave could not own property because the master owned not only his slave's body but everything that body might accumulate. This did not mean, however, that slaves could not possess and accumulate property but only that their owners had legal title to whatever the slaves had. In a host of other societies, such as ancient and Roman Egypt, Babylonia, Assyria, Talmudic Palestine, Gortyn, much of medieval Germany, Thailand, Mongol and Ch'ing China, medieval Spain, and the northern Nigerian emirates, slaves had the right of property ownership. Some places, such as Rome, allowed slaves to accumulate, manage, and use property in a peculium that was legally revocable but could be used to purchase their freedom. This provision gave slaves an incentive to work as well as the hope of eventual manumission.

Considerable research has been done on the treatment of slaves, and the consensus is that, while the law may have spelled out the desired social standards of master–slave relations, it did not necessarily define the reality for any particular situation. Sadists, even psychopaths, who could not cope with their right of total dominance over another human being, might appear anywhere, as might kindly masters. More determining than the law were the conditions of the society itself. At one extreme, among the Tuareg of North Africa, the slave owners themselves often lived badly, and so, of course, did their slaves. At the other extreme, in the American South material conditions were sufficiently favourable to provide comparative comfort for both masters and slaves. Moreover, slaves born of already enslaved parents usually were treated much better than those purchased or captured from foreign groups. The treatment of slaves in expansive, dynamic societies was likely to be worse than in more stable ones.

Legal relationships between slave owners. There was more uniformity across systems regarding legal relationships between slave owners. All societies had provisions for the recovery of runaways, and most imposed sanctions on owners who stole others' slaves (a capital offense in some systems) or helped them to flee. There also were relatively uniform laws about passing slaves from one generation to another.

There was considerable variability among societies in the law of slave transactions. Whereas Roman-law societies had elaborate norms on contracts, Muscovy had essentially none. Whereas legal systems from Babylonia, Athens, Rome, early Germany, China, and Ethiopia to Islāmic societies and Louisiana allowed guarantees by the sellers that slaves would not flee, were free from disease, or had certain skills, no such laws existed in places such as Muscovy.

Legal relationships between slaves and free strangers. Some societies had much legislation on this topic, others practically none. Where the slave was completely dependent on his owner, few laws existed beyond the normal rules governing any form of property; it was the owner's responsibility to recover damages if a third party killed or assaulted either his cow or his slave. The owner, moreover, was held equally or even more responsible for the slave's actions, ranging from homicide to theft, than was the slave himself, for the society desired that the former control his property and there was no assurance that sanctions, especially money fines, could be enforced against slaves.

Attitudes toward homicide of a slave

Homicide of a slave by a stranger was a revealing test of a society's attitude toward the slave. In Mesopotamia and in Islāmic practice the killer of a slave merely had to compensate the owner for the loss of his property. Elsewhere, however, it was different. Roman law introduced the idea in the Lex Cornelia de Sicariis et Veneficis (the dictator Sulla's enactment on murders and poisoners of 81 BC) that a slave was a person and thus that killing a slave could be a crime. That provision found its way into the Code of Justinian. In North America in the period from 1770 to 1830 the killing of a slave was equated in common law with the murder of a white person. Laws were uniformly harsh when a slave killed a stranger who was a freeman.

Some societies did not allow third parties to assault slaves with impunity. In Muscovy, for example, a slave might have honour and could recover from a third party who injured his honour. Societies elsewhere, however, such as the North American Yurok, Tlingit, and other neighbouring Indians, as well as in the American South, explicitly stated that slaves could have no honour, personal status, or prestige. South Carolina law noted that the slave was not "within the peace of the state, and therefore the peace of the state [was] not broken by an assault and battery on him." Conversely, when a slave assaulted a freeman, the latter often recovered from the slave's owner. Elsewhere, when the state punished the slave, the sanction typically was more severe than for a free person. For example, in Ch'ing China a slave was punished one degree more severely than free citizens for offenses against a freeman.

Most societies, such as those in Athens, Rome, Kievan Rus, Thailand, and Louisiana, did not allow slaves to contract independently with third parties, although some allowed the slave to make a contract on his owner's behalf. The brutal deprivation of rights was expressed in the Alabama case *Creswell's Executor* v. *Walter* (1860); the slave, said the court, had "no legal mind, no will which the law can recognize. . . . Because they are slaves, they are incapable of performing civil acts." On the other hand, in a few societies, as in the ancient Middle East, slaves were allowed to contract with third parties. Roman slaves were allowed to make contracts in regard to third peculium.

A few societies, such as late Assyria and Muscovy, allowed slaves to testify in court, but most did not. It was a rare society that permitted a slave to serve as a witness against his owner, but some societies, such as ancient Nuzi and Muscovy, allowed slaves to testify against, even to sue, third parties. That was particularly likely to be the case when slaves played a major role in the society, because disputes could not be resolved by the freemen alone without resort to evidence provided by slaves.

Laws of manumission. Laws of manumission varied widely from society to society and within societies across time. They are often viewed as the litmus test of a particular society's views of the slave, that is, of the capacities the slave was likely to exhibit as a free human being. Many Islāmic societies, broadly interpreting the Hebrew prescription, generally prescribed that slave owners had to free their slaves after the passage of a number of years, essentially the length of time they considered it took for an "outsider" to become an "insider." Most other societies allowed masters to free their slaves whenever they wished, although there were exceptions. Some legal systems prescribed manumission when the slave adopted the religion of his owner. It is hardly surprising that manumission was more frequent in systems of household slavery, for intimate relations between master and slave soon converted the outsider into an insider. With notable exceptions, such as Athens, Rome, Muscovy, and some circum-Caribbean societies, many societies required manumission after three generations.

Manumission upon birth

Birth was occasionally a route to manumission. In thriving slave systems such as those of the New World, in harsh systems such as those among the Northwest Coast Indians and the medieval Germanic peoples, or even in milder systems such as those of the Chinese and the Muscovites, a slave's offspring simply added to the slave population. But that was not universally the case; African slave societies, such as the Dahomeans of West Africa, the Ashanti of Ghana, or the Azande living between the Congo and the Nile, prescribed that the offspring of slaves should be free, as part of the process of incorporation into a new lineage. Although Islāmic law did not require manumission upon birth, the Qur'ān recommended it, and slave owners were often inclined to follow the religious tenet. The Aztecs freed all children born in slavery except the offspring of traitors. In Thailand emancipation was considered a pious act, and at their death many owners freed their slaves.

Rates of manumission

The rate of manumission did not necessarily correspond to the legal ease of manumission. It should be noted, however, that in Rome manumission was relatively easy and was widely practiced, even though there was a 5 percent tax on manumission in the Republic, and the Lex Fufia Caninia of 2 BC forbade manumission by testament of more than a fifth to a half of one's slaves, depending on the number owned. In much of sub-Saharan Africa, manumission was common in most periods, and the freed person typically became a kind of relative in a process of assimilation. In Neo-Babylonia, in Late and Middle Assyria, and in Muscovy manumission was easy but rare; in the American South manumission was comparatively difficult and almost never happened after the prohibition on importing new slaves. The factors of institutional dynamism, expansionism, and profitability, as well as race (see below), may have been the most crucial variants for the South, where manumission was even forbidden in South Carolina in 1820, Mississippi in 1822, Arkansas in 1858, and Maryland and Alabama in 1860; other factors were at work in the ancient Middle East and Muscovy.

There was considerable variation among societies as to whether a slave was allowed to accumulate property that he might keep after manumission. One form of such accumulation was the Roman peculium, which legally belonged to the master. One of its heirs was called *coartación*, the self-purchase system, widely used 1,500 years later in Latin America.

After manumission, most societies prescribed a period of legal transition to freedom. In the Roman Empire, China, and elsewhere, this period took three generations and might mean that the grandchild of a slave owner (the "patron") was legally responsible for the grandchild of a slave (his "client"). Thereafter the descendants of the freedman became full members of society, although perhaps still despised. The reason for the legally mandated period of transition to freedom was clear: the slave initially was not a member of the society but an outsider (see below), and it took time to become integrated into the new society. Equally important, the slave was dependent on his owner, and it took time for the freedman and his heirs to become fully self-reliant members of society. If the slave owner and his heirs were not responsible for the freedmen, the fear was, as expressed in the Louisiana Slave Code of 1824, that the latter might otherwise become public wards.

THE SOCIOLOGY OF SLAVERY

The slave as outsider. The slave generally was an outsider. He ordinarily was of a different race, ethnicity, nationality, and religion from his owner. The general rule, as enunciated by the specialist on classical slavery Moses I. Finley, was that "no society could withstand the tension inherent in enslaving its own members." In most cases, the slave was an outsider because he was enslaved against his will in one society and then taken by force to another.

As with nearly all rules, there were exceptions, however. Korea, for reasons that are not understood, was one. India was another exception, because of ritual requirements that the social origins of intimate associates be known; there slaves were ritually distanced from their owners. Muscovite Russia, which had outsider slaves as well, was yet another exception, perhaps because the boundaries between insiders and outsiders were blurred. A number of scholars have pointed out that, although the status of the slaves was uniformly lower than that of comparable free people in every society, the material and sometimes other conditions of slaves were frequently better than those of free people; thus it is not surprising that free people occasionally volunteered to be slaves. What is somewhat more surprising is that so few societies found that form of social welfare to be acceptable; most took measures to prohibit or inhibit it. Solon in 594 BC, for example, forbade enslavement for debt in Athens, and the Lex Poetelia Papiria did the same for Rome, *c.* 326 BC. Muscovy in 1597 prevented self-sale into slavery from becoming hereditary by mandating manumission of such slaves on their owners' deaths.

Regardless of the slave's origin, he was nearly always a marginal person in the society in which he was enslaved. In Africa slaves were despised, and their low status, which was passed on to freedmen, persists to the present time. In most societies most slaves were at the very bottom of society.

Attitudes toward slavery: the matter of race. Slaves in most societies were despised. This is best seen in the homology for slaves. The favourite homology was the woman or wife, then the minor child or an animal. Other terms for slaves were the apprentice, the pauper, the harlot, the felon, the actor, and the complex image of the Southern "Sambo" or Caribbean "Quashee." Throughout history slaves have often been considered to be stupid, uneducable, childlike, lazy, untruthful, untrustworthy, prone to drunkenness, idle, boorish, lascivious, licentious, and cowardly. In China slaves were considered to be "mean" and "base"; in India they were fed table scraps.

Christianity's tolerance of slavery

The attitudes of the world's great religions toward slavery are of special interest. The Judeo-Christian-Islāmic tradition has been the most tolerant of slavery. Judaic and Islāmic canonical texts refer frequently to slavery and treat it as a natural condition that might befall anyone. But they view it as a condition that should be gotten over quickly. Islāmic practice was based on the assumption that the outsider rapidly became an insider and consequently had to be manumitted after six years. New Testament Christianity, on the other hand, had no prescriptions that slaves be manumitted. Canon law sanctioned slavery. This was attributable at least partially to Christianity's primary focus on spiritual values and salvation after death rather than on temporal conditions and the present life. Under such a regime it mattered little whether someone was a slave or a free person while living on earth.

A major issue in the topic of attitudes toward slavery is that of race. Although slaves were usually outsiders and often despised, there nevertheless were different kinds of outsiders and different degrees of contempt. Studies have shown that race made a difference. In Rome, where most owners and slaves were white, manumission was frequent. In Africa, where most owners and slaves were black, lineage incorporation was the primary purpose of slavery, and in most societies slaves were allowed to participate in many aspects of social life. In the American South, however, where the owners were of northern European stock and the slaves of African stock, the degree of social isolation of and contempt for slaves was extraordinary. Southern slaves were forbidden to engage in occupations that might demonstrate their capacities, intermarriage almost never occurred, and manumission was almost unheard of as the reigning publicists proclaimed ever more loudly that blacks lacked any capacity to maintain themselves as free individuals.

Slave occupations. Throughout history the range of occupations held by slaves has been nearly as broad as that held by free persons, but it varied greatly from society to society. The actual range did not depend upon whether the slave lived in a slave-owning or a slave society, although the greatest restrictions appeared in the latter.

To start at the top, the highest position slaves ever attained was that of slave minister, or *ministerialis. Ministeriales* existed in the Byzantine Empire, Merovingian France, 11th-century Germany during the Salian dynasty, medieval Muscovy, and throughout the Ottoman Empire. A few slaves even rose to be monarchs, such as the slaves who became sultans and founded dynasties in Islām.

At a level lower than that of slave ministers were other slaves, such as those in the Roman Empire, the Central Asian Samanid domains, Ch'ing China, and elsewhere, who worked in government offices and administered provinces. Some of those slaves were government property, whereas others belonged to private individuals who employed them for government work.

On a level similar to that of slaves working in government were the so-called temple slaves. They were employed by religious institutions in Babylonia, Rome, and elsewhere. Unless they were ultimately destined for sacrifice to the gods, temple slaves usually enjoyed a much easier life than other slaves. They served in occupations ranging from priestess to janitor.

The Ottoman janissary

Slaves fought as soldiers and usually were considered of high status. In some societies military slaves belonged to private individuals, in others to the government. In 16th-century Muscovy, for example, cavalrymen purchased slaves who fought alongside them on horseback; in the later 17th century Muscovite slaves were relegated to guarding the baggage train. A special type of slave soldier was the Ottoman janissary. The Islāmic Ottoman Turks confiscated Christian children (called "the tribute children"), took them to Istanbul, and raised them to be professional soldiers, or janissaries. Some janissaries served as members of the palace guard and became involved in the succession struggles of the Ottoman Empire. The Egyptian Mamlūks were also professional soldiers of slave origin who rose to run the entire country. The African Hausa of Zaria and most Sudanic regimes included slaves in all ranks of the soldiery and command. The canoe crews of the West African coast were usually slaves. The British even had detachments of slave soldiers in the Caribbean.

Societies that explicitly refused to employ slaves in combat, such as Athens in its fleet, Rome in its infantry legions, or the American South in the Civil War, were rare. They took such action because fighting was done by freemen, and it was feared that it would be necessary to free the slaves if they could fight. In fact, all of those slave societies occasionally resorted to using slave soldiers when their military situations became desperate.

Slave estate managers

In many societies slaves were employed as estate managers or bailiffs. This was especially likely to be the case when it was deemed unfitting for freemen to take or give orders involving other freemen. Where such cultural taboos existed, managers were almost always either real outsiders (imported foreigners) or fictive outsiders (slaves). In Muscovy estate managers were a special category of slave, and they were the first whose registration with the central authorities was required.

Still other high-status slaves worked as merchants. Before the invention of the corporation, using slaves was one way to expand the family firm. The practice seems to have begun in Babylonia and was perpetuated in Rome, Spain, the Islāmic world, China, and Africa. Slaves were entrusted with large sums of money and were given charge of long-distance caravans. A few slaves in Muscovy were similarly employed in the Siberian fur trade. Other societies, particularly in the American South, forbade slaves to engage in commerce out of fear that they would sell stolen goods.

In nearly all societies possessing slaves, some slaves were found in what might be termed urban occupations ranging from petty shopkeepers to craftsmen. In the Tredegar Iron Works of Richmond, Va., much of the labour force consisted of slaves. In the American South, ancient Rome, Muscovy, and many other societies, slaves worked as carpenters, tailors, and masons. In Bursa, Tur., some of the finest weaving ever done was by slave craftsmen, who often contracted to fulfill a certain amount of work in exchange for emancipation. The stereotype that slaves were careless and could only be trusted to do the crudest forms of manual labour was disproved countless times in societies that had different expectations and proper incentives.

Only a small portion of slaves throughout history were fortunate enough to be employed in elite or prestige occupations. Most were assigned to strictly physical labour, sometimes the most degrading a society had to offer.

Galley slaves

Among the worst forms of slave employment were prostitution and occupations demanding hard physical labour. Mining, often conducted in dangerous conditions causing high death rates, seems to have been the worst. The silver mines at Laurium employed as many as 30,000 slaves, who contributed to the prosperity on which Athenian democracy was based. Slaves were also used in gold mining in Africa and in gold and silver mining in Latin America. Gold and coal mining employed (and killed) millions of state slaves of the Gulag in the Soviet Union between the 1920s and 1956. Slaves have been used on great construction projects such as military fortifications, roads, irrigation projects, and temples from Babylonian to Soviet times. Timber felling for lumber and firewood was another form of hard slave labour, as in the Gulag. Yet another form of brutal slave labour was rowing in the galleys, particularly those that belonged to the Ottoman Empire and sailed the Mediterranean. Tens of thousands of Slavs, victims of Crimean Tatar slave raids, first suffered a hellish existence in the Crimea itself and then ended their days rowing on Ottoman triremes.

Large numbers of slaves were employed in agriculture. As a general rule, slaves were considered suitable for working some crops but not others. Slaves rarely were employed in growing grains such as rye, oats, wheat, millet, and barley, although at one time or another slaves sowed and especially harvested all of these crops. Most favoured by slave owners were commercial crops such as olives, grapes, sugar, cotton, tobacco, coffee, and certain forms of rice that demanded intense labour to plant, considerable tending throughout the growing season, and significant labour for harvesting. The presence or absence of such crops and their relative profitability were among the major determinants of whether or not a slave-owning society became a slave society. In the Roman Empire employment in olive groves and vineyards occupied many slaves. Sugar cultivation made 9th-century Iraq into a slave society. Rice, coconut, coffee, clove, kola nut, peanut, and sesame cultivation were central occupations in some African societies.

The great discovery in Brazil in the second half of the 16th century was the gang labour system, which was so cost-effective that it made Brazilian sugar cheaper in Europe than the sugar produced in the islands off Africa. A plantation using gang labour could produce, on average, 39 percent more output from comparable inputs than could free farms or farms employing non-gang slave labour. The secret of success was that slaves could be driven, whereas free labour could not; this led to the creation of very profitable gangs of slaves supervised by white overseers and black drivers. Tobacco and coffee cultivation also used gang labour, but cultivation of these crops was less physically demanding than that of sugar and cotton and led to much lower mortality rates than did sugar and rice.

Domestic service and concubinage

Throughout history domestic service was probably the major slave occupation. Drawing water, hewing wood, cleaning, cooking, waiting on table, taking out the garbage, shopping, child-tending, and similar domestic occupations were the major functions of slaves in all slave-owning societies. In a major productive slave system, the Roman Empire at the time of Augustus and later, the richest 5 percent of Italy's population owned 1,000,000 house slaves (another 2,000,000 were employed elsewhere, out of a total population of about 7,500,000 people). In yet another productive slave system, the American South, large numbers of slaves also worked in their owners' houses. A related function was concubinage, unquestionably one of the major uses of female slaves since the beginning of the institution and particularly prevalent in China. Some societies prescribed that a concubine who bore her owner children was to be freed; others, ranging from the ancient Middle East to the European Middle Ages, specified that the offspring of free–slave unions were to be freed. Rome and the American South were unusual in believing that all concubines and offspring should remain enslaved. Added to this in Africa was the function of lineage expansion, one of the major purposes of slavery in the sub-Saharan region.

Slave marriages

Slave demography. It is sometimes alleged that slavery and marriage were totally incompatible, for recognition of the husband–wife bond would have limited intolerably the slave owner's authority and his right to dispose of his property. Historically, however, such a view is incorrect. Limitations on the right to dispose of property have been frequent throughout history, and slaves were no exception. Thus, slave marriages were recognized in a number of slave-owning societies, including Carthage, Hellenistic Greece, late Byzantium, most of the Roman Catholic medieval world, Ch'ing China, Hindu India, Thailand, the Tlingit and Kwakiutl, and Oregon coast tribes. Ḥanbalī Muslims stated that a slave could insist that his master provide him with a spouse, and Ming Chinese masters were obliged to choose mates for their female slaves when the latter were in their teens and for males around the age of 20. In Russia marriage between a free person and a slave was recognized legally, but according to one of the oldest Russian laws the free person became enslaved by marrying a slave. In Muscovy if a married slave fled, remarried, and was subsequently apprehended, he was to be rejoined to the first spouse.

In the majority of slave societies (the Danish Virgin Islands excepted), on the other hand, slave marriages were not recognized in law and were not something that slave owners had to think about legally when disposing of slaves. For example, the Louisiana Code of 1824 explicitly stated that a slave had no right to be married. Nevertheless, even in these societies, including Rome, the American South, and West Indian Barbuda, slaves formed what they considered marriages and had children. Southern slave owners often recognized such marriages (even across estate boundaries) and their offspring because to have done otherwise would have interfered with production. In Brazil slave marriages were recognized by the Roman Catholic Church and recognized by law in 1869, but in 1875 only one-sixth of the slaves of marriageable age were recorded as married or widowed.

Slave demography was frequently determined by the occupational employment of the slaves. Consequently, sexual imbalance was not at all unusual. In 9th-century France on the Abbey of Saint Germain des Prés' territory there were nearly three male slaves for every female, presumably because of the demand for agricultural labourers. In late medieval Europe, on the other hand, there was a great demand for female slaves as domestics and concubines. The same was true in China, where by the end of the Ch'ing era the institution of slavery had become primarily a female one. In early modern Russia there were two male slaves for every female because of a market demand for cavalrymen, military body servants, and domestics who could perform heavy labour. Concubinage, moreover, was illegal, and those who sold themselves into slavery practiced female infanticide before selling themselves. In many parts of Africa the demand was primarily for women and children for the purpose of incorporation into and expansion of lineages. Adult males were often killed unless they could be exported abroad. Such export conveniently fit into the circum-Caribbean demand for productive slaves to work in sugar, tobacco, and cotton production. Consequently, twice as many males as females and relatively few children under age 10 were shipped to the New World.

One of the notions about slavery has been that slaves rarely reproduced themselves in bondage. Given the skewed demographic profile of many slave societies, it is not surprising that they failed to do so. The slaves of the Athenian Laurium silver mines or the Cuban sugar plantations, for example, lived in largely male societies. In Islāmic slave-owning societies, castration and infibulation curtailed slave reproduction.

Slave population growth

The major exception to the rule was North America, where slaves began to procreate in significant numbers in the mid-18th century. This fact helped the slave owners survive the cutting off of imports in 1808. Between the censuses of 1790 and 1860 the slave population of the South expanded enormously—from 657,327 to 3,838,765—one of the fastest rates of population growth ever recorded prior to the advent of modern medicine. Paradoxically, although the Southern slave regime was one of the most dehumanizing ever recorded, it was one of the most favourable on record demographically, because the nutritional and general living environments were highly conducive to explosive population growth. Without significant imports the Southern slave population increased fourfold between the early 1800s and 1860.

The ages of slave populations also were determined partially by productive requirements. As mentioned above, in Africa children were preferred for incorporation into lineages, whereas in much of the circum-Caribbean world adults were demanded for production. As a consequence, the age pyramids of both societies were skewed; in Africa children predominated, in much of the New World people over age 15. In Muscovy, to take another example, the age structure was skewed toward young adults, for it was primarily young adult males (aged 15–25) who sold themselves into slavery.

Slave protest. Throughout history human beings have objected to being enslaved and have responded in myriad ways ranging from individual shirking, alcoholism, flight, and suicide to arson, murdering owners, and mass rebellion. Perhaps the most common individual response to enslavement was sluggishness, passivity, and indifference. A nearly universal stereotype of the slave was of a lying, lazy, dull brute who had to be kicked or whipped. There probably were three mutually reinforcing factors at work: an unconscious response to overcontrol and absence of freedom, a conscious effort to sabotage the master's desires, and a conditioned response to the expectation of stereotypical behaviour. Some owners tried to overcome such behaviour by a system of incentives or by strict regimentation, such as the gang system, but historically they were in a minority. Less frequent was suicide. A number of slaves are known to have jumped overboard during the Middle Passage because they feared that the transatlantic voyage was taking them to be eaten by witches or barbarians, a fate that seemed worse than drowning.

Slave flight

Flight, either individually or in groups, was one of the most visible forms of protest against enslavement. The rates of flight, which varied greatly from society to society throughout history, usually depended less on individual slave-owner conduct than on the likelihood of success. Immediate conditions, such as the brutality of an overseer or master or a temporary lapse of supervision, often precipitated slave flight, but willingness to undertake such a form of rebellion against the system was usually determined by such factors as the accessibility of refuge or the ability to blend in with the free population (some societies marked slaves to inhibit such blending). Slave flight was infrequent in societies such as the peacetime American South or in West Africa, where a refuge of freedom was very distant. In East Africa, where flight was curtailed by slave owners united in their desire to prevent it in spite of a high demand for labour, runaways joined neighbouring communities and then raided their former masters. For more than two centuries fugitive slaves in Brazil known as maroons set up independent polities, or *quilombos*, that lasted for years. Maroon communities were found in many other places in Latin America and the Caribbean as well. In Muscovy, where most of the slaves were natives or of similar origin (Poles and Swedes), where there was an open frontier, and where masters had no compunction about taking in other owners' slaves, the rate of flight was very high; and as many as a quarter to a third of the slaves ran away. In China flight by male slaves was also common. During the American Revolution, when the slave owners were occupied with fighting the British, fugitive slaves numbered in the tens of thousands.

Direct, personal attacks on slave owners often were determined by the nature of the slave regime. Where owners believed they enjoyed automatic sexual access to female slaves, both the women and their "husbands" were prone to respond by assaulting the owners or their agents. In Hausaland, killings by concubines instilled great fear in slave owners. Where slaves were driven, assault on the drivers was a not uncommon response. As a result, overseers in the Mississippi Valley feared for their lives and constantly carried arms.

Slave uprisings

The most dramatic form of slave protest was outright rebellion. Slave uprisings varied enormously in frequency, size, intensity, and duration. Perhaps the calmest of all known slave societies were those of West Africa, where the predominance of women and children caused rebellions to be very few. Slave rebellions in North America were also noticeably few and involved only a handful of participants: the New York revolt of 1712, the Stono rebellion of South Carolina (1739), the Gabriel plot in Richmond, Va. (1800), the Denmark Vesey conspiracy in Charleston, S.C. (1822), and Nat Turner's uprising in Jerusalem, Va. (1831) are the best known. Southern slave uprisings were so few and so small because of the absolute certainty that they would be brutally repressed. All the Southern uprisings were savagely repressed, and the Turner rebellion is usually given as the reason for the marked increase in the severity of the slave regime after 1831.

Other slave revolts were on a much grander scale than those of West Africa and North America. One of the most famous slave uprisings was the Gladiatorial War led by Spartacus against Rome in 73–71 BC. The Spartacus rebellion was brutally repressed (the roads leading into Rome were lined with gibbets from which rebel corpses hung). Slaves led the Khlopko and Bolotnikov uprisings in Muscovy in 1603 and 1606, respectively, a time of dynastic crisis. Another great slave rebellion was that of the Zanj (black slaves imported from Zanzibar) in Iraq and Khuzistan in the years 869–883. It was joined by fair-skinned slaves as well and was on a larger scale than the Spartacus revolt. Slave rebellion in China at the end of the 17th and the beginning of the 18th century was so extensive that owners eventually eschewed male slaves and converted the institution into a female-dominated one.

Slave rebellions occurred in every slave society in the Americas from the 16th century onward. Prominent slave revolts occurred in Jamaica in 1760, 1798, and 1831–32, in Barbados in 1816, and in British Guiana in 1823. Perhaps the most famous Caribbean rebellion, in Saint-Domingue, began in 1791 and was subsequently led to victory by the freedman Toussaint-Louverture; it produced the emancipation of its slaves while the French were preoccupied with their own revolution and ultimately led to the independent state of Haiti.

SLAVE CULTURE

The institution of slavery usually tried to deny its victims their native cultural identity. Torn out of their own cultural milieus, they were expected to abandon their heritage and to adopt at least part of their enslavers' culture. Nonetheless, studies have shown that there were aspects of slave culture that differed from the master culture. Some of these have been interpreted as a form of resistance to oppression, while other aspects were clearly survivals of a native culture in the new society. Most of what is known about this topic comes from the circum-Caribbean world, but analogous developments may have occurred wherever alien slaves were concentrated in numbers sufficient to prevent their complete absorption by the host slave-owning or slave society. Thus slave culture was probably very different on large plantations from what it was on small farms or in urban households, where slave culture (and especially Creole slave culture) could hardly have avoided being very similar to the master culture. Slave cultures grew up within the perimeters of the masters' monopoly of power but separate from the masters' institutions.

Religion and slave culture

Religion, which performed the multiple function of explanation, prediction, control, and communion, seems to have been a particularly fruitful area for the creation of slave culture. Africans perceived all misfortunes, including enslavement, as the result of sorcery, and their religious practices and beliefs, which were often millennial, were formulated as a way of coping with it. Myalism was the first religious movement to appeal to all ethnic groups in Jamaica, Voodoo in Haiti was the product of African culture slightly refashioned on that island, and syncretic Afro-Christian religions and rituals appeared nearly everywhere throughout the New World. Slave religions usually had a supreme being and a host of lesser spirits brought from Africa, borrowed from the Amerindians, and created in response to local conditions. There were no firm boundaries between the secular and the sacred, which infused all things and activities. At least initially African slaves universally believed that posthumously they would return to their lands and rejoin their friends.

African elements in New World slave cultures

Black slaves preserved some of their culture in the New World. African medicine was practiced in America by slaves. The poisoning of masters and other hated individuals was a particularly African method of coping with evil. Throughout the circum-Caribbean world slaves and free blacks had electoral procedures, adapted from West African customs, to choose governors, sheriffs, and judges to maintain order among themselves. Objects of material culture, such as rugs, mats, baskets, thatched roofs, and walking canes, were modeled on African examples. Nevertheless, relatively few African social practices or plastic arts survived in the New World. On the other hand, Afro-American music and dance are known to have many African roots, and they differed dramatically from the practices of the European master culture; the use of drum and banjo were especially significant. Songs and spirituals borrowed their strong call-and-response patterns from the West African style. Furthermore, slaves created tales to amuse themselves, and the African element is most evident in animal tales; the tar-baby story is among the best known of the genre. Afro-American stories and songs often featured the devil, who was a demon and a trickster, terrifying, a friend in need, and a source of mirth.

Slave culture also developed beliefs and customs that were at odds with those of the master culture. One such belief was that what the masters called theft was something else; thus stealing from the master was not theft at all but merely a process of channeling his property from one use to another, as in taking his corn and feeding it to his pigs. Polygamous domestic arrangements were a further aspect of slave culture brought from Africa. Yet another aspect of slave culture, especially prevalent in the Caribbean, involved the market. Slaves there were often required to provide their own food, which they raised on provision grounds. If they had any surplus, they were permitted by their owners to sell it in the market. As a result slaves developed an autonomy and an individualism that contrasted starkly with the rigid control of the work gang system and the putative stifling control of slave law.

BIBLIOGRAPHY

General works: A bibliography of more than 6,000 references on slavery has been compiled in JOSEPH C. MILLER, *Slavery: A Comparative Teaching Bibliography* (1977), and continued in the issues of the journal *Slavery & Abolition* (three times a year) in several supplements, most of the references from which have been consolidated in the same author's *Slavery: A Worldwide Bibliography, 1900–1982* (1985). Theoretical works

on slavery include H.J. NIEBOER, *Slavery as an Industrial System: Ethnological Researches,* 2nd rev. ed. (1910, reprinted 1971); and ORLANDO PATTERSON, *Slavery and Social Death: A Comparative Study* (1982). Several cultures are presented in DAVID BRION DAVIS, *The Problem of Slavery in Western Culture* (1966, reissued 1971), *The Problem of Slavery in the Age of Revolution, 1770–1823* (1975), and *Slavery and Human Progress* (1984). Articles of interest are EVSEY D. DOMAR, "The Causes of Slavery or Serfdom: A Hypothesis," *The Journal of Economic History,* 30(1):18–32 (March 1970); MOSES I. FINLEY, "Slavery," in the *International Encyclopedia of the Social Sciences,* ed. by DAVID L. SILLS, vol. 14, pp. 307–313 (1968); and BERNARD J. SIEGEL, "Some Methodological Considerations for a Comparative Study of Slavery," *American Anthropologist,* 47:357–392 (1945).

Slavery in early history: On slavery in the ancient Middle East, see MUHAMMAD A. DANDAMAEV, *Slavery in Babylonia: From Nabopolassar to Alexander the Great (626–331 B.C.),* rev. ed. (1984; originally published in Russian, 1974); and ISAAC MENDELSOHN, *Slavery in the Ancient Near East: A Comparative Study of Slavery in Babylonia, Assyria, Syria, and Palestine, from the Middle of the Third Millennium to the End of the First Millennium* (1949, reprinted 1978).

Central works on slavery in classical antiquity are MOSES I. FINLEY, "Was Greek Civilization Based on Slave Labor?" in his *Slavery in Classical Antiquity: Views and Controversies* (1960); the same author's *Ancient Economy,* 2nd ed. (1985); and *Ancient Slavery and Modern Ideology* (1980). See also MOSES I. FINLEY (ed.), *Classical Slavery* (1987). Other works of major interest include KEITH R. BRADLEY, *Slaves and Masters in the Roman Empire: A Study in Social Control* (1984, reissued 1987); WILLIAM W. BUCKLAND, *The Roman Law of Slavery: The Condition of the Slave in Private Law from Augustus to Justinian* (1908, reprinted 1970); YVON GARLAN, *Slavery in Ancient Greece* (1988; originally published in French, 1982); KEITH HOPKINS, *Conquerors and Slaves* (1978); RAMSAY MacMULLEN, "Late Roman Slavery," *Historia: Zeitschrift für alte Geschichte,* 36(3):359–382 (1987); ALAN WATSON, *Roman Slave Law* (1987); and WILLIAM L. WESTERMANN, *The Slave System of Greek and Roman Antiquity* (1955).

Slavery in Europe: The following works study slavery in Europe, including Russia, after the fall of the Roman Empire: MARC BLOCH, *Slavery and Serfdom in the Middle Ages: Selected Essays* (1975; originally published in French, 1963); PIERRE DOCKÈS, *Medieval Slavery and Liberation* (1982; originally published in French, 1979); PETER FOOTE and DAVID M. WILSON, *The Viking Achievement: The Society and Culture of Early Medieval Scandinavia* (1970, reprinted 1984); ANNA CHATZĒNIKOLAOU-MARAVA, *Recherches sur la vie des esclaves dans le monde byzantyn* (1950); CHARLES VERLINDEN, *L'Esclavage dans l'Europe mědiěvale* (1955); AGNES MATHILDE WERGELAND, *Slavery in Germanic Society During the Middle Ages* (1916); CARL O. WILLIAMS, *Thraldom in Ancient Iceland* (1937); RICHARD HELLIE, *Slavery in Russia, 1450–1725* (1982); RICHARD HELLIE (ed. and trans.), *The Muscovite Law Code (Ulozhenie) of 1649* (1988); and, on the modern period, S. SWIANIEWICZ, *Forced Labour and Economic Development: An Enquiry into the Experience of Soviet Industrialization* (1965, reprinted 1985); and DAVID J. DALLIN and BORIS I. NICOLAEVSKY, *Forced Labor in Soviet Russia* (1947, reprinted 1974).

Slavery in Africa and Asia: JAMES L. WATSON (ed.), *Asian and African Systems of Slavery* (1980), is a collection of informative essays. Slavery in the non-African Islāmic world is explored in DAVID AYALON, *L'Esclavage du mamelouk* (1951); R. BRUNSCHVIG, "ʿAbd," in *The Encyclopedia of Islam,* vol. 1, pp. 24–40 (1960); PATRICIA CRONE, *Slaves on Horses: The Evolution of the Islamic Polity* (1980); HALIL INALCIK, "Servile Labor in the Ottoman Empire," in *The Mutual Effects of the Islamic and Judeo-Christian Worlds: The East European Pattern,* ed. by ABRAHAM ASCHER, TIBOR HALASI-KUN, and BÉLA K. KIRÁLY (1979); DANIEL PIPES, *Slave Soldiers and Islam: The Genesis of a Military System* (1981); and EHUD R. TOLEDANO, *The Ottoman Slave Trade and Its Suppression, 1840–1890* (1982).

Important works on slavery in other parts of Asia are D.R. BANAJI, *Slavery in British India* (1933); BENEDICTE HJEJLE, "Slavery and Agricultural Bondage in South India in the Nineteenth Century," *The Scandinavian Economic History Review,* 15:71–126 (1967); UTSA PATNAIK and MANJARI DINGWANEY (eds.), *Chains of Servitude: Bondage and Slavery in India* (1985); E.G. PULLEYBLANK, "The Origins and Nature of Chattel Slavery in China," *Journal of the Economic and Social History of the Orient,* 1:185–220 (April 1958); JAMES L. WATSON, "Chattel Slavery in Chinese Peasant Society: A Comparative Analysis," *Ethnology,* 15(4):361–375 (October 1976); MARINUS J. MEIJER, "Slavery at the End of the Ch'ing Dynasty," in *Essays on China's Legal Tradition,* ed. by JEROME ALAN COHEN, R. RANDLE EDWARDS, and FU-MEI CHANG CHEN (1980); C. MARTIN WILBUR, *Slavery in China During the Former Han Dynasty, 206 B.C.–A.D. 25* (1943, reprinted 1968); and RICHARD HELLIE, "Slavery Among the Early Modern Peoples on the Territory of the USSR," *Canadian-American Slavic Studies,* 17:454–465 (Winter 1983).

For the study of slavery in sub-Saharan Africa, see the following collections: PAUL E. LOVEJOY (ed.), *Africans in Bondage: Studies in Slavery and the Slave Trade* (1986), and *The Ideology of Slavery in Africa* (1981); CLAUDE MEILLASSOUX (ed.), *L'Esclavage en Afrique précoloniale* (1975); SUZANNE MIERS and IGOR KOPYTOFF (eds.), *Slavery in Africa: Historical and Anthropological Perspectives* (1977); and CLAIRE C. ROBERTSON and MARTIN A. KLEIN (eds.), *Women and Slavery in Africa* (1983). Also recommended are the following individual works: FREDERICK COOPER, *Plantation Slavery on the East Coast of Africa* (1977); PHILIP D. CURTIN, *Economic Change in Precolonial Africa: Senegambia in the Era of the Slave Trade* (1975); ALLAN G.B. FISHER and HUMPHREY J. FISHER, *Slavery and Muslim Society in Africa: The Institution in Saharan and Sudanic Africa, and the Trans-Saharan Trade* (1970); JOHN GRACE, *Domestic Slavery in West Africa, with Particular Reference to the Sierra Leone Protectorate, 1896–1927* (1975); PAUL E. LOVEJOY, "Indigenous African Slavery," *Historical Reflections,* 6(1)19–61 (1979), and his *Transformations in Slavery: A History of Slavery in Africa* (1983).

Slavery in the New World: Three significant works that deal with the entire circum-Caribbean slave environment or that compare different New World countries in both North and South America are CARL N. DEGLER, *Neither Black nor White: Slavery and Race Relations in Brazil and the United States* (1971, reprinted 1986); STANLEY L. ENGERMAN and EUGENE D. GENOVESE (ed.), *Race and Slavery in the Western Hemisphere* (1975); and VERA RUBIN and ARTHUR TUDEN (eds.), *Comparative Perspectives on Slavery in New World Plantation Societies* (1977).

On slavery in the United States, see JOHN W. BLASSINGAME, *The Slave Community: Plantation Life in the Antebellum South,* rev. ed. (1979); HELEN TUNNICLIFF CATTERALL (ed.), *Judicial Cases Concerning American Slavery and the Negro,* 5 vol. (1926–37, reprinted 1968); PAUL FINKELMAN, *An Imperfect Union: Slavery, Federalism, and Comity* (1981), and his *Slavery in the Courtroom: An Annotated Bibliography of American Cases* (1985); ROBERT WILLIAM FOGEL and STANLEY L. ENGERMAN, *Time on the Cross: The Economics of American Negro Slavery* (1974); DAVID W. GALENSON, *White Servitude in Colonial America: An Economic Analysis* (1981); EUGENE D. GENOVESE, *The Political Economy of Slavery: Studies in the Economy & Society of the Slave South* (1965), *Roll, Jordan, Roll: The World the Slaves Made* (1974), and *The World the Slaveholders Made,* rev. ed. with a new introduction (1988); CLAUDIA DALE GOLDIN, *Urban Slavery in the American South, 1820–1860* (1976); HERBERT G. GUTMAN, *The Black Family in Slavery and Freedom, 1750–1925* (1976); WINTHROP D. JORDAN, *White over Black: American Attitudes Toward the Negro, 1550–1812* (1968, reissued 1977); PETER KOLCHIN, *Unfree Labor: American Slavery and Russian Serfdom* (1987); ALLAN KULIKOFF, *Tobacco and Slaves: The Development of Southern Cultures in the Chesapeake, 1680–1800* (1986); LAWRENCE W. LEVINE, *Black Culture and Black Consciousness: Afro-American Folk Thought from Slavery to Freedom* (1977); LEON F. LITWACK, *Been in the Storm So Long: The Aftermath of Slavery* (1979); EDMUND S. MORGAN, *American Slavery, American Freedom: The Ordeal of Colonial Virginia* (1975); GERALD W. MULLIN, *Flight and Rebellion: Slave Resistance in Eighteenth-Century Virginia* (1972); JAMES OAKES, *The Ruling Race: A History of American Slaveholders* (1982); ULRICH BONNE PHILLIPS, *American Negro Slavery: A Survey of the Supply, Employment, and Control of Negro Labor as Determined by the Plantation Régime* (1918, reissued 1966); KENNETH M. STAMPP, *The Peculiar Institution: Slavery in the Ante-bellum South* (1956, reprinted 1975); ERIC WILLIAMS, *Capitalism & Slavery* (1944, reissued 1980); and PETER H. WOOD, *Black Majority: Negroes in Colonial South Carolina from 1670 Through the Stono Rebellion* (1974).

For information on slavery in the rest of the circum-Caribbean world and the transatlantic slave trade, see the following: FREDERICK P. BOWSER, *The African Slave in Colonial Peru, 1524–1650* (1974); ROGER NORMAN BUCKLEY, *Slaves in Red Coats: The British West India Regiments, 1795–1815* (1979); ROBERT CONRAD, *The Destruction of Brazilian Slavery, 1850–1888* (1972); MICHAEL CRATON, *Sinews of Empire: A Short History of British Slavery* (1974); PHILIP D. CURTIN, *The Atlantic Slave Trade: A Census* (1969); HENRY A. GEMERY and JAN S. HOGENDORN (eds.), *The Uncommon Market: Essays in the Economic History of the Atlantic Slave Trade* (1979); DAVID ELTIS, *Economic Growth and the Ending of the Transatlantic Slave Trade* (1987); DAVID W. GALENSON, *Traders, Planters, and Slaves: Market Behavior in Early English America* (1986); B.W. HIGMAN, *Slave Population and Economy in Jamaica, 1807–1834* (1976), and *Slave Populations of the British Caribbean, 1807–*

1834 (1984); HERBERT S. KLEIN, *The Middle Passage: Comparative Studies in the Atlantic Slave Trade* (1978); FRANKLIN W. KNIGHT, *Slave Society in Cuba During the Nineteenth Century* (1970, reprinted 1977); ROLANDO MELLAFE, *Negro Slavery in Latin America* (1975; originally published in Spanish, 1974); ORLANDO PATTERSON, *The Sociology of Slavery: An Analysis of the Origins, Development, and Structure of Negro Slave Society in Jamaica* (1967); RICHARD PRICE, *Maroon Societies: Rebel Slave Communities in the Americas,* 2nd ed. (1979); STUART B. SCHWARTZ, *Sugar Plantations in the Formation of Brazilian Society: Bahia, 1550–1835* (1985); RICHARD B. SHERIDAN, *Sugar and Slavery: An Economic History of the British West Indies, 1623–1775* (1974); and STANLEY J. STEIN, *Vassouras, a Brazilian Coffee County, 1850–1900: The Roles of Planter and Slave in a Plantation Society* (1957, reprinted 1985).

(R.He.)

Sleep and Dreams

Sleep is a normal, easily reversible, recurrent, and spontaneous state of decreased and less efficient responsiveness to external stimulation. The state contrasts with that of wakefulness, in which there is an enhanced potential for sensitivity and an efficient responsiveness to external stimuli. The sleep–wakefulness alternation is the most striking manifestation in higher vertebrates of the more general phenomenon of periodicity in the activity or responsivity of living tissue (see BEHAVIOUR, ANIMAL). There is no single, perfectly reliable criterion of sleep. Sleep is defined by the convergence of observations satisfying several different motor, sensory, and physiological criteria. Occasionally, one or more of these criteria may be absent during sleep or present during wakefulness, but even in such cases there usually is little difficulty in achieving agreement among observers in the discrimination between the two behavioral states.

Dreaming, a common and distinctive phenomenon of sleep, has since the dawn of human history given rise to myriad beliefs, fears, and conjectures, both imaginative and experimental, regarding its mysterious nature. While any effort toward classification must be subject to inadequacies, beliefs about dreams fall into various classifications depending upon whether dreams are held to be reflections of reality, sources of divination, curative experiences, or evidence of unconscious activity.

This article treats first the psychological and physiological characteristics of the sleeping state. Next, dreaming is examined from both historical-cultural and scientific points of view.

For coverage of related topics in the *Macropædia* and *Micropædia,* see the *Propædia,* Part Four, Division II, Section 422 and Division III, Section 433.

The article is divided into the following sections:

Sleep

THE NATURE OF SLEEP

Motor and sensory criteria used in defining sleep

Sleep usually requires the presence of flaccid or relaxed skeletal muscles and the absence of the overt, goal-directed behaviour of which the waking organism is capable. Part of the recurring fascination with sleep talking and sleepwalking stems from their apparent violation of this latter criterion. Were these phenomena continuous, rather than intermittent, during a behavioral state, it is indeed questionable whether the designation "sleep" would continue to be appropriate. The characteristic posture associated with sleep in man and in many but not all animals is that of horizontal repose. The relaxation of the skeletal muscles in this posture and its implication of a more passive role toward the environment are symptomatic of sleep.

Indicative of the decreased sensitivity of the human sleeper to his external environment are the typical closed eyelids (or the functional blindness associated with sleep while the eyes are open) and the presleep activities that include seeking surroundings characterized by reduced or monotonous levels of sensory stimulation. Three additional criteria—reversibility, recurrence, and spontaneity—distinguish the insensitivity of sleep from that of other states. Compared to that of hibernation or coma, the insensitivity of sleep is more easily reversible. Although the occurrence of sleep is not perfectly regular under all conditions, it is at least partially predictable from a knowledge of the duration of prior sleep periods and of the intervals between periods of sleep; and, although the onset of sleep may be facilitated by a variety of environmental or chemical means, sleep states are not thought of as being absolutely dependent upon such manipulations.

In experimental studies, both with subhuman vertebrates and with humans, sleep also has been defined in terms of physiological variables generally associated with recurring periods of inactivity identified behaviorally as sleep. For example, the typical presence of certain electroencephalogram (EEG) patterns (brain patterns of electrical activity as recorded in tracings) with behavioral sleep has led to the designation of such patterns as "signs" of sleep. Conversely, in the absence of such signs (as, for example, in a hypnotic trance) it is felt that true sleep is absent. Such signs as are now employed, however, are not invariably discriminating of the behavioral states of sleep and wakefulness. Advances in the technology of animal experimentation have made it possible to extend the physiological approach from externally measurable manifestations of sleep such as the EEG to the underlying neural (nerve) mechanisms presumably responsible for such manifestations. As a result, it may finally become possible to identify structures or functions that are invariably related to behavioral sleep and to trace the evolution of sleep through comparative anatomic and physiological studies of structures found to be critical in the maintenance of sleep behaviour in the higher vertebrates.

In addition to the behavioral and physiological criteria already mentioned, subjective experience (in the case of the self) and verbal reports of such experience (in the case of others) are used at the human level to define sleep. Upon being alerted, one may feel or say, "I was asleep just then," and such judgments ordinarily are accepted as evidence for identifying a pre-arousal state as sleep, but such subjective evidence can be at variance with behaviouristic classifications of sleep.

Problems in defining sleep

More generally, problems in defining sleep arise when evidence for one or more of the several criteria of sleep is lacking or when the evidence generated by available criteria is inconsistent. Do subhuman species sleep? Other mammalian species whose EEG and other physiological correlates are akin to those observed in human sleep demonstrate recurring, spontaneous, and reversible periods of inactivity and decreased critical reactivity. There is general acceptance of the designation of such states as sleep. As one descends the evolutionary scale below the birds and reptiles, however, and such criteria are successively less well satisfied, the unequivocal identification of sleep becomes more difficult. Bullfrogs (*Rana catesbeiana*), for example, seem not to fulfill sensory threshold criteria of sleep during resting states. Tree frogs (genus *Hyla*), on the other hand, show diminished sensitivity as they move from a state of behavioral activity to one of rest. Yet the EEGs of the alert rest of the bullfrog and the sleeplike rest of the tree frog are the same. There are parallel problems in defining sleep at different stages in the development of a single individual. At full-term birth in the human being, for instance, a convergence of nonsubjective criteria clearly seems to justify the identification of periods of sleep, but it is more difficult to justify the attribution of sleep to the human fetus.

Problems in defining sleep may arise from the effects of artificial manipulation. For example, the EEG patterns commonly used as signs of sleep can be induced in an otherwise waking organism by the administration of certain drugs. Sometimes, also, there is conflicting evidence: a person who is "awakened" from a spontaneously assumed state of immobility with all the EEG criteria of sleep may claim that he had been awake prior to this event. In such troublesome cases and more generally, it is becoming common to qualify attributions of sleep with the criteria upon which such attributions rest—*e.g.*, "behavioral sleep," "physiological sleep," or "self-described sleep." Such terminology accurately reflects both the multiplicity of criteria available for the identification of sleep and the possibility that these criteria may not always agree with one another.

DEVELOPMENTAL PATTERNS OF SLEEP AND WAKEFULNESS

Length of time spent in sleep

How much sleep does a person need? While the physiological bases of the need for sleep remain conjectural, rendering definitive answers to this question impossible, much evidence has been gathered on how much sleep people do in fact obtain. Perhaps the most important conclusion to be drawn from this evidence is that there is great variability among individuals in total sleep time. For adults, anything between six and nine hours of sleep as a nightly average is not unusual, and $7\frac{1}{2}$ hours probably best expresses the norm. Such norms, of course, inevitably vary with the criteria of sleep employed. The most precise and reliable figures on sleep time, including those cited here, come from studies in sleep laboratories, where EEG criteria are employed.

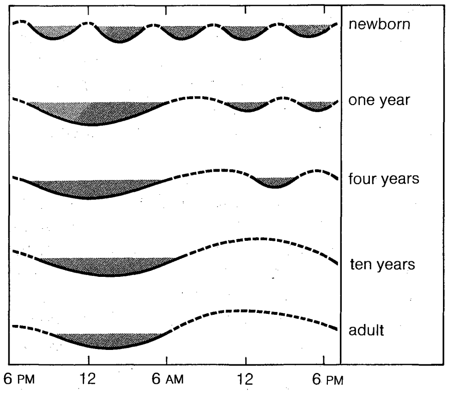
Age consistently has been associated with the varying amount, quality, and patterning of electrophysiologically defined sleep. The newborn infant may spend an average of about 16 hours of each 24-hour period in sleep, although there is wide variability among individual babies. During the first year of life, total sleep time drops sharply; by two years of age, it may range from nine to 12 hours. Decreases to approximately six hours have been observed among the elderly.

As will be elaborated below, EEG sleep studies have indicated that sleep can be considered to consist of several different stages. Developmental changes in the relative proportion of sleep time spent in these sleep stages are as striking as age-related changes in total sleep time. For example, the newborn infant may spend 50 percent of total sleep time in a stage of EEG sleep that is accompanied by intermittent bursts of rapid eye movements (REMs) indicative of a type of sleep that in some respects bears more resemblance to wakefulness than to other forms of sleep (see below *Rapid eye movement sleep*), while the comparable figure for adults is approximately 25 percent, and for the aged is less than 20 percent. There is also a decline with age of EEG stage 4 (deep slumber).

Aspects of sleep patterning

Sleep patterning consists of (1) the temporal spacing of sleep and wakefulness within a 24-hour period and (2) the ordering of different sleep stages within a given sleep period. In both senses, there are major developmental changes in the patterning of sleep. In alternations between sleep and wakefulness, there is a developmental shift from polyphasic sleep to monophasic sleep (*i.e.,* from intermittent to uninterrupted sleep). At birth, there may be five or six periods of sleep per day alternating with a like number of waking periods. With the dropping of nocturnal feedings in infancy and of morning and afternoon naps in childhood, there is an increasing tendency to the concentration of sleep in one long nocturnal period (see the Figure). The trend to monophasic sleep probably reflects some blend of the effects of maturing and of pressures from a culture geared to daytime activity and nocturnal rest. Among the elderly there may be a partial return to the polyphasic sleep pattern of infancy and early childhood, namely, more frequent daytime napping and less extensive periods of nocturnal sleep because of the loss of zeitgebers, time markers that provide cues. These include the need to arise at a set time for work or to get children off to school. Significant developmental effects also have been observed in spacing of stages within sleep. In the adult, REM sleep rarely occurs at sleep onset, while, in newborn infants, sleep-onset REM sleep is typical.

From E.J. Murray, *Sleep, Dreams and Arousal,* p. 297 (1965); Appleton-Century-Crofts



Alternations of sleep and wakefulness at specified ages. The shading indicates periods of sleep.

It would be difficult to overestimate the significance of the various age-related changes in sleep behaviour for a general theory of sleep. In the search for the functional significance of sleep or of particular stages of sleep, the shifts in sleep variables can be linked with variations in waking developmental needs, in the total capacities of the individual, and in environmental demands. It has been suggested, for instance, that the high frequency and priority in the night of REM sleep in the newborn infant may reflect a need for stimulation from within to permit orderly maturation of the central nervous system (CNS). Another interpretation of age-related changes in REM sleep stresses its possible role in processing new information, the rate of acquisition for which is assumed to be relatively high in childhood but reduced in old age. As these views illustrate, developmental changes in the electrophysiology of sleep are germane not only to sleep but also to the role of CNS development in behavioral adaptation.

PSYCHOPHYSIOLOGICAL VARIATIONS IN SLEEP

That there are different kinds of sleep has long been recognized. In everyday discourse there is talk of "good" sleep or "poor" sleep, of "light" sleep and "deep" sleep; yet, only in the second half of the 20th century have scientists paid much attention to qualitative variations within sleep. Sleep was formerly conceptualized by scientists as a unitary state of passive recuperation. Revolutionary changes have occurred in scientific thinking about sleep, the most important of which has been increased sensitivity to its heterogeneity.

This revolution may be traced back to the discovery of sleep characterized by rapid eye movement (REM sleep), first reported by the physiologists Eugene Aserinsky and Nathaniel Kleitman in 1953. REM sleep proved to have characteristics quite at variance with the prevailing model of sleep as recuperative deactivation of the central nervous system. Various central and autonomic nervous system measurements seemed to show that the REM stage of sleep is more nearly like activated wakefulness than it is like other sleep. It now has become conventional to consider REM ("paradoxical") and non-REM (NREM or "orthodox") sleep as qualitatively different. Thus, the earlier assumption that sleep is a unitary and passive state has yielded to the viewpoint that there are two different kinds of sleep, a relatively deactivated NREM phase and an activated REM phase.

Stages of NREM sleep

Non-rapid eye movement sleep. NREM sleep itself is conventionally subdivided into several different stages on the basis of EEG criteria. In the adult, stage 1 is observed at sleep onset or after momentary arousals during the night and is defined as a low-voltage mixed-frequency EEG tracing with a considerable representation of theta-wave (four to seven hertz, or cycles per second) activity. Stage 2 is a relatively low-voltage EEG tracing characterized by intermittent, short sequences of waves of 12–14 hertz ("sleep spindles") and by formations called K-complexes—biphasic wave forms that can be induced by external stimulation, as by a sound, but that also occur spontaneously during sleep. Stages 3 and 4 consist of relatively high-voltage (more than 50-microvolt) EEG tracings with a predominance of delta-wave (one to two hertz) activity; the distinction between the two stages is based on an arbitrary criterion of amount of delta-wave activity, with greater amounts classified as stage 4. Unlike the basic distinction between NREM and REM, differences among NREM sleep stages generally are regarded as quantitative rather than qualitative.

The EEG patterns of NREM sleep, particularly of stages 3 and 4 (tracings of slower frequency and higher amplitude), are those associated in other circumstances with decreased vigilance. Furthermore, after the transition from wakefulness to NREM sleep, most functions of the autonomic nervous system decrease their rate of activity and their moment-to-moment variability. Thus, NREM sleep is the kind of seemingly restful state that appears capable of supporting the recuperative functions assigned to sleep. There are, in fact, several lines of evidence suggesting such functions for NREM stage 4: (1) increases in such sleep, in both man and laboratory animals, have been observed after physical exercise; (2) the concentration of such sleep in the early portion of the sleep period (*i.e.,* immediately after wakeful states of activity) in human beings; and (3) the relatively high priority that such sleep has, among human beings, in "recovery" sleep following abnormally extended periods of wakefulness.

Rapid eye movement sleep. REM sleep is a state of diffuse bodily activation. Its EEG patterns (tracings of faster frequency and lower amplitude than in NREM stages 2–4) are at least superficially similar to those of wakefulness. Most autonomic variables exhibit relatively high rates of activity and variability during REM sleep; for example, there are higher heart and respiration rates and more short-term variability in these rates than in NREM sleep, increased blood pressure, and, in males, full or partial penile erection. In addition, REM sleep is accompanied by a relatively low rate of gross body motility, but with some periodic twitching of the muscles of the face and extremities, relatively high levels of oxygen consumption by the brain, increased cerebral blood flow, and higher brain temperature. An even more impressive demonstration of the activation of REM sleep is to be found in the firing rates of individual cerebral neurons, or nerve cells, in experimental animals: during REM sleep such rates exceed those of NREM sleep and often equal or surpass those of wakefulness. Another distinguishing feature of REM sleep, of course, is the intermittent appearance of bursts of the rapid eye movements, whence the term is derived.

The three events defining REM sleep

For both humans and animals, REM sleep now is defined by the concurrence of three events: low-voltage, mixed-frequency EEG; intermittent REMs; and suppressed tonus of the muscles of the facial region (*i.e.,* suppression of the continuous slight tension otherwise normally present). This decrease in muscle tonus and a similarly observed suppression of spinal reflexes are indicative of heightened motor inhibition during REM sleep. Animal studies have identified the locus ceruleus, in the pons, as the probable source of this inhibition. (The pons is in the brain stem directly above the medulla oblongata; the locus ceruleus borders on the brain cavity known as the fourth ventricle.) When this structure is surgically destroyed in experimental animals, they periodically engage in active, apparently goal-directed behaviour during REM sleep, although they still show the unresponsivity to external stimulation characteristic of the stage. It has been suggested that such behaviour may be the acting out of the hallucinations of a dream.

An important theoretical distinction is that between REM sleep phenomena that are continuous and those that are intermittent. Tonic (continuous) characteristics of REM sleep include the low-voltage EEG and the suppressed muscle tonus; intermittent events in REM sleep include the REMs themselves and, as observed in the cat, spikelike electrical activity in those parts of the brain concerned with vision and in other parts of the cerebral cortex. The various intermittent events of REM sleep tend to occur together, and it seems to be these moments of intermittent activation that are responsible for much of the difference between REM sleep and NREM sleep. The spiking mentioned is observed occasionally in NREM sleep, an occurrence that has been interpreted by some theorists as suggesting that REM sleep is not qualitatively unique in its capacity to support intermittent activation and that between NREM and REM sleep the differences may be less striking than the differences in eye movement and in EEG have indicated.

Sequences of NREM and REM sleep. The usual temporal progression of the two kinds of sleep in the adult human is for a period of approximately 70–90 minutes of NREM sleep (the stages being ordered 1–2–3–4–3–2) to precede the first period of REM sleep, which may last from approximately five to 15 minutes. NREM–REM cycles of roughly equivalent total duration then recur through the night, with the REM portion lengthening somewhat, and the NREM portion shrinking correspondingly, as sleep continues. Approximately 25 percent of total accumulated sleep is spent in REM sleep and 75 percent in NREM sleep. Most of the latter is EEG stage 2. The high proportion of stage 2 NREM sleep is attributable to the loss of stages 3 and 4 in the NREM portion of the NREM–REM cycles after the first two or three.

Light and deep sleep. Which of the various NREM stages is light sleep and which is deep sleep? The criteria used to establish sleep depth are the same as those used to distinguish sleep from wakefulness. In terms of motor behaviour, motility decreases (depth increases) from stages 1 through 4. By criteria of sensory responsivity, thresholds generally increase (sleep deepens) from stages 1 through 4. By most physiological criteria, NREM stages 3 and 4 are particularly deactivated (deep). Thus, gradations within NREM sleep do seem fairly consistent, with a continuum extending from the "lightest" stage 1 to the "deepest" stage 4.

Relative to NREM sleep, is REM sleep light or deep? The answer seems to be that by some criteria REM sleep

Problems of defining depth in REM sleep

is light and by others it is deep. For example, in terms of muscle tone, which is at its lowest point during sleep in REM sleep, it is deep. In terms of its increased rates of intermittent fine body movements, REM sleep would have to be considered light. Arousal thresholds during REM sleep are variable, apparently as a function of the meaningfulness of the stimulus (and of the possibility of its incorporation into an ongoing dream sequence). With a meaningful stimulus (*e.g.,* one that cannot be ignored with impunity), the capacity for responsivity can be demonstrated to be roughly equivalent to that of "light" NREM sleep (stages 1 and 2). With a stimulus having no particular significance to the sleeper, thresholds can be rather high. The discrepancy between these two conditions suggests an active shutting out of irrelevant stimuli during REM sleep. By most physiological criteria related to the autonomic and central nervous systems, REM sleep clearly is more like wakefulness than like NREM sleep, but drugs that cause arousal in wakefulness, such as amphetamine, suppress REM sleep. In terms of subjective response, recently awakened sleepers often describe REM sleep as having been "deep" and NREM sleep as having been "light." The subjectively felt depth of REM sleep may reflect the immersion of the sleeper in the vivid dream experiences of this stage.

Thus, as was true in defining sleep itself, there are difficulties in achieving unequivocal definitions of sleep depth. Several different criteria may be employed, and they are not always in agreement. REM sleep is particularly difficult to classify along any continuum of sleep depth. The current tendency is to consider it a unique state, sharing properties of both light and deep sleep. The fact that selective deprivation of REM sleep (elaborated below) results in a selective increase in such sleep on recovery nights is consistent with this view of REM sleep as unique.

Autonomic variables. Some autonomic physiological variables have a characteristic pattern relating their activity to cumulative sleep time, without respect to whether it is REM or NREM sleep. These variables are viewed by some authorities as incidental rather than essential features of the state of sleep, which is conceived in terms of the central nervous system. Such variables presumably reflect constant or slowly changing features of both kinds of sleep, such as the cumulative effects of immobility and of relaxation of skeletal muscles on metabolic processes. Body temperature, for example, drops during the early hours of sleep, reaching a low point after five or six hours, then rises toward the morning awakening.

Sleep learning, sleep-walking, sleep talking

Behavioral variables. Behaviorally, it has been shown that already established motor responses can be evoked in all stages of sleep, but it has proved much more difficult to demonstrate that new responses can be acquired during sleep. When EEG criteria of sleep are employed, it appears that "sleep learning" of verbal material takes place only to the degree that the person being tested is partially awake during the presentation of the stimuli. Another line of behavioral study is the observation of spontaneously occurring integrated behaviour patterns, such as walking and talking during sleep. In keeping with the idea of a heightened tonic (continuous) motor inhibition during REM sleep, but contrary to the idea that such behaviour is an acting out of especially vivid dream experiences or a substitute for them, sleep talking occurs primarily in NREM sleep and sleepwalking exclusively in NREM sleep. Talking in one's sleep is particularly characteristic of lighter NREM sleep (stage 2), while sleepwalking is initiated from deeper NREM sleep (stage 4). Episodes of NREM sleepwalking generally do not seem to be associated with any remembered dreams, nor is NREM sleep talking consistently associated with reported dreams of appropriate content.

For a discussion of dreaming, see below *Dreams and dreaming.*

EFFECTS OF SLEEP DEPRIVATION

One time-honoured approach to determining the function of an organ or process is to deprive an organism of the organ or process. In the case of sleep, the deprivation approach to function has been applied, both experimentally and naturalistically, to sleep as a unitary state (general sleep deprivation), and, experimentally only, to particular kinds of sleep (selective sleep deprivation). General sleep deprivation may be either total (*e.g.,* a person has had no sleep at all for a period of days) or partial (*e.g.,* over a period a person obtains only three or four hours of sleep per night). The method of general deprivation studies is enforced wakefulness. Selective deprivation has been reported for two stages of sleep: stage 4 of NREM sleep and REM sleep. Both typically occur after the appearance of other sleep stages, REM sleep after all four NREM stages and stage 4 after the lighter NREM stages. The general idea of selective deprivation studies is to allow the sleeper to have natural sleep until the point at which he enters the stage to be deprived and then to prevent the stage, either by experimental awakening or by other manipulations such as application of a mildly noxious stimulus or prior administration of a drug known to suppress it. The hope is that total sleep time will not be altered but that increased occurrence of some other stage will substitute for the loss of the one selectively eliminated.

Types of sleep deprivation

General sleep deprivation. On a three-hour sleep schedule, partial deprivation does not reproduce, in miniaturized form, the same relative distribution of sleep patterns achieved in a seven- or eight-hour sleep period. Some increase is observed in absolute amounts of REM sleep during the three-hour sleep period as compared to the first three hours of normal sleep, and there also is a significant increase in the amount of stage 4 of NREM sleep. Lighter NREM sleep (*e.g.,* stage 2) seems to have a particularly low priority under partial sleep deprivation. Although the REM sleep percentage increases somewhat under partial deprivation, the person is still far from achieving his usual quota in absolute minutes of sleep time. On uninterrupted recovery nights following the termination of the deprivation, there is more REM sleep than there was before the deprivation. This change is viewed as a compensatory "rebound" of REM sleep such that at least some of the quota is made up. Most if not all of the nightly quota of stage 4 of NREM sleep can be achieved on a three-hour nightly schedule. Because partial deprivation on a three-hour nightly regimen also tends to be selective deprivation (the person receives most of his quota of stage 4 NREM sleep but relatively little of his quota of stage REM), the behavioral effects of such deprivation may be relevant to the question of the adaptive functions served by REM sleep. One study has reported no effects from deprivation of REM sleep on the capacity for performance on a perceptual discrimination task but decreased motivation. When a schedule of partial deprivation began to interfere with the routine accumulation of stage 4 (*i.e.,* less than three hours of sleep per night), on the other hand, the capacity for performance seemed to be adversely affected.

Effects of total sleep deprivation

In view of several obvious practical considerations, many general deprivation studies have used animals rather than human beings as experimental subjects. Waking effects routinely observed in these studies have been of deteriorated physiological functioning, sometimes including actual tissue damage. Long-term sleep deprivation in the rat (six to 33 days), accomplished by enforced locomotion of both experimental and control animals but timed to coincide with any sleep of the experimental animals, has been shown to result in severe debilitation and death of the experimental but not the control animals. This supports the view that sleep serves a vital physiological function. There is some suggestion that age is related to sensitivity to the effects of deprivation, younger organisms proving more capable of withstanding the stress than mature ones.

Among human subjects, the champion non-sleeper apparently was a 17-year-old student who voluntarily undertook a 264-hour sleep deprivation experiment. Effects noted during the deprivation period included irritability, blurred vision, slurring of speech, memory lapses, and confusion concerning his identity. No long-term (*i.e.,* post-recovery) effects were observed on either his personality or his intellect. More generally, although brief hallucinations and easily controlled episodes of bizarre behaviour have been observed after five to 10 days of continuous sleep deprivation, these symptoms do not occur in most subjects and thus offer little support to the hypothesis that sleep loss induces psychosis. In any event, these symptoms rarely persist beyond the period of sleep that follows the period of deprivation. When inappropriate behaviour does persist, it generally seems to be in persons known to have a tendency toward such behaviour. Generally, upon investigation, injury to the nervous system has not been discovered in persons who have been deprived of sleep for many days. This negative result must be understood in the context of the limited duration of these studies and should not be interpreted as indicating that sleep loss is either safe or desirable. The short-term effects observed with the student mentioned are typical and are of the sort that, in the absence of the continuous monitoring his vigil received, might well have endangered his health and safety.

Other commonly observed behavioral effects during total sleep deprivation include fatigue, inability to concentrate, and visual or tactile illusions and hallucinations. These effects generally become intensified with increased loss of sleep, but they also wax and wane in a cyclic fashion in line with 24-hour fluctuations in EEG alpha-wave (eight to 12 hertz) phenomena and with body temperature, becoming most acute in the early morning hours. Changes in intellectual performance during moderate sleep loss can, to a certain extent, be compensated for by increased effort and motivation. In general, tasks that are work paced (the subject must respond at a particular instant of time not of his own choice) tend to be affected more adversely than tasks that are self-paced. Errors of omission are common with the former kind of task and are thought to be associated with "microsleep"—momentary lapses into sleep. Changes in body chemistry and in workings of the autonomic nervous system sometimes have been noted during deprivation, but it has proved difficult to establish either consistent patterning in such effects or whether they should be attributed to sleep loss per se or to the stress or other incidental features of the deprivation manipulation. In general, involuntary bodily functions seem relatively more impervious to effects of short-term deprivation than are adaptive, or voluntary, ones. The length of the first recovery sleep session for the student mentioned above, following his 264 hours of wakefulness, was slightly less than 15 hours. His sleep demonstrated increased amounts of both stage 4 NREM and stage REM sleep.

Effects of selective sleep deprivation

Selective sleep deprivation. Studies of selective sleep deprivation have confirmed the attribution of need for both stage 4 NREM and REM sleep, because an increasing number of experimental arousals is required each night to suppress both stage 4 and REM sleep on successive nights of deprivation, and because both show a clear rebound effect following deprivation. Rebound from stage 4 NREM-sleep deprivation occurs only on the night following termination of the deprivation regardless of the length of the deprivation, whereas the duration of the rebound effect following REM-sleep deprivation is related to the length of the prior deprivation. Little is known of the consequences of stage 4 deprivation.

Particular interest has attached to the selective deprivation of REM sleep, partly because of its unique and somewhat puzzling properties as an activated state of sleep and partly because of the association of this stage with vivid dreaming. REM-sleep-deprivation studies once were considered also to be "dream-deprivation" studies. This psychological view of REM sleep deprivation has become less pervasive since the experimental demonstration of the occurrence of dreaming during NREM-sleep stages, and because, contrary to the Freudian position that the dream is an essential safety valve for the release of emotional tensions, it has become evident that REM-sleep deprivation is not psychologically disruptive and may in fact be helpful in treating depression. REM-sleep-deprivation studies have focused more upon the presumed functions of the REM state than upon those of the vivid dreams that accompany it. The evidence from these studies has proved to be partially supportive of a number of different theoretical positions concerning REM sleep. Some animal studies have reported deleterious effects of REM-sleep deprivation on learning or other cognitive tasks (*i.e.,* tasks concerned with thinking, remembering, perceiving, and the like), in line with the view that cognitive processing may be one function of REM sleep. Other animal studies have shown heightened levels of sexuality and aggressiveness after a period of deprivation, suggesting a drive-regulative function for REM sleep. Other observations suggest increased sensitivity of the central nervous system to auditory stimuli and to electroconvulsive shock following deprivation, as might have been predicted from the theory that REM sleep somehow serves to maintain CNS integrity.

Effects of REM-sleep deprivation

Although there is a need for REM sleep, apparently it is not absolute. Animals have been deprived of REM sleep for as long as two months without showing behavioral or physiological evidence of injury. Several problems arise in connection with the methods of most REM-sleep-deprivation studies. Controls for factors such as stress, sleep interruption, and total sleep time are difficult to manage. Thus, it is unclear whether observed effects of REM-sleep deprivation are the result of REM-sleep loss or the result of such factors as stress and general sleep loss. It also is unclear whether it is the loss of continuous REM sleep or of the intermittent events that accompany it that is crucial in REM-sleep deprivation. Preliminary research indicates the latter, suggesting that REM-deprivation studies are more relevant to the function of separate intermittent events occurring in sleep than to the function of the continuous REM sleep.

PATHOLOGICAL ASPECTS

It is important, at the outset, to emphasize that, as dramatic and reliable as the various stages of sleep are, their functions or relations to waking performance, mood, or health are still largely unknown. Thus, association of a sleep abnormality with a certain stage of sleep (either in the sense that an abnormal event occurs during a certain stage or in the sense that an abnormal condition is associated with an increase or decrease in the proportion of total sleep time spent in that stage) is difficult to interpret when the function or necessity of that stage is uncertain. The pathology of sleep includes: (1) primary disturbances of sleep–wakefulness mechanisms, such as seem to characterize encephalitis lethargica ("sleeping sickness"), narcolepsy (irresistible brief episodes of sleep), and hypersomnia (sleep attacks of lesser urgency but greater duration than those of narcolepsy); (2) minor episodes occurring during sleep, such as bed-wetting and nightmares; (3) medical disorders such as sleep apnea whose symptoms occur during sleep; (4) sleep symptoms of the major psychiatric disorders; and (5) disorders of sleep schedule.

Primary disturbances. Epidemic lethargic encephalitis is produced by viral infections of sleep–wakefulness mechanisms in the hypothalamus, a structure at the upper end of the brain stem. The disease often passes through several stages: fever and delirium; hyposomnia (loss of sleep); and hypersomnia (excessive sleep), sometimes bordering on a coma. Inversions of 24-hour sleep–wakefulness patterns also are commonly observed, as are disturbances in eye movements.

Narcolepsy, like encephalitis, is thought to involve specific abnormal functioning of subcortical sleep regulatory centres. Some people who experience attacks of narcolepsy also have one or more of the following auxiliary symptoms: cataplexy, a sudden loss of muscle tone often precipitated by an emotional response such as laughter or startle and sometimes so dramatic as to cause the person to fall down; hypnagogic (sleep onset) and hypnopompic (awakening) visual hallucinations of a dreamlike sort; and hypnagogic or hypnopompic sleep paralysis, in which the person is unable to move voluntary muscles (except respiratory muscles) for a period ranging from several seconds to several minutes. When narcolepsy includes one or more of the accessory symptoms, some of the sleep attacks consist of periods of REM at onset of sleep. This precocious triggering of REM sleep (which occurs in adults generally only after 70–90 minutes of NREM sleep) may indicate that the accessory symptoms are dissociated aspects of REM sleep; *i.e.,* the cataplexy and the paralysis represent the active motor inhibition of REM sleep and the hallucinations represent the dream experience of REM sleep. Thus, narcolepsy involves REM sleep, and it is thought that it probably involves a failure of wakefulness mechanisms to inhibit the REM-sleep mechanisms.

Hypersomnia; hyposomnia

Hypersomnia may involve either excessive daytime sleep and drowsiness or a nocturnal sleep period of greater than normal duration but does not include sleep-onset REM periods. One reported concomitant of hypersomnia, the failure of heart rate to decrease during sleep, suggests that hypersomniac sleep may not be as restful per unit of time as is normal sleep. In its primary form, hypersomnia is probably hereditary in origin (as is also narcolepsy) and is thought to involve some disruption of the functioning of hypothalamic sleep centres. Narcolepsy and hypersomnia are not characterized by grossly abnormal EEG sleep patterns. The abnormality seems to involve a failure in "turn on" and "turn off" mechanisms regulating sleep, rather than in the sleep process itself. Narcoleptic and hypersomniac symptoms can be managed by administration of drugs. Several forms of hypersomnia are periodic rather than chronic. One rare disorder of periodically excessive sleep, the Kleine-Levin syndrome, is characterized by periods of two to four weeks of excessive sleep along with a ravenous appetite and psychotic-like behaviour during the few waking hours. The "Pickwickian syndrome" (in reference to the fat boy, Joe, in Dickens' *Pickwick Papers*), another form of periodically excessive sleep, is associated with obesity and respiratory insufficiency.

Hyposomnia (this word, meaning "too little sleep," is chosen in preference to "insomnia," or "lack of sleep," because some sleep invariably is present) is less clearly understood than the conditions already mentioned. It has been demonstrated that, by physiological criteria, self-described poor sleepers generally sleep much better than they imagine. Their sleep, however, does show signs of disturbance: frequent body movement, enhanced levels of autonomic functioning, reduced levels of stage REM, and in some the intrusion of waking rhythms (alpha waves) throughout the various sleep stages. Although hyposomnia in a particular situation is common and without pathological import, chronic hyposomnia may be related to psychological disturbance. Hyposomnia conventionally is treated by administration of drugs but often with substances that are potentially addictive and otherwise dangerous when used over long periods. Newer treatments involve behavioral programs such as the temporary restriction of sleep time and its gradual reinstatement.

Minor episodes. Among the minor episodes sometimes considered abnormal in sleep are: somniloquy (sleep talking) and somnambulism (sleepwalking), enuresis (bedwetting), bruxism (tooth grinding), snoring, and nightmares. Sleep talking seems more often to consist of inarticulate mumblings than of extended, meaningful utterances. It occurs at least occasionally for many people and at this level cannot be considered pathological. Sleepwalking is not uncommon in children, but its continuation into adulthood is suggestive of persistent immaturity of the central nervous system. Enuresis may be a secondary symptom of a variety of organic conditions or, more frequently, a primary disorder in its own right. In the latter case, it seems to involve some immaturity in neural control of bladder muscles. While mainly a disorder of early childhood, enuresis persists into adulthood for a small number of persons. Treatment generally has been directed either toward sensitizing the sleeper to bladder distention, so that he will awaken and urinate according to appropriate social norms, or toward increasing bladder capacity. Primary enuresis does not seem to be an abnormality of sleep, sleep cycles of bed-wetting and of normal children being roughly the same. Tooth grinding is not consistently associated with any particular stage of sleep, nor does it appreciably affect overall sleep patterning; it, too, seems to be an abnormality in rather than of sleep.

Sleep talking; sleepwalking; bedwetting; tooth grinding

Varieties of nightmare

A variety of frightening experiences associated with sleep, at one time or another, have been called nightmares. Because not all such phenomena have proved to be identical in their associations with sleep stages or with other variables, several distinctions need to be made among them. Incubus, the classic nightmare of adult years, consists of arousal from stage 4 NREM sleep with a sense of heaviness over the chest, with diffuse anxiety, but with little or no dream recall. Night terrors (*pavor nocturnus*) are disorders of early childhood. Delta-wave NREM sleep is suddenly interrupted with a scream; the child may sit up in apparent terror and be incoherent and inconsolable. After a period of minutes, he returns to sleep, often without ever having been fully alert or awake. Dream recall generally is absent, and the entire episode may be forgotten in the morning. Anxiety dreams most often seem associated with spontaneous arousals from REM sleep. There is remembrance of a dream whose content is in keeping with the disturbed awakening. While their persistent recurrence probably indicates waking psychological disturbance or stress caused by a difficult situation, anxiety dreams occur occasionally in many otherwise healthy persons.

Disorders accentuated during sleep. A variety of medical symptoms may be accentuated by the conditions of sleep. Attacks of angina (spasmodic, choking pain), for example, apparently can be augmented by the activation of the autonomic nervous system in REM sleep; the same is true of gastric acid secretions in persons who have duodenal ulcers. NREM sleep, on the other hand, can increase the likelihood of certain kinds of epileptic discharge.

Rhythmic snoring, which can occur throughout sleep, indicates the partial muscular relaxation of sleep, and its occasional occurrence is not abnormal. When snoring is of the loud, laboured, snorting variety, however, and is accompanied by pauses in respiration of more than 10 seconds in duration, broken by gasping sounds, the respiratory disorder called sleep apnea may be present. This disorder can occur at any age but is most common in the elderly. It results in hypoxia and sleep fragmentation, both of which contribute to excessive daytime sleepiness and cognitive deficits. Treatment approaches include behaviour change (reduction of alcohol consumption and body weight), sleep position training, mechanical appliances to keep the airway unobstructed, and surgery.

The resemblance of dream consciousness to waking psychotic experience often has been noted, and the psychotic has been considered a "waking dreamer." Thus, it has been theorized that waking psychotic symptoms may be generated by a spontaneous, or REM-sleep-deprivation-induced, shift of REM phenomena from sleep to the waking state. Symptomatically, schizophrenics have shown neither the exacerbation of psychotic symptoms under experimental REM-sleep deprivation nor the consistent or large deviations from normal EEG sleep patterning that would seem to be required by the hypothesis that sleep mechanisms play some critical role in bringing on psychotic episodes. Depressed people do sleep less and have an earlier first REM period than normal people. The first REM period, occurring 40–60 minutes after sleep onset, is often longer than normal with more eye movement activity. This suggests a disruption in the drive-regulation function, affecting such things as sexuality, appetite, or aggressiveness, all of which are reduced in such persons. REM deprivation by pharmacological agents (tricyclic antidepressants) or by REM-awakening techniques appears to reverse this sleep abnormality and to relieve the waking symptoms.

Disorders of sleep schedule. There are two prominent types of sleep-schedule disorder: phase-advanced sleep and phase-delayed sleep. In the former the sleep onset and offset occur earlier than the social norms, and in the latter sleep onset is delayed and waking is also later in the day than is desirable. These alterations in the sleep–wake cycle may occur in shift workers or following international travel across time zones. They can be treated by gradual readjustment of the timing of sleep.

DRUGS AND SLEEP

Various chemical substances long have been employed to induce or prolong sleep, but there have been few controlled, double-blind studies (neither the physician who evaluates the results nor the patient knows whether the latter has received a drug or placebo—an inert substitute) of alleged hypnotics (sleep-inducing drugs) in which sleep has been assessed by physiological measurement; and the mechanisms of sleep themselves are only now beginning to be isolated. The little research that has been done makes it clear that the manner in which a drug affects sleep can be extremely complex, with different effects sometimes attributable to different dosages of the same substance and with different effects sometimes observed for short-term and long-term administration of the same substance.

Drugs that reduce REM sleep

Many pharmacological agents tend to reduce the absolute amount and relative proportion of sleep spent in REM sleep. In this sense, REM sleep has been called a fragile state. Specifically, most effective hypnotics, particularly the barbiturates (*e.g.,* pentobarbital, secobarbital), decrease both total REM time and the proportion of sleep spent in REM sleep, with enhanced amounts of NREM sleep. Amphetamine, an analeptic (stimulant), decreases REM sleep. Many tranquilizers also slightly reduce REM sleep. There is evidence that the withdrawal symptoms of persons taken off addictive drugs of any variety (*e.g.,* barbiturates, amphetamines, narcotics) are accompanied by relatively high percentages of REM sleep. It has been suggested that the drugs in question are REM-sleep deprivers, that the elevated periods spent in REM sleep on withdrawal represent REM-sleep rebound, that the withdrawal syndrome may be functionally related to high pressure for REM sleep, and that the vivid, unpleasant dreams associated with REM-sleep rebound may be responsible for some patients' return to the use of the REM-sleep-depriving agents. Caffeine seems to have little effect on normal sleep patterning, but the effects of alcohol are variable: the short-term effect is to reduce the time spent in REM sleep, but, with continued use, there may be a REM-sleep rebound. Not all drug effects are on REM sleep; some of the more recently developed tranquilizers and hypnotics have been found to reduce stage 4 of NREM sleep.

Much interest has attached to the search for hypnotic substances that are not REM-sleep deprivers, that is, that induce or prolong sleep without altering natural sleep patterns. While some such hypnotics have been found, they most often either have adverse side effects or have not been fully evaluated. Theoretically, the most interesting substances are those few that have been found to increase REM sleep. In certain dosage ranges and under certain conditions, such an effect has been noted for reserpine, a tranquilizer, and for D-lysergic acid diethylamide (LSD), a hallucinogen. Both substances have important interactions with neurohumours (serotonin and norepinephrine—substances formed in nerve cells), and their effects may offer clues to the mechanisms underlying REM sleep.

THEORIES OF SLEEP

Two kinds of sleep theory of contemporary interest may be distinguished. One begins with the peripheral physiology of sleep and relates it to underlying neural (nervous system) or biochemical mechanisms. Such theories most often rely on experiments with animals by means of drugs or surgery. Alternatively, sleep theories may start with behavioral observations of sleep and may attempt to specify the functions of such a state of lethargy and insensitivity from an evolutionary or adaptive point of view. The question here is not so much how people sleep, or even why they sleep, but what good it does.

Mechanistic theories. Historically, mechanistic theories of sleep have focused on a succession of organs or structures in a manner reflective of the degree of access different civilizations have had to the inner workings of the human body. Thus, the relatively perceptible processes of circulation, digestion, and secretion played large roles in the theories of classical antiquity, and modern theories have been concerned with the central nervous system, particularly the brain, although various peripheral factors in the induction of sleep are not ruled out. Proposals that blood composition, metabolic changes, or internal secretions regulate sleep are necessarily incomplete to the extent that they ignore the contributions of environment and intent to the onset of sleep. It also has been noted that in two-headed human monsters one "twin" may seem asleep while the other is awake, despite their sharing a circulatory system.

Neural theories. Among neural theories of sleep, there are certain issues that each must face. Is the sleep–wakefulness alternation to be considered a property of individual neurons (nerve cells), making unnecessary the postulation of specific regulative centres, or is it to be assumed that there are some aggregations of neurons that play a dominant role in sleep induction and maintenance? The Russian physiologist Ivan Petrovich Pavlov adopted the former position, proposing that sleep is the result of irradiating inhibition among cortical and subcortical neurons (nerve cells in the outer brain layer and in the brain layers beneath the cortex). Microelectrode studies, on the other hand, have revealed high rates of discharge during sleep from many neurons in the motor and visual areas of the cortex, and thus it seems that, as compared with wakefulness, sleep must consist of a different organization of cortical activity rather than some general, overall decline.

Pavlov's theory

Another issue has been whether there is a waking centre, fluctuations in whose level of functioning are responsible for various degrees of wakefulness and sleep, or whether the induction of sleep requires another centre, actively antagonistic to the waking centre. Early speculation favoured the passive view of sleep. A *cerveau isolé* preparation, an animal in which a surgical incision high in the midbrain has separated the cerebral hemispheres from sensory input, demonstrated chronic somnolence. It has been reasoned that a similar cutting off of sensory input, functional rather than structural, must characterize natural states of sleep. Other supporting observations for the stimulus-deficiency theory of sleep included presleep rituals, such as turning out the lights, regulating stimulus input, and the facilitation of sleep induction by muscular relaxation. With the discovery of the ascending reticular activating system (ARAS, a network of nerves in the brain stem), it was found that it is not the sensory nerves themselves that maintain cortical arousal but rather the ARAS, which projects impulses diffusely to the cortex from the brain stem. Presumably sleep would result from interference with the active functioning of the ARAS. Injuries to the ARAS were, in fact, found to produce sleep. Sleep thus seemed passive, in the sense that it was the absence of something (ARAS support of sensory impulses) characteristic of wakefulness.

Theory has tended to depart from this belief and to move toward conceiving of sleep as an actively produced state. Two kinds of observation primarily have been responsible for the shift. First, earlier studies showing that sleep can be induced directly by electrical stimulation of certain areas in the hypothalamus have been confirmed and extended to other areas in the brain. Second, the discovery of REM sleep has been even more significant in leading theorists to consider the possibility of actively produced sleep. REM sleep, by its very active nature, defies description as a passive state. As is noted below, REM sleep can be eliminated in experimental animals by the surgical destruction of a group of nerve cells in the pons, the active function of which appears to be necessary for REM sleep. Thus, it is difficult to imagine that the various manifestations of REM sleep reflect merely the deactivation of wakefulness mechanisms.

One- and two-system theories

The REM–NREM-sleep dichotomy poses a third issue for the theories of sleep mechanisms, or at least for those that accept the idea of sleep as an active phenomenon. Does one hypnogenic (sleep-causing) system serve both kinds of sleep, or are there two antagonistic sleep systems, one for REM sleep and one for NREM sleep? Opinion is sharply divided. One group of theorists states that there must be two sleep systems. It is noted that NREM sleep is not affected, but that REM sleep is abolished, by injuries to the pontine tegmentum (the posterior part of the pons) and that NREM sleep is suppressed in animals whose brain stem has been severed at the midpoint of the pons, suggesting that an NREM-sleep centre behind this section no longer is capable of suppressing the effect of the ARAS. It is further observed that the neurohumour serotonin is localized in the brain-stem regions presumed to be responsible for NREM sleep; that destruction of serotonin-containing nerve cells in the brain stem may produce insomnia; that, in some species, reductions of serotonin by chemical interference with its production produces an amount of sleep loss correlated with the reduction of serotonin; that administration of a serotonin precursor (a substance from which serotonin is formed) after interference with production of serotonin produces a sleeplike state and that artificially induced increases in brain serotonin increase NREM sleep; that the neurohumour norepinephrine is localized in the brain-stem regions presumed to be responsible for REM sleep; and that substances interfering with the synthesis of norepinephrine suppress REM sleep. Other theorists have proposed that REM and NREM sleep are served by a common hypnogenic system. Chemical stimulation of certain brain structures, assumed to constitute a hypnogenic system, has been found capable of inducing both stages of sleep. It also is argued that different varieties of sleep should require different mechanisms no more than do different varieties of wakefulness (*e.g.,* alertness, relaxation).

Functional theories. Functional theories stress the recuperative and adaptive value of sleep. Sleep arises most unequivocally in animals that maintain a constant body temperature and that can be active at a wide range of environmental temperatures. In such forms, increased metabolic requirements may find partial compensation in periodic decreases in body temperature and metabolic rate (*i.e.,* during NREM sleep). Thus, the parallel evolution of temperature regulation and NREM sleep has suggested to some authorities that NREM sleep may best be viewed as a regulatory mechanism conserving energy expenditure in species whose metabolic requirements are otherwise high. As a solution to the problem of susceptibility to predation that comes with the torpor of sleep, it has been suggested that the periodic reactivation of the organism during sleep better prepares it for fight or flight, and that the possibility of enhanced processing of significant environmental stimuli during REM sleep may even reduce the need for sudden confrontation with danger. Other functional theorists agree that NREM sleep may be a state of "bodily repair," while suggesting that REM sleep is one of "brain repair" or restitution, a period, for example, of increased cerebral protein synthesis or of "reprogramming" the brain so that information achieved in wakeful functioning is most efficiently assimilated. In their specification of functions and provision of evidence for such functions, such theories are necessarily vague and incomplete. The function of stage 2 NREM sleep is still unclear, for example. Such sleep is present in only rudimentary form in subprimate species yet consumes approximately half of human sleep time. Comparative, physiological, and experimental evidence is unavailable to suggest why so much human sleep is spent in this stage. In fact, poor sleepers whose laboratory sleep records show high proportions of stage 2 and little or no REM sleep often report feeling they have not slept at all. (D.F./R.D.C.)

Dreams and dreaming

DIVERSE VIEWS ON THE NATURE OF DREAMS

Dreams as reflecting reality. Philosophers continue to argue about the differences between reality and dreams. The English philosopher Bertrand Russell (1872–1970) wrote, "It is obviously possible that what we call waking life may be only an unusual and persistent nightmare," and he further stated that "I do not believe that I am now dreaming but I cannot prove I am not." Philosophers generally try to resolve the question by saying that so-called waking experience seems vivid and coherent. As the French philosopher René Descartes (1596–1650) put it: ". . . memory can never connect our dreams one with the other or with the whole course of our lives as it unites events which happen to us while we are awake"; or, as Russell stated succinctly, "Certain uniformities are observed in waking life, while dreams seem quite erratic."

Dreams in various cultures

Members of many cultures have variously coped with this dilemma; for example, among the Eskimo of Hudson Bay and the Patani Malay people, it is believed that during sleep one's "soul" leaves the body to live in a special dreamworld. Believers often consider it dangerous to wake someone lest his "soul" be lost. On these grounds the Tajal people of Luzon, for example, severely punish for awakening a sleeping person. In other cultures, dream events are held to be identical with reality; thus, a Macusi Indian of Guyana is reported to have become enraged at the European leader of an expedition when he dreamed that the leader had made him haul a canoe up dangerous cataracts. He awoke exhausted and could not be persuaded that the dream was not real. There is a tradition in Borneo that if a man dreams that his wife is an adulteress, her father must take her back. A Zulu man is said to have broken off a friendship after dreaming that the friend meant him harm. A Paraguayan Indian, reportedly having dreamed that a missionary shot at him, attempted to kill the missionary.

In other instances, dream events may be believed to demand fulfillment. Jesuit priests in the 1700s reported that among Iroquois Indians it was obligatory to carry out dreams as soon as possible; one Indian was said to have dreamed that 10 friends dove into a hole in the ice of a lake and came up through another. When told of the dream, the friends duly enacted their roles in it, but unfortunately, only nine of them succeeded. After dreaming of something valuable, Kurdish people were traditionally expected to take it, by force if necessary. Among some natives of Kamchatka a man need only dream of a girl's favour for her to owe him her sexual favours.

Such interpretations in which the dream is given a status of reality need not imply that the two are indistinguishable. In some instances, the dream may be differentiated from reality, but dreams are accorded a superior status to the banal activities of wakefulness.

Dreams in the Bible

Dreams as a source of divination. There is an ancient belief that dreams predict the future; the Chester Beatty Papyrus is a record of Egyptian dream interpretations dating from the 12th dynasty (1991–1786 BC). In the *Iliad,* Agamemnon is visited in dream by a messenger of the god Zeus to prescribe his future actions. From India, a document called the Atharvaveda, attributed to the 5th century BC, contains a chapter on dream omens. A Babylonian dream guide was discovered in the ruins of the city of Nineveh among tablets from the library of the emperor Ashurbanipal (668–627 BC). The Old Testament is rife with prophetic dreams, those of the pharaohs and of Joseph and Jacob being particularly striking. Among pre-Islāmic peoples dream divination so heavily influenced daily life that the practice was formally forbidden by Muḥammad (*c.* 570?–632), founder of the Muslim religion.

Ancient and religious literatures express the most confidence about so-called message dreams. Characteristically, a god or some other respected figure appears to the dreamer (typically a king, a hero, or a priest) in time of crisis and states a message. Such reports are found on ancient Sumerian and Egyptian monuments; frequent examples appear in the Old and New testaments. Joseph Smith (1805–44), the founder of the Mormon religion, said that an angel had directed him to the location of buried golden tablets that described American Indians as descendants of the tribes of Israel.

Not all dream prophecies are so readily accepted. In the *Odyssey,* for example, dreams are classed as false ("passing through the Gate of Ivory") and as true ("passing the Gate of Horn"). Furthermore, prophetic meaning may be attributed to dream symbolism. In the Bible, Joseph interpreted sheaves of grain and the Moon and stars as symbols of himself and of his brethren. In general, the social status of dream interpreters varies; in cultures for which dreams loom important, their interpretation frequently is an occupation of priests, elders, or medicine men.

Perhaps the most famous dream interpretation book is that of the Greek soothsayer Artemidorus Daldianus (*c.* 2nd century AD), the *Oneirocritica* (from the Greek *oneiros,* "a dream"). Dream books remain widely available today. They continue to enjoy profitable sales everywhere among people who follow them in affairs of the heart, in gambling, and in matters of health and work.

Dreams as curative. So-called prophetic dreams in the Middle Eastern cultures of antiquity often were combined with other means of prophecy, such as animal sacrifice, and with efforts to heal the sick. In classical Greece, dreams became directly associated with healing; ailing people came to dream in oracular temples where priests and priestesses advised about the cures dreams were held to provide. Similar practices, known as dream incubation, are recorded for Babylon and Egypt. In a widespread cult, suffering petitioners came to at least 600 temples of the Greek god of medicine to perform rites or sacrifices in efforts to dream appropriately, sleeping in wait of the appearance of the god or his emissary to deliver a cure. Many stone monuments placed at the entrances of the temples survive to record dream cures.

Dreams as extensions of the waking state. Even in early human history dreams also were interpreted as reflections of waking experiences and of emotional needs. Aristotle (384–322 BC), despite his contemporaries who practiced divination and incubation, in his work *Parva Naturalia* (*On the Senses and Their Objects*) attributed dreams to sensory impressions from "external objects . . . pauses within the body . . . eddies . . . of sensory movement often remaining like they were when they first started, but often too broken into other forms by collision with obstacles." In anticipation of psychoanalyst Sigmund Freud (1856–1939), Aristotle wrote that sensory function is reduced in sleep, favouring the susceptibility of dreams to emotional subjective distortions.

In spite of Aristotle's unusually modern views and even after a devastating attack by the Roman statesman Marcus Tullius Cicero (106–43 BC) on dream divination (*De divinatione;* "On Divination"), views that dreams have supernatural attributes persisted vigorously until the 1850s and the classic work of the French physician Alfred Maury, who studied more than 3,000 reported recollections of dreams. Maury concluded that dreams arose from external stimuli, instantaneously accompanying such impressions as they acted upon the sleeping person. He wrote that part of his bed once fell on the back of his neck and woke him, leaving the memory of dreaming that he had been brought before a French revolutionary tribunal, questioned, condemned, led to the scaffold, and bound by the executioner, and that the guillotine blade had fallen.

Creative dreaming

The English poet Samuel Taylor Coleridge reported that he had written "Kubla Khan" as the result of creative thinking in a dream. Having fallen asleep while reading about that Mongol conqueror, he woke to write down a fully developed poem he seemed to have composed while dreaming. Novelist Robert Louis Stevenson said that much of his writing was developed by "little people" in his dreams, and specifically cited the story of Dr. Jekyll and Mr. Hyde in this context. The German chemist F.A. Kekulé von Stradonitz attributed his interpretation of the ring structure of the benzene molecule to his dream of a snake with its tail in its mouth. Otto Loewi, the German physiologist, attributed to a dream inspiration for an experiment with a frog's nerve that helped him win the Nobel Prize. In all of these cases the dreamers reported having thought about the same topics over considerable periods while they were awake.

Psychoanalytic interpretations. Among Freud's earliest writings was *The Interpretation of Dreams* (1899). His insistence that dreams are "the royal road to the unconscious" continued from it down to his last published statement on dreams, printed about a year before he died. Freud held that dreams reflect waking experience; he offered a theoretical explanation for their bizarre nature, invented a system for their interpretation, and elaborated on their curative potential.

Dreamwork

Freud theorized that thinking during sleep tends to be primitive and regressive and that the effects of forgetting (repression) are reduced. Repressed wishes, particularly those associated with sex and hostility, were said to be released in dreams when the inhibitory demands of wakefulness diminished. The content of the dream was said to derive from such stimuli as urinary pressure in the bladder, traces of experiences from the previous day (day residues), and associated infantile memories. The specific dream details were called their manifest content; the presumably repressed wishes being expressed were called the latent content. Freud suggested that the dreamer kept himself from waking and avoided unpleasant awareness of repressed wishes by disguising them as bizarre manifest content in an effort called dreamwork. He held that impulses one fails to satisfy when awake are expressed in dreams as sensory images and scenes. In dreaming, Freud believed:

> All of the linguistic instruments . . . of subtle thought are dropped . . . and abstract terms are taken back to the concrete . . . . The copious employment of symbols . . . for representing certain objects and processes is in harmony (with) the regression of the mental apparatus and the demands of censorship.

Freud theorized that one aspect of manifest content could come to represent a number of latent elements (and vice versa) through a process called condensation. Further displacement of emotional attitudes toward one object or person theoretically could be displaced in dreaming to another object or person or not appear in the dream at all. Freud further observed a process called secondary elaboration, which occurs when people wake and try to remember dreams. They may recall inaccurately in a process of elaboration and rationalization and provide "the dream, a smooth facade, (or by omission) display rents and cracks." This waking activity he called secondary revision.

In seeking the latent meaning of a dream, Freud advised the individual to associate freely about it. From listening to the associations, the analyst was supposed to determine what the dream represented, in part through an understanding of the personal needs of the dreamer.

Carl Jung (1875–1961) disagreed with Freud's view of dreams as being complementary to waking mental life with respect to specific instinctual impulses. Jung felt that dreams are instead compensatory, that they balance whatever elements of character are underrepresented in the way people are living their lives. Dreaming, to Jung, represents a continuous 24-hour flow of mental activity that surfaces in sleep when conditions are right, but which affects waking life when a person's behaviour denies important elements of his true personality.

Thus, dreams are constructed not to conceal or disguise forbidden wishes but to bring the under-attended areas to attention. This function is carried out unconsciously in sleep when people are living well-balanced lives. If this is not the case there may be first bad moods, then symptoms in waking. Then and only then do dreams need to be interpreted. This is best done not with a single dream and multiple free associations but with a series of dreams so that the repetitive elements become apparent.

EFFORTS TO STUDY DREAMING

Dream reports. Though each person seems to know his own private dreams, the manner in which people dream obviously defies direct observation. It has been said that each dream "is a personal document, a letter to oneself" and must be inferred from the observable behaviour of people. Furthermore, observational methods and purposes clearly affect conclusions to be drawn about the inferred dreams. Reports of dreams collected from people after morning awakenings at home tend to exhibit more content of an overt sexual and emotional nature than do those from laboratory subjects. Such experiences as dreaming in colour seldom are spontaneously mentioned but often emerge under careful questioning. Reports of morning dreams are typically richer and more complex than those collected early at night. Immediate recall differs from what is reported after longer periods of wakefulness; psychoanalysts seem to elicit more recollections of overt sexual dreams than do laboratory investigators. In spite of these complications, there have been substantial efforts to describe the general characteristics of what people say they have dreamed.

Length of dream reports

The reported length of dreams varies widely between and within individuals (and by inference, so does the length of the presumed dreams themselves). Spontaneously reported dreams among laboratory subjects are typically short; about 90 percent of these reports are less than 150 words long, although some may exceed 1,000. With additional probing, about a third of such reports are longer than 300 words.

Some investigators have been surprised by repeated findings that suggest dreams may be less fantastic or bizarre than generally supposed. In the language of modern art, one investigator stated that visual dreams are typically faithful to reality (representational) with little, if any, abstractionist or surrealistic dreaming. Any variations from the representational were characterized as impressionistic. Except for those that are very short, dreams are reported to take place in ordinary physical settings, about half of them seeming quite familiar to the dreamer; only rarely is the setting said to be exotic or peculiar.

Apparently dreams are quite egocentric, the dreamer perceiving himself as a participant, though the presence of others is typically recalled. Seldom does the person remember an empty, unpopulated dreamworld, and individuals seem to dream roughly two-thirds of the time about people they know. Usually they are close acquaintances; family members are mentioned in about 20 percent of dream reports. Recollections of notables or weird representations of people are generally rare.

The typical report is of visual imagery; indeed, in its absence, the person may say only that he had been thinking rather than "dreaming" while asleep. Rare statements about dreams dominated by auditory experience commonly are made with claims of actually having been awake. It is unusual, however, to hear of dreams without some auditory characteristics. One typically is told of bland dreams; when there are emotional overtones, they tend to be unpleasant about two-thirds of the time. Fear and anxiety are most commonly mentioned, followed by anger; pleasant feelings are most often those of friendliness. Reports of overtly erotic dreams, particularly among those gathered in laboratories, are infrequent.

Despite their generally representational nature, dreams seem somehow odd or strange. Perhaps this is related to discontinuities in time and purpose. One suddenly may dream of himself in a familiar auditorium viewing a fencing match rather than hearing a lecture and abruptly in the "next scene" walking beside a swimming pool. Or a person may have the experience of lying in a hallway listening to two people standing by an elevator; he may be looking at a bleeding hand and walk across an empty room to a liquor cabinet to find a roll of adhesive tape. These sudden transitions contribute to the dreamer's feeling of strangeness, and this is enhanced by his waking statements that the bulk of his dreams cannot be clearly recalled, giving them a dim, mysterious quality.

Rapid eye movement

Physiological dream research. A new era of dream research began in 1953 with the discovery that rapid eye movements during sleep seem often to signal that a person is dreaming. In that year it was observed that, about an hour or so after falling asleep, laboratory subjects are apt to experience a burst of rapid eye movement (REM) under their closed lids, accompanied by a change in brain waves detected by an electroencephalograph as an electrical pattern resembling that of an alert waking person (see above *Rapid eye movement sleep*). When subjects were awakened during REM, they reported vivid dreams 20 out of 27 times; when roused during non-REM sleep, they recalled dreams in only four of 23 instances. Subsequent systematic study confirmed this relationship between REM, activated brain waves (EEG), and dream recall. Several thousand experimental studies utilizing these observable indexes of dreaming have since been conducted.

A major finding is that the usual report of a vivid, visual dream is primarily associated with REM and activated EEG. On being aroused while exhibiting these signs, people recall dreams with visual imagery about 80 percent of the time. When awakened in the absence of them, however, people still report some kind of dream activity, though only about 30 to 50 percent of the time; in such cases they are apt to remember their sleep experiences as being relatively "thoughtlike" and realistic and as resembling the experiences of wakefulness.

Dreaming in animals

D-state (desynchronized or dreaming) sleep has been reported for all mammals studied; it has been observed, for example, among monkeys, dogs, cats, rats, elephants, shrews, and opossums; these signs also have been reported in some birds and reptiles.

Surgical destruction of selected brain structures among laboratory animals has clearly demonstrated that the D-state depends on an area within the brain stem known as the pontine tegmentum. Evidence indicates that D-state sleep is associated with a mechanism involving a bodily chemical called norepinephrine; other stages of sleep seem to involve another chemical (serotonin) in the brain. Among other physiological changes found intimately related to D-state sleep are increased variability in breathing and heart rate, relaxation of skeletal muscles (in lower animals), and, in humans, reduction of electrical activity in muscles near the base of the tongue and penile erections or increase in vaginal blood flow and uterine contractions.

When people are chronically deprived of the opportunity to manifest D-state activity (by awakening them whenever there is EEG evidence of dreaming), it appears increasingly difficult to prevent them from dreaming. On recovery nights (after such deprivation) when the subject can sleep without interruption, there is a substantial increase in the number of reports of dreaming. This rebound effect continues in some degree on subsequent recovery nights, depending on how badly the person has been deprived.

Dream recall

During D-states in the last 6½ to 7½ hours of sleep people are likely to wake by themselves about 40 percent of the time. This figure is about the same as that for dream recall, people saying they had a dream the previous night about 35 percent of the time (roughly once every three or four nights). Evidence concerning the amount and kind of dreaming also depends on how rapidly one is roused and on the intensity of his effort to recall. Some people recall dreams more often than the average, while others rarely report them. While these two groups of people show little difference in amount of D-state sleep, evidence suggests that non-recall reflects a general tendency on the part of the individual to repress or to deny his experiences.

The psychoanalytic literature is rich with reports indicating that what one dreams about reflects his needs and his immediate and remote past experience. Nevertheless, when someone in D-state sleep is stimulated (*e.g.*, by spoken word or by drops of water on his skin), the chances that he will say he has dreamed about the stimulus, or anything similar, are quite low. Studies in which people have watched vivid movies before falling asleep also indicate some possibilities of influencing dreams but again clearly emphasize the limitations of such influences. Highly suggestible people seem likely to dream as they are told to do while under hypnosis, but the influence of direct suggestion during ordinary wakefulness seems quite limited.

Variations within the usual range of about 18 to 30 percent of D-state sleep apparently are unrelated to differences in the amount or content of dreaming. The amount of D-state further seems generally independent of wide variations in the daily activities or personality characteristics of different people; groups of scientists, athletes, housewives, and artists, for example, cannot be reliably distinguished from one another in terms of D-state activity. Such disorders as schizophrenia and mental retardation appear to have no clearly discernible effect on the amount of time sufferers spend in REM-activated EEG sleep.

DREAMLIKE ACTIVITIES

Related states of awareness may be distinguished from the dream experiences typically reported; these include dreamlike states experienced as a person falls asleep and as he awakens, respectively called hypnagogic and hypnopompic reveries. During sleep itself there are nightmares, observable signs of sexual activity (*e.g.*, nocturnal emissions of sperm), and sleepwalking. Even people who ostensibly are awake may show evidence of such related phenomena as hallucinating, trance behaviour, and reactions to drugs.

Rapid eye movement is not characteristic of sleep onset; nevertheless, as people drift (as inferred from EEG activity) from wakefulness through drowsiness into sleep, they report dreamlike hypnagogic experiences about 90 percent of the time on being awakened. Most of these experiences (about 80 percent) are said to be visual. If dreaming is defined as at least partly hallucinatory and somewhat dramatic, then awakening from drowsiness or at the onset of sleep yields recall of experiences that may be classified as dreams for about 75 percent of the occasions. These "dreamlets" seem to differ from dream-associated REM sleep in being less emotional (neither pleasant nor unpleasant), more transient, and less elaborate. Such hypnagogic experiences apparently tend to incorporate abstract thinking and recall of recent events (day residues) and to be quite typical of falling asleep. Systematic studies remain to be made of the hypnopompic reveries commonly reported mornings before full arousal, but it seems likely that they include recollections of the night's dreams, or represent one's drifting back into transient REM sleep.

Extreme behavioral manifestations during sleep—night terrors, nightmares, sleepwalking, enuresis—all have been found generally unrelated to ordinary dreaming. Night terrors are characterized by abrupt awakening, sometimes with a scream; a sleeping child may sit up in bed, apparently terror-stricken, with wide-open eyes, and often with frozen posturing that may last several minutes. Afterward there typically is no recollection of dreamlike experience. Observed in about 2 or 3 percent of children, roughly half of the attacks of night terror occur between the ages of four and seven; about 10 percent of them are seen among youngsters as old as 12 to 14 years. Nightmares typically seem to be followed by awakening with feelings of suffocation and helplessness and expressions of fearful or threatening thoughts. Evidence of nightmares is observed for 5 to 10 percent of children, primarily about eight to 10 years of age. Studies have suggested that signs of spontaneously generated night terrors and nightmares may be related to abrupt awakening from deep sleep that experimentally appears dreamless. This suggests that the vividly reported fears well may be produced by emotional disturbances that first occur on awakening.

Nightmares

Sleepwalking, observed in about 1 percent of children, predominantly appears between ages 11 and 14. Apparently sleeping individuals rise and walk from their beds, eyes open, usually avoiding obstacles, and expressing no recollection of the episode when they wake. Studies of EEG data indicate that sleepwalking occurs only in deep sleep when dreams seem essentially absent; the behaviour remains to be reported for REM sleep. Enuresis occurs in about one-fourth of children over age four. These episodes seem not to be associated with REM as much as they do with deep sleep in the absence of D-state signs.

Nocturnal emission of sperm remains to be described in terms of any distinguishing EEG pattern; such events are quite rarely observed among sleeping laboratory subjects. Among a large sample of males who were interviewed about their sexual behaviour about 85 percent reported having experienced emissions at some time in their lives, typical frequency during the teens and 20s being about once a month. Of the females interviewed 37 percent reported erotic dreams, sometimes with orgasm, averaging about three to four times a year. Most often, however, openly sexual dreams are said not to be accompanied by orgasm in either sex. Males not infrequently could recall no dreams associated with emission, although most implicated erotic dreaming.

Dreamlike experiences induced as trances, deliriums, or hallucinatory behaviour by drugs seem attributable to lowered efficiency of the central nervous system in processing sensory stimuli from the external environment. The result seems to be that one's physiological activities begin to escape environmental constraint to the point that internalized, uncritical thinking and perceiving prevail.

Since antiquity, dreams have been viewed as a source of divination, as a form of reality, as a curative force, and as an extension or adjunct of the waking state. Psychoanalytic theorists stress the individual meaningfulness of dreams and their relation to personal hopes and fears. Contemporary research focuses on efforts to discover and describe unique, complex biochemical and neurophysiological bases of dreaming. Among the plethora of theories ranging from those that assert dreaming to be awareness of a god's voice to those that reduce the dream to physical activity in the nervous system, no single, encompassing theory seems to be available. (W.B.W./R.D.C.)

BIBLIOGRAPHY

Sleep: The standard reference on the physiology of sleep is NATHANIEL KLEITMAN, *Sleep and Wakefulness*, rev. and enl. ed. (1963). Also see ERNEST HARTMANN, *The Functions of Sleep* (1973), an excellent summary of psychological and biological research on REM and NREM sleep functions; ANTHONY KALES and JOYCE D. KALES, *Evaluation and Treatment of Insomnia* (1984), the most comprehensive work on this sleep disorder; and JERROLD S. MAXMEN, *A Good Night's Sleep: A Step-By-Step Program for Overcoming Insomnia and Other Sleep Problems* (1981), a popular book that covers a range of sleep problems.

Dreams and dreaming: Two differing classic theories of dream interpretation are found in SIGMUND FREUD, *The Interpretation of Dreams,* vol. 4 and 5 in *The Standard Edition of the Complete Psychological Works of Sigmund Freud*, edited by JAMES STRACHEY (1953, reprinted 1981; originally published in German, 8th ed., 1930), also available in other translations; and C.G. JUNG, *Dream Analysis: Notes of the Seminar Given in 1928–1930* (1984). Also see G.E. VON GRUNEBAUM and ROGER CAILLOIS (eds.), *The Dream and Human Societies* (1966), a scholarly work, with a chapter on dream research; RICHARD M. JONES, *The New Psychology of Dreaming* (1970, reissued 1978), which attempts to coordinate experimental findings on dreams with classic theories; ROSALIND DYMOND CARTWRIGHT, *Night Life: Explorations in Dreaming* (1977), which covers a series of studies approaching a laboratory-based understanding of dreams and dreaming; MONTAGUE ULLMAN and NAN ZIMMERMAN, *Working with Dreams* (1979, reissued 1985), which helps the general reader understand and work with dream material; and ERNEST HARTMANN, *The Nightmare: The Psychology and Biology of Terrifying Dreams* (1984), a comprehensive work.
(D.F./W.B.W./R.D.C.)

Adam Smith

After two centuries, Adam Smith remains a towering figure in the history of economic thought. Known primarily for a single work, *An Inquiry into the nature and causes of the Wealth of Nations* (1776), the first comprehensive system of political economy, Smith is more properly regarded as a social philosopher whose economic writings constitute only the capstone to an overarching view of political and social evolution. If his masterwork is viewed in relation to his earlier lectures on moral philosophy and government, as well as to allusions in *The Theory of Moral Sentiments* (1759) to a work he hoped to write on "the general principles of law and government, and of the different revolutions they have undergone in the different ages and periods of society," then *The Wealth of Nations* may be seen not merely as a treatise on economics but as a partial exposition of a much larger scheme of historical evolution.

Early life. Unfortunately, much more is known about Smith's thought than about his life. Though the exact date of his birth is unknown, he was baptized on June 5, 1723, in Kirkcaldy, a small (population 1,500) but thriving fishing village near Edinburgh, the son by second marriage of Adam Smith, comptroller of customs at Kirkcaldy, and Margaret Douglas, daughter of a substantial landowner. Of Smith's childhood nothing is known other than that he received his elementary schooling in Kirkcaldy and that at the age of four years he was said to have been carried off by gypsies. Pursuit was mounted, and young Adam was abandoned by his captors. "He would have made, I fear, a poor gypsy," commented his principal biographer.

At the age of 14, in 1737, Smith entered the university of Glasgow, already remarkable as a centre of what was to become known as the Scottish Enlightenment. There, he was deeply influenced by Francis Hutcheson, a famous professor of moral philosophy from whose economic and philosophical views he was later to diverge but whose magnetic character seems to have been a main shaping force in Smith's development. Graduating in 1740, Smith won a scholarship (the Snell Exhibition) and traveled on horseback to Oxford, where he stayed at Balliol College.

Influence of Hutcheson

Smith, paste medallion by James Tassie, 1787. In the Scottish National Portrait Gallery, Edinburgh.

By courtesy of the Scottish National Portrait Gallery, Edinburgh

Compared to the stimulating atmosphere of Glasgow, Oxford was an educational desert. His years there were spent largely in self-education, from which Smith obtained a firm grasp of both classical and contemporary philosophy.

Returning to his home after an absence of six years, Smith cast about for suitable employment. The connections of his mother's family, together with the support of the jurist and philosopher Lord Henry Kames, resulted in an opportunity to give a series of public lectures in Edinburgh—a form of education then much in vogue in the prevailing spirit of "improvement."

The lectures, which ranged over a wide variety of subjects from rhetoric to history and economics, made a deep impression on some of Smith's notable contemporaries. They also had a marked influence on Smith's own career, for in 1751, at the age of 27, he was appointed professor of logic at Glasgow, from which post he transferred in 1752 to the more remunerative professorship of moral philosophy, a subject that embraced the related fields of natural theology, ethics, jurisprudence, and political economy.

Glasgow. Smith then entered upon a period of extraordinary creativity, combined with a social and intellectual life that he afterward described as "by far the happiest, and most honourable period of my life." During the week he lectured daily from 7:30 to 8:30 AM and again thrice weekly from 11 AM to noon, to classes of up to 90 students, aged 14 to 16. (Although his lectures were presented in English, following the precedent of Hutcheson, rather than in Latin, the level of sophistication for so young an audience today strikes one as extraordinarily demanding.) Afternoons were occupied with university affairs in which Smith played an active role, being elected dean of faculty in 1758; his evenings were spent in the stimulating company of Glasgow society.

His friends and acquaintances

Among his wide circle of acquaintances were not only members of the aristocracy, many connected with the government, but also a range of intellectual and scientific figures that included Joseph Black, a pioneer in the field of chemistry, James Watt, later of steam-engine fame, Robert Foulis, a distinguished printer and publisher and subsequent founder of the first British Academy of Design, and, not least, the philosopher David Hume, a lifelong friend whom Smith had met in Edinburgh. Smith was also introduced during these years to the company of the great merchants who were carrying on the colonial trade that had opened to Scotland following its union with England in 1707. One of them, Andrew Cochrane, had been a provost of Glasgow and had founded the famous Political Economy Club. From Cochrane and his fellow merchants Smith undoubtedly acquired the detailed information concerning trade and business that was to give such a sense of the real world to *The Wealth of Nations.*

The Theory of Moral Sentiments. In 1759 Smith published his first work, *The Theory of Moral Sentiments.* Didactic, exhortative, and analytic by turns, the *Theory* lays the psychological foundation on which *The Wealth of Nations* was later to be built. In it Smith described the principles of "human nature," which, together with Hume and the other leading philosophers of his time, he took as a universal and unchanging datum from which social institutions, as well as social behaviour, could be deduced.

One question in particular interested Smith in *The Theory of Moral Sentiments.* This was a problem that had attracted Smith's teacher Hutcheson and a number of Scottish philosophers before him. The question was the source of the ability to form moral judgments, including judgments on one's own behaviour, in the face of the seemingly overriding passions for self-preservation and self-interest. Smith's answer, at considerable length, is the presence within each of us of an "inner man" who plays the role of the "impartial spectator," approving or condemning our own and others' actions with a voice impossible to disregard. (The theory may sound less naive if the question is reformulated to ask how instinctual drives are socialized through the superego.)

Smith's view of human nature

The thesis of the impartial spectator, however, conceals a more important aspect of the book. Smith saw humans as creatures driven by passions and at the same time self-regulated by their ability to reason and—no less important—by their capacity for sympathy. This duality serves both to pit individuals against one another and to provide them with the rational and moral faculties to create institutions by which the internecine struggle can be mitigated and even turned to the common good. He wrote in his *Moral Sentiments* the famous observation that he was to repeat later in *The Wealth of Nations:* that self-seeking men are often "led by an invisible hand . . . without knowing it, without intending it, [to] advance the interest of the society."

It should be noted that scholars have long debated whether *Moral Sentiments* complemented or was in conflict with *The Wealth of Nations,* which followed it. At one level there is a seeming clash between the theme of social morality contained in the first and the largely amoral explication of the economic system in the second. On the other hand, the first book can also be seen as an explanation of the manner in which individuals are socialized to become the market-oriented and class-bound actors that set the economic system into motion.

Travels on the Continent. The *Theory* quickly brought Smith wide esteem and in particular attracted the attention of Charles Townshend, himself something of an amateur economist, a considerable wit, and somewhat less of a statesman, whose fate it was to be the chancellor of the exchequer responsible for the measures of taxation that ultimately provoked the American Revolution. Townshend had recently married and was searching for a tutor for his stepson and ward, the young Duke of Buccleuch. Influenced by the strong recommendations of Hume and his own admiration for *The Theory of Moral Sentiments,* he approached Smith to take the charge.

The terms of employment were lucrative (an annual salary of £300 plus traveling expenses and a pension of £300 a year thereafter), considerably more than Smith had earned as a professor. Accordingly, Smith resigned his Glasgow post in 1763 and set off for France the next year as the tutor of the young duke. They stayed mainly in Toulouse, where Smith began working on a book (eventually to be *The Wealth of Nations*) as an antidote to the excruciating boredom of the provinces. After 18 months of ennui he was rewarded with a two-month sojourn in Geneva, where he met Voltaire, for whom he had the profoundest respect, thence to Paris, where Hume, then secretary to the British embassy, introduced Smith to the great literary salons of the French Enlightenment. There he met a group of social reformers and theorists headed by François Quesnay, who called themselves *les économistes* but are known in history as the physiocrats. There is some controversy as to the precise degree of influence the physiocrats exerted on Smith, but it is known that he thought sufficiently well of Quesnay to have considered dedicating *The Wealth of Nations* to him, had not the French economist died before publication.

The stay in Paris was cut short by a shocking event. The younger brother of the Duke of Buccleuch, who had joined them in Toulouse, took ill and perished despite Smith's frantic ministrations. Smith and his charge immediately returned to London. Smith worked in London until the spring of 1767 with Lord Townshend, a period during which he was elected a fellow of the Royal Society and broadened still further his intellectual circle to include Edmund Burke, Samuel Johnson, Edward Gibbon, and perhaps Benjamin Franklin. Late that year he returned to Kirkcaldy, where the next six years were spent dictating and reworking *The Wealth of Nations,* followed by another stay of three years in London, where the work was finally completed and published in 1776.

The Wealth of Nations. Despite its renown as the first great work in political economy, *The Wealth of Nations* is in fact a continuation of the philosophical theme begun in *The Theory of Moral Sentiments.* The ultimate problem to which Smith addresses himself is how the inner struggle between the passions and the "impartial spectator"—explicated in *Moral Sentiments* in terms of the single individual—works its effects in the larger arena of history itself, both in the long-run evolution of society and in terms of the immediate characteristics of the stage of history typical of Smith's own day.

Stages of social organization

The answer to this problem enters in Book V, in which Smith outlines the four main stages of organization through which society is impelled, unless blocked by deficiencies of resources, wars, or bad policies of government: the original "rude" state of hunters, a second stage of nomadic agriculture, a third stage of feudal or manorial "farming," and a fourth and final stage of commercial interdependence.

It should be noted that each of these stages is accompanied by institutions suited to its needs. For example, in the age of the huntsman, "there is scarce any property . . . ; so there is seldom any established magistrate or any regular administration of justice." With the advent of flocks there emerges a more complex form of social organization, comprising not only "formidable" armies but the central institution of private property with its indispensable buttress of law and order as well. It is the very essence of Smith's thought that he recognized this institution, whose social usefulness he never doubted, as an instrument for the protection of privilege, rather than one to be justified in terms of natural law: "Civil government," he wrote, "so far as it is instituted for the security of property, is in reality instituted for the defence of the rich against the poor, or of those who have some property against those who have none at all." Finally, Smith describes the evolution through feudalism into a stage of society requiring new institutions, such as market-determined rather than guild-determined wages and free rather than government-constrained enterprise. This later became known as laissez-faire capitalism; Smith called it the system of perfect liberty.

There is an obvious resemblance between this succession of changes in the material basis of production, each bringing its requisite alterations in the superstructure of laws and civil institutions, and the Marxian conception of history. Though the resemblance is indeed remarkable, there is also a crucial difference: in the Marxian scheme the engine of evolution is ultimately the struggle between contending classes, whereas in Smith's philosophical history the primal moving agency is "human nature" driven by the desire for self-betterment and guided (or misguided) by the faculties of reason.

Society and the "invisible hand." The theory of historical evolution, although it is perhaps the binding conception of *The Wealth of Nations,* is subordinated within the work itself to a detailed description of how the "invisible hand" actually operates within the commercial, or final, stage of society. This becomes the focus of Books I and II, in which Smith undertakes to elucidate two questions. The first is how a system of perfect liberty, operating under the drives and constraints of human nature and intelligently designed institutions, will give rise to an orderly society. The question, which had already been considerably elucidated by earlier writers, required both an explanation of the underlying orderliness in the pricing of individual commodities and an explanation of the "laws" that regulated the division of the entire "wealth" of the nation (which Smith saw as its annual production of goods and services) among the three great claimant classes—labourers, landlords, and manufacturers.

This orderliness, as would be expected, was produced by the interaction of the two aspects of human nature, its response to its passions and its susceptibility to reason and sympathy. But whereas *The Theory of Moral Sentiments* had relied mainly on the presence of the "inner man" to provide the necessary restraints to private action, in *The Wealth of Nations* one finds an institutional mechanism that acts to reconcile the disruptive possibilities inherent in a blind obedience to the passions alone. This protective mechanism is competition, an arrangement by which the passionate desire for bettering one's condition—"a desire that comes with us from the womb, and never leaves us until we go into the grave"—is turned into a socially beneficial agency by pitting one person's drive for self-betterment against another's.

The role of competition

It is in the unintended outcome of this competitive struggle for self-betterment that the invisible hand regulating the economy shows itself, for Smith explains how mutual vying forces the prices of commodities down to their "natural" levels, which correspond to their costs of production. Moreover, by inducing labour and capital to move from less to more profitable occupations or areas, the competitive mechanism constantly restores prices to these "natural" levels despite short-run aberrations. Finally, by explaining that wages and rents and profits (the constituent parts of the costs of production) are themselves subject to this same discipline of self-interest and competition, Smith not only provided an ultimate rationale for these "natural" prices but also revealed an underlying orderliness in the distribution of income itself among workers, whose recompense was their wages; landlords, whose income was their rents; and manufacturers, whose reward was their profits.

Economic growth. Smith's analysis of the market as a self-correcting mechanism was impressive. But his purpose was more ambitious than to demonstrate the self-adjusting properties of the system. Rather, it was to show that, under the impetus of the acquisitive drive, the annual flow of national wealth could be seen steadily to grow.

Division of labour

Smith's explanation of economic growth, although not neatly assembled in one part of *The Wealth of Nations,* is quite clear. The core of it lies in his emphasis on the division of labour (itself an outgrowth of the "natural" propensity to trade) as the source of society's capacity to increase its productivity. *The Wealth of Nations* opens with a famous passage describing a pin factory in which 10 persons, by specializing in various tasks, turn out 48,000 pins a day, compared with the few, perhaps only 1, that each could have produced alone. But this all-important division of labour does not take place unaided. It can occur only after the prior accumulation of capital (or stock, as Smith calls it), which is used to pay the additional workers and to buy tools and machines.

The drive for accumulation, however, brings problems. The manufacturer who accumulates stock needs more labourers (since labour-saving technology has no place in Smith's scheme), and in attempting to hire them he bids up their wages above their "natural" price. Consequently his profits begin to fall, and the process of accumulation is in danger of ceasing. But now there enters an ingenious mechanism for continuing the advance. In bidding up the price of labour, the manufacturer inadvertently sets into motion a process that increases the supply of labour, for "the demand for men, like that for any other commodity, necessarily regulates the production of men." Specifically, Smith had in mind the effect of higher wages in lessening child mortality. Under the influence of a larger labour supply, the wage rise is moderated and profits are maintained; the new supply of labourers offers a continuing opportunity for the manufacturer to introduce a further division of labour and thereby add to the system's growth.

Here then was a "machine" for growth—a machine that operated with all the reliability of the Newtonian system with which Smith was quite familiar. Unlike the Newtonian system, however, Smith's growth machine did not depend for its operation on the laws of nature alone. Human nature drove it, and human nature was a complex rather than a simple force. Thus, the wealth of nations would grow only if individuals, through their governments, did not inhibit this growth by catering to the pleas for special privilege that would prevent the competitive system from exerting its benign effect. Consequently, much of *The Wealth of Nations,* especially Book IV, is a polemic against the restrictive measures of the "mercantile system" that favoured monopolies at home and abroad. Smith's system of "natural liberty," he is careful to point out, accords with the best interests of all but will not be put into practice if government is entrusted to, or heeds, "the mean rapacity, the monopolizing spirit of merchants and manufacturers, who neither are, nor ought to be, the rulers of mankind."

The attack on mercantilism

The Wealth of Nations is therefore far from the ideological tract it is often supposed to be. Although Smith preached laissez-faire (with important exceptions), his argument was directed as much against monopoly as government; and although he extolled the social results of the acquisitive process, he almost invariably treated the manners and maneuvers of businessmen with contempt. Nor did he see the commercial system itself as wholly admirable. He wrote with discernment about the intellectual degradation of the worker in a society in which the division of labour has proceeded very far; for by comparison with the alert intelligence of the husbandman, the specialized worker "generally becomes as stupid and ignorant as it is possible for a human being to become."

In all of this, it is notable that Smith was writing in an age of preindustrial capitalism. He seems to have had no real presentiment of the gathering Industrial Revolution, harbingers of which were visible in the great ironworks only a few miles from Edinburgh. He had nothing to say about large-scale industrial enterprise, and the few remarks in *The Wealth of Nations* concerning the future of joint-stock companies (corporations) are disparaging. Finally, one should bear in mind that, if growth is the great theme of *The Wealth of Nations,* it is not unending growth. Here and there in the treatise are glimpses of a secularly declining rate of profit; and Smith mentions as well the prospect that when the system eventually accumulates its "full complement of riches"—all the pin factories, so to speak, whose output could be absorbed—economic decline would begin, ending in an impoverished stagnation.

Later life

The Wealth of Nations was received with admiration by Smith's wide circle of friends and admirers, although it was by no means an immediate popular success. The work finished, Smith went into semiretirement. The year following its publication he was appointed commissioner both of customs and of salt duties for Scotland, posts that brought him £600 a year. He thereupon informed his former charge that he no longer required his pension, to which Buccleuch replied that his sense of honour would never allow him to stop paying it. Smith was therefore quite well off in the final years of his life, which were spent mainly in Edinburgh with occasional trips to London or Glasgow (which appointed him a rector of the university). The years passed quietly, with several revisions of both major books but with no further publications. On July 17, 1790, at the age of 67, full of honours and recognition, Smith died; he was buried in the churchyard at Canongate with a simple monument stating that Adam Smith, author of *The Wealth of Nations,* was buried there.

Beyond the few facts of his life, which can be embroidered only in detail, exasperatingly little is known about the man. Smith never married, and almost nothing is known of his personal side. Moreover, it was the custom of his time to destroy rather than to preserve the private files of illustrious men, with the unhappy result that much of Smith's unfinished work, as well as his personal papers, was destroyed (some as late as 1942). Only one portrait of Smith survives, a profile medallion by Tassie; it gives a glimpse of the older man with his somewhat heavy-lidded eyes, aquiline nose, and a hint of a protrusive lower lip. "I am a beau in nothing but my books," Smith once told a friend to whom he was showing his library of some 3,000 volumes.

Personal qualities

From various accounts, he was also a man of many peculiarities, which included a stumbling manner of speech (until he had warmed to his subject), a gait described as "vermicular," and above all an extraordinary and even comic absence of mind. On the other hand, contemporaries wrote of a smile of "inexpressible benignity," and of his political tact and dispatch in managing the sometimes acerbic business of the Glasgow faculty.

Certainly he enjoyed a high measure of contemporary fame; even in his early days at Glasgow his reputation attracted students from nations as distant as Russia, and his later years were crowned not only with expressions of admiration from many European thinkers but by a growing recognition among British governing circles that his work provided a rationale of inestimable importance for practical economic policy.

Over the years, Smith's lustre as a social philosopher has escaped much of the weathering that has affected the reputations of other first-rate political economists. Although he was writing for his generation, the breadth of his knowledge, the cutting edge of his generalizations, the boldness of his vision, have never ceased to attract the admiration of all social scientists, and in particular economists. Couched in the spacious, cadenced prose of his period, rich in imagery and crowded with life, *The Wealth of Nations* projects a sanguine but never sentimental image of society. Never so finely analytic as David Ricardo nor so stern and profound as Karl Marx, Smith is the very epitome of the Enlightenment: hopeful but realistic, speculative but practical, always respectful of the classical past but ultimately dedicated to the great discovery of his age—progress.

BIBLIOGRAPHY. The complete works have appeared in a definitive edition, "The Glasgow Edition of the Works and Correspondence of Adam Smith," 6 vol. in 7 (1976–83), including vol. 1, *The Theory of Moral Sentiments,* ed. by D.D. RAPHAEL and A.L. MACFIE (1976), vol. 2, *An Inquiry into the Nature and Causes of the Wealth of Nations,* 2 vol., ed. by R.H. CAMPBELL and A.S. SKINNER, vol. 3, *Essays on Philosophical Subjects,* ed. by W.P.D. WIGHTMAN and J.C. BRYCE (1980), which contains the interesting "The History of Astronomy," vol. 4, *Lectures on Rhetoric and Belles Lettres,* ed. by J.C. BRYCE (1983), and vol. 5, *Lectures on Jurisprudence,* ed. by R.L. MEEK, D.D. RAPHAEL, and P.G. STEIN (1978). For the nonspecialist, ROBERT L. HEILBRONER (ed.), *The Essential Adam Smith* (1986), offers fairly extensive readings and short discussions of Smith's main works.

Among biographical works are JOHN RAE, *Life of Adam Smith* (1895, reprinted 1965); WILLIAM R. SCOTT, *Adam Smith as Student and Professor* (1937, reprinted 1965), including "An Early Draft of Part of *The Wealth of Nations,*" various documents, and correspondence; and DUGALD STEWART, *Biographical Memoirs of Adam Smith. . . ,* vol. 10 in *The Collected Works of Dugald Stewart* (1858, reprinted 1966).

DONALD WINCH, *Adam Smith's Politics: An Essay in Historiographic Revision* (1978), reinterprets Smith's place in the history of economic and political thought. ANDREW S. SKINNER and THOMAS WILSON, *Essays on Adam Smith* (1975), contains discussion by well-known scholars of various aspects of Smith's work. KNUD HAAKONSSEN, *The Science of a Legislator: The Natural Jurisprudence of David Hume and Adam Smith* (1981), compares their philosophical systems. Useful articles include ADOLPH LOWE, "The Classical Theory of Economic Growth," *Social Research,* 21(2):127–158 (Summer 1954); NATHAN ROSENBERG, "Adam Smith on the Division of Labour: Two Views or One?" *Economica,* 32(126):127–139 (May 1965), and "Some Institutional Aspects of the *Wealth of Nations,*" *The Journal of Political Economy,* 68(6):557–570 (December 1960); JOSEPH J. SPENGLER, "Adam Smith's Theory of Economic Growth," *Southern Economic Journal,* 25(4):397–415, 26(1):1–12 (April and July 1959); the entry by JACOB VINER, "Adam Smith," in DAVID L. SILLS (ed.), *International Encyclopaedia of the Social Sciences,* vol. 14, pp. 322–329 (1968); and the entry by ANDREW S. SKINNER, "Adam Smith," in *The New Palgrave: A Dictionary of Economics,* ed. by JOHN EATWELL MURRAY MILGATE and PETER NEWMAN, vol. 4 (1987), pp. 357–375, with a bibliography.
(R.L.He.)

Social Differentiation

Social differentiation refers to the recognition and establishment in society of differences between groups or categories of individuals. It is evident that not all differences between individuals give rise to social differentiation. On the one hand there are personal qualities or combinations of qualities that distinguish a particular individual from all others; and on the other hand there are some characteristics that may be possessed in similar degree or kind by a number of individuals—for example, height or weight, colour of hair or eyes, temperament—but which never, or very rarely, lead to social differentiation. Many differences between individuals, however, are invested with social significance (or may be said to be created by society), with the result that the individual is seen as belonging to specific groups and categories that define his position in society. Thus he can be described as male or female, middle class or working class, young or old, married or unmarried, or belonging to a particular ethnic, linguistic, or religious group; and these classifications will refer to typical or expected kinds of behaviour, styles of life, and opportunities.

Another way of describing the same phenomena is to say that each individual occupies or performs "social roles." A society is thus regarded as a system of differentiated roles, and the position of an individual in the society is, in some sense, a product of his various roles.

Whether social differentiation is conceived in terms of roles or in terms of the groups and categories to which an individual belongs or is assigned, it has diverse aspects that must be clearly distinguished. In the first place, differentiation is closely associated with stratification; many social differences are invidiously ranked, so that some categories or groups of individuals have higher status or belong to higher classes than others. Second, the social importance of the various recognized differences may be unequal, and it is generally the case that those differences accompanying social stratification have the greatest prominence. In society as a whole, the distinction between married and unmarried, or between young and old, is usually less significant than the distinction between an aristocrat and a commoner or between an executive of a large corporation and a manual worker. Third, the various positions that an individual occupies may be nonaggregative, incongruous, or conflicting. It does not seem possible, for example, to aggregate the roles of father and corporation executive into an intelligible whole; and from this it follows that differentiation needs to be observed in several distinct spheres of social life and that a distinction has to be made between those roles that fit together in a significant whole and those that do not. Incongruous or conflicting roles appear particularly in connection with stratification; thus, an individual may be highly educated but poor and socially despised, or on the other hand he may be wealthy or powerful but uneducated and lacking the cultural attributes of established members of an upper class. Fourth, it should be noted that social differentiation varies in character from one type of society to another, and even from one society to another within the same type, and that it changes over time. Indeed, one of the principal features to which sociologists have drawn attention is the growth of social differentiation (especially in the form of a more extensive division of labour) in the course of social development.

This article is divided into the following sections:

The criteria of social differentiation

A great many factors enter into the process of differentiation. Initially, a broad distinction may be made between two kinds of factors: those that depend upon biological differences among human beings and those that arise wholly or largely from social or cultural conditions.

BIOLOGICAL FACTORS

Sex differentiation. One of the most obvious differences among human beings is that between the sexes. In all societies men and women are treated differently (often unequally) and have different functions, or roles. Many social scientists from the 19th century onward (among them Karl Marx and Herbert Spencer) have suggested that the economic division of labour began with the division of tasks between the sexes; and this, it is generally held, would arise from biological differences—in particular, men's greater physical strength, which equipped them for the activities of hunting and warfare, and women's reproductive function, which tended to confine them to the area of the home or settlement. Many other social

Social differences between men and women and cultural differences between men and women in early societies, and notably the male dominance in political life, may have followed from this original division of labour; and these differences were perpetuated and developed in later societies. Thus, in spite of considerable variations in the social position of women from one society to another, there has been a general male predominance which has continued up to the present time. In the new industrial societies of the 19th century, women still lacked many of the civil rights (especially property rights) and political rights enjoyed by men, and although they have now gained these rights in most modern societies, considerable differences remain: women workers are concentrated in the less skilled and lower paid occupations; disproportionately few women are in the higher professions; and women usually play little part in politics at the national level. Nevertheless, there have been gradual changes in all these spheres during the present century, and at the same time movements for women's rights—from the woman suffragists to the Women's Liberation Movement—have been a significant element in the more general movements of radicalism and reform.

The social differentiation of the sexes poses an important problem that is relevant to all discussion of the biological factors in social life. Some writers have argued that the physical differences between men and women determine the major psychological differences—of intellect and temperament—that largely account for the differentiation of social roles. Against this view it has been argued that biological elements acquire importance only when they are socially recognized and established; that there are in fact great variations, between societies and between historical periods, in the relative positions of men and women; and that in modern times the differences between the social roles of men and women have tended to diminish. Moreover, the existence of substantial, innate psychological differences between men and women is not securely established, and such differences as do appear can be explained as a consequence of social and cultural influences rather than biological inheritance. In this case, as in many others, the question of the relative importance of nature and nurture is still unsettled. At least, however, it can be said that the differences between men and women need not necessarily lead to economic and social inequality; and it may be claimed further that there is no longer any reason in advanced industrial societies for these differences to produce any large degree of differentiation at all outside the sexual sphere.

Age differentiation. Another biological factor that leads to differentiation is age. In all societies a distinction is made between the young, the mature, and the old; and different roles are assigned to them. A more elaborate classification exists in many societies—for example, according to Hindu doctrine, four stages are to be distinguished in the individual's life (a period of preparation, a period as a householder, a period as a recluse, and a final period of renunciation of the world). In many modern societies it is common to distinguish the periods of childhood, youth (including the transitional stage of adolescence), middle age, and old age. The extent to which these stages are clearly defined or formally recognized, however, varies considerably from one type of society to another, as do the social attitudes toward the members of each category. In some tribal societies there are organized age sets that mark the progress of the individual toward full adult membership of the tribe. On the other hand, in modern industrial societies the points of transition from one age category to another tend to be less well defined, to change over time, and to vary from one sphere of social life to another. Childhood may be regarded as lasting until puberty (or, in social terms, adolescence), and it has corresponded also, to some extent, with the period of formal school education; however, the continuing extension of the period of education has meant that childhood now merges into youth while a condition of financial and social dependency continues, without the sharp break that occurred in earlier societies when the individual moved at an earlier age from school or family to adult work. At the same time, there has developed in modern societies in recent decades a more distinctive "teenage" culture within a broader youth culture. In the legal and political spheres there has been a tendency to recognize earlier maturation by reducing the age of majority and of acquisition of political rights (in Great Britain in 1970, for example, from 21 to 18 years), but there is still not a complete correspondence between the ages at which, in different spheres, individuals assume adult roles.

Inequalities among age groups

The evaluation of different age groups varies widely from one kind of society to another. In many tribal societies and traditional societies, the elders have considerable prestige and power as the possessors of a store of valuable knowledge and experience. In modern industrial societies, on the other hand, they may be regarded as old-fashioned and out-of-date, because of the rapid growth and transformation of knowledge.

Differentiation by age is accompanied by inequalities. In most societies economic resources and political power are in the hands of the middle-aged or old, though there are considerable variations and in periods of social upheaval the young may increase their influence or power. The inequalities arising from age differences, like those arising from the differences between the sexes, are in any case less prominent in most societies than those that emerge from property ownership and other directly social elements in stratification.

Racial differentiation. Race differences have been another important source of social differentiation. In the first place, it may be claimed that the various races of mankind have created widely different civilizations. But this assertion raises difficulties of the kind already encountered in discussing the differences between men and women, namely, that it is not possible to establish a strict relation between the physical characteristics and the psychological and cultural features. In this particular case it seems clear that the diverse forms of nations and civilizations are mainly the product of geographical, social, and historical influences, although the actual creation of nations may have owed something to a sense of racial distinctiveness, among many other factors.

Effects of modern colonialism and imperialism upon race relations

Perhaps the most important observation to be made about differences of race is that they may be either emphasized or treated as unimportant in different times and places, according to the prevailing cultural and political circumstances. Race relations in modern times have undoubtedly been most strongly affected by the colonial situation that followed the overseas expansion of the western European countries. Ideas of racial superiority and inferiority accompanied colonization, and these ideas have persisted up to the present time, sustained by the great discrepancies in wealth and power between the nations of western Europe and North America and the countries of the Third World. Not only were race differences and inequalities emphasized on a world scale during the period of modern imperialism, but they also became prominent within particular societies in which, for one reason or another, different racial groups lived together. Thus, in the Republic of South Africa, a white minority which possesses economic and political power maintains, through the institutions of apartheid, a strict separation between white, black, and coloured members of society in order to perpetuate its own supremacy. In the United States, a black minority originating in the African slave trade of the early period of colonialism has enjoyed formal political rights for a century, but it still occupies in practice a situation of economic, cultural, and political inferiority that is increasingly a source of major political conflict. More recently, in Great Britain, the growth of an immigrant population drawn mainly from the West Indies, India, and Pakistan has resulted in more widespread manifestations of racism and, on the other side, attempts by legislation and by other means to prevent racial discrimination.

The principal aspect of race differences in modern times, therefore, is that such differences have taken the form not merely of differentiation but of gross inequality. It is possible to conceive of a state of affairs in which the differences between races, and the cultural differences that have become associated with them, would be generally regarded as creating a pleasurable diversity of human life;

but this is far from being the case at the present time. The colonial conquests established inequality as a principal element in the relations between races, and the dissolution of the colonial empires is still too recent for their historical consequences to have been eliminated.

Intellectual differentiation. Besides the major biological factors of sex, age, and race, there are other genetic differences between individuals which may lead to social differentiation. One of those most frequently discussed is differences in intelligence. In this case social differentiation is directly linked with stratification; those who are intellectually more able, it is suggested, will attain the leading positions in society. This thesis is clearly most readily applicable to the modern industrial societies, in which social mobility (that is, movement from one level to another in the social hierarchy) is possible and even encouraged and in which, moreover, such mobility occurs to a large extent through the acquisition of educational qualifications. To say this, however, is to recognize that social differentiation and stratification do not by any means correspond with the variations in intellectual ability in all societies. Where there exists a rigid system of stratification (as in a caste system), an individual's position in society is fairly strictly determined by descent, and intellectual or other personal qualities can have little effect upon it. Again, intellectual ability may be more or less highly valued in different types of society; if the principal social activity is hunting or warfare, then physical strength, energy, and courage may be much more important. Even in modern industrial societies, the relation between intelligence and social position is far from straightforward. The inheritance of class position limits the possibility of social differentiation in terms of intelligence; moreover, it is generally acknowledged that there are social, as well as genetic, factors in intelligence itself, so that it may be said that those individuals born into high social positions will display, on average, high intelligence and hence are likely to maintain themselves in these positions, while those born into lower social strata will experience much greater difficulty in developing their intellectual qualities.

This consideration of the biological factors in social differentiation brings to light a question of crucial importance. Although sex, age, race, and individual qualities such as intelligence obviously play some part in social differentiation, they do not have a directly determining influence. Their effect depends upon the way in which they are conceived and treated in each society, and this, in turn, depends upon a great variety of social, cultural, and historical influences.

SOCIAL AND CULTURAL FACTORS

National or societal differentiation

Among the most obvious forms of social differentiation arising from social and cultural factors are (1) the diversity of nations and civilizations, and (2) the distinctions within societies between classes or status groups. The second of these forms will be considered later in discussing stratification. The distinctions between nations, or more generally between the numerous autonomous communities into which mankind has always been divided, are closely associated with differences of language and of culture. Primitive societies, relatively isolated from one another even in the same geographical region, differed in language, in kinship structure, in religious belief and practice, and in economic system—the latter depending partly upon the nature of the physical environment and partly upon the level of technological development in a particular society. Many of these features persist in present-day tribal societies, although they have been greatly modified by the impact of the industrial countries.

It is impossible now to trace the ways in which this differentiation of early human societies came about; and, even in the case of ancient societies that have left historical records, it is not easy to follow in detail the processes by which distinct political communities were created. War and conquest, associated with migrations, played an important part in creating new and larger scale societies, but there were also contrary processes of disintegration and further differentiation. At a later stage, the world religions played a large part in forming distinct areas of civilization. The modern nation-states in Europe were formed largely on the basis of language and historical traditions, while in other parts of the world, more recently, new independent nations have been created on the basis of a religious or cultural identity or have established themselves as the successors to somewhat arbitrarily defined colonial territories (and consequently face difficult problems of social integration).

In considering social differentiation on a world scale, social scientists have been concerned not only with the distinctive features of particular societies and civilizations but also with *types* of society as they have appeared in different historical periods. The French sociologist Émile Durkheim in 1893, for example, distinguished between societies characterized by "mechanical solidarity" and those characterized by "organic solidarity." In the former (primitive and early societies) individuals differ little from each other and are uniformly subjected to the authority of the social group; in the latter (especially modern societies) there is great internal differentiation, individualism flourishes, and individuals are bound to each other by mutual dependence, especially through the extensive division of labour. A distinction that is similar in certain respects is that made by the German sociologist Ferdinand Tönnies in 1887 between *Gemeinschaft* (community) and *Gesellschaft* (society); in the former type of social system, men live in small-scale groups and have direct, face-to-face relations with one another, whereas in the latter type they belong to large-scale groups and are related indirectly and impersonally. This distinction, like Durkheim's, is, at least in part, one between primitive and early societies on the one hand and modern societies on the other.

Other writers have established quite different distinctions between types of society, often making use of economic criteria. One of the best known classifications along these lines is that of Karl Marx, who distinguished several types in terms of their mode of production and the class system arising from it: primitive communism, ancient society, Asiatic society, feudal society, modern capitalist society. More recently, another kind of economic distinction has been made, between industrial societies and nonindustrial or preindustrial societies, the latter being referred to quite frequently as the "developing countries." At the same time, a distinction is now increasingly made between socialist and capitalist economic systems and between the societies that develop upon these different economic bases, even though there are important similarities between all the industrial countries as a social type, on one hand, and the developing countries as a contrasting type, on the other, in spite of differences of economic regime within each type.

Differentiation within societies

The consideration of social and cultural differences between whole societies, or types of society, or civilizations, involves treating these units as if they were themselves relatively homogeneous and undifferentiated. The aspect of social differentiation that has attracted most attention, however, is that which occurs within each society. Some kinds of internal differentiation related to biological characteristics, notably those resulting from sex and age differences, have already been mentioned; but the most prevalent types of social differentiation are those arising directly from social and cultural influences. Many social scientists have regarded the division of labour as the most important factor; one of the most obvious differences between individuals and groups in society is to be found in the kind of work that they perform. In the late 18th century the Scottish economist Adam Smith discussed not only the economic consequences of the extending division of labour but also its effects in determining the characteristics and life-styles of individuals. In the late 19th century the English philosopher Herbert Spencer conceived of the division of labour as a primary element in social differentiation and traced its development in the various parts of society. Similarly, Durkheim considered that the division of labour was the source of the individualism to be found in modern societies and that, although it posed problems of fragmentation and social disintegration, it offered the possibility of a new form of social solidarity based upon the interdependence of individuals and groups.

The great importance of the economic division of labour in producing social differentiation cannot be denied. Moreover, it is closely connected, as will be seen later, with the establishment and modification of social stratification. Nevertheless, it is now generally acknowledged that social differentiation extends beyond the economic division of labour and its effects; as one scholar noted, it "should be thought of as a differentiation of activity of every sort that has significance for the group." Spencer, for example, wrote in terms of the "specialization of functions" which resulted in a clearer distinction between the major parts of society—the governmental, the military, the ecclesiastical, the professional, the industrial—as well as a more intensive division of labour in the productive sphere. But the writer who brought out most plainly the complexity of social differentiation in modern societies was the German sociologist Georg Simmel. In his volume of essays *Über soziale Differenzierung* (1890; "On Social Differentiation") and in other works, he considered the great variety of influences that had contributed to the growth of individualism and the diversification of social groups in the western European countries during the 19th century. The rapid development of a money economy, the growth of cities, the mobility of individuals, the emergence of new social and cultural interests all have tended to produce more varied styles of life. In particular, city life offered the stimulus of diverse and competing intellectual and cultural outlooks, from which new kinds of differentiation could again arise, while the increase in the number of associations of all kinds, concerned with specific interests, permitted the individual to develop particular aspects of his own character and purposes. According to Simmel, in early societies the individual was more or less completely absorbed in the family group, but later he was able to enter into relationships with persons outside his circle on the basis of a similarity in character, tendencies, and activities. Thus new "social circles" arose which intersected the existing "natural groups." In modern societies a vast superstructure of social circles has grown up, and this process has been accompanied by an increase in individual liberty; in Simmel's view, the number of social circles to which the individual belongs is, to some extent, a measure of the development of civilization.

The twin processes of individualism and diversification

Although Simmel attached great significance to social differentiation, he also observed that differentiation was accompanied by a process of integration; and he suggested that human nature shows a fundamental need for both kinds of association—that association which differentiates individuals on the basis of competition and that which brings them together on the basis of cooperation. Thus, he observed that although there is a continual proliferation of more specialized circles because of the appearance of new occupations, new scientific pursuits, and new cultural styles, there is also a continual formation of large circles out of smaller ones—for example, the formation of a working class out of diverse occupational groups. This relationship between differentiation and integration can best be studied by looking at the historical development of social differentiation.

The development of social differentiation

The process of social differentiation has often been taken as one of the principal indexes of social development. Thus, in 1926 the American sociologist Cecil C. North observed that the most significant distinction between a primitive and a highly civilized society was probably the degree to which functional differentiation had taken place. This followed the view of many earlier sociologists. Herbert Spencer for instance, had argued that it is

> a character of social bodies, as of living bodies, that while they increase in size they increase in structure. . . . At first the unlikenesses among [society's] groups of units are inconspicuous in number and degree but as population augments, divisions and subdivisions become numerous and more decided . . . progressive differentiation of structures is accompanied by progressive differentiation of functions.

Spencer's ideas have been revived recently by the American sociologist Talcott Parsons, who has treated social change and development wholly as a process of increasing social differentiation.

It should be recognized, however, that although increasing social differentiation is one important aspect of social development, there are other features that have been regarded by many writers as having equal or greater significance in defining or explaining historical transformations of society. Spencer, for example, saw as one major element in social differentiation the emergence of dominant classes; this, too, was the crucial element in Marx's theory, which—in contrast to the idea of social development through a gradual, progressive differentiation of function—conceived of historical change as proceeding by revolutionary leaps whenever a new ruling class arose as a result of new modes of production. Many recent studies by sociologists and economists have emphasized particularly the distinction between industrial and nonindustrial societies and have considered industrialization, in which social differentiation is only one element, as the major form of social change.

Another important qualification must be applied to the view that social change consists essentially in the process of social differentiation. It is true that in many respects—especially in their occupational structures—modern societies are more differentiated and more heterogeneous than were earlier societies; but forces are also at work making for greater homogeneity. The rise of the European nation-states in the 19th century, the development of national systems of education, the improvement of communications, the increasing size of economic enterprises, and more recently the growth of the mass media have all contributed, in the opinion of many writers, to a greater uniformity of national culture and to the disappearance or attenuation of local and regional diversity. Such developments have been commented upon, from various points of view, in writings on "mass society" and the "consumer society." It has also been argued that increasing social equality in the industrial countries (a diminution of the differences between social strata and greater social mobility) has begun to create a more homogeneous national culture. The factual basis of this argument, however, is disputed, and the most that can be said probably is that the general rise of levels of living, rather than equalization, has had some effect in producing more uniformity in styles of life. The tendencies toward homogeneity are to be seen not only at the national level but also on an international scale, and it has been suggested that all the advanced industrial countries show certain fundamental similarities in their structure and culture—primarily because of industrialism, urbanism, vastly improved communications systems, and the influence of international economic organizations. It would be very difficult, nevertheless, to establish in a precise way the degree to which present-day societies are becoming more or less heterogeneous. What is important is to recognize that the two processes of differentiation and homogenization proceed together, both within and between societies.

Contrasting processes of differentiation and homogenization

THE PROCESSES OF DIFFERENTIATION AND STRATIFICATION

It will be evident from the preceding account that many socially differentiated roles are evaluated as being "higher" or "lower" in some scale of prestige. Social differentiation, that is, overlaps with social stratification, the arrangement of groups and categories of individuals in a hierarchical order. Men and women, different age categories, ethnic groups, religious groups, or occupations may be treated not merely as different but as unequal. Stratification within societies is one of the most important forms of differentiation, but it has been conceived in diverse ways. Most students of the subject would agree that its basis is primarily economic, but economic differentiation itself can be viewed from a variety of aspects. According to Marxist theory, the most important distinction in all societies beyond the most primitive was that between the owners of the means of production and those who supplied their labour power. These two groups constituted the principal classes in every society (though Marx also recognized the existence of intermediate and transitional strata), and the relations between the classes determined

Contrasting processes of differentiation and homogenization the general course of social events in all spheres of life. Depending upon the character of the means of production and, thus, to a large extent, upon the level of technology, different forms of class society have existed. Marx distinguished, in the Western world, ancient society, feudal society, and modern bourgeois (or capitalist) society; and he envisaged the advent of a new form of society, socialism, in the future. In all cases, the rise of new classes and the struggle between classes were the principal causes of major changes in social structure. In the modern capitalist societies, according to Marx, it was the industrial working class, in the various forms of labour movement that are its political expression, that would accomplish such changes and inaugurate a new society.

Marx was not alone in recognizing the importance of social classes; many 19th-century historians, political scientists, and sociologists made similar observations, but they did not all explain the origins and development of classes in the same way, nor did they attach the same political importance to class conflict or envisage the future, as Marx did, in terms of a working class revolution to establish socialism. Spencer drew attention to the role of dominant classes, and Durkheim recognized the pervasiveness of class conflict in the modern European societies, but both considered that classes could coexist harmoniously in the industrial societies, so long as these societies continued to develop their relatively open and democratic character. Other writers, such as the Italian economist Vilfredo Pareto, while adopting the conception of class from Marxist thought, argued that the existence of a governing class or elite was unavoidable and were chiefly concerned, as many later writers have been, with the effectiveness of the dominant class in maintaining a stable social order.

The Marxist scheme has, at various times, been ignored (largely because of its ideological orientation), criticized, and revised. One point of criticism concerns its applicability to all the diverse forms of stratification. It seems most relevant to the 19th-century European societies and to the transition from feudalism to capitalism in western Europe, but much less useful in dealing with such a phenomenon as the Indian caste system. Caste has undoubtedly an economic basis: the major caste categories (the *varṇas*) resemble feudal estates, while the effective caste groups (the *jātis*) are closely related to differences of occupation. Nevertheless, the complexities of the caste system and the sanctions maintaining it can hardly be understood without taking account of the influence of religious ideas; and the historical relations between castes seem far removed from those which characterize the conflict between classes.

Marxist view of class

The interrelations of class, status, and power. A more general criticism of the Marxist theory leads to an alternative conception of social stratification. In writings published posthumously in 1922 the German sociologist Max Weber made a distinction between class, status, and power, arguing that these represented different forms of stratification that coexisted but did not necessarily coincide in modern societies. According to Weber, class situation was mainly determined by the ownership or nonownership of property (more or less along the lines that Marx had described); status situation was determined by the attribution of "social honour" on the basis of consumption or style of life; and the power situation was determined by possession or nonpossession of the means of command. Thus the relative importance of class and status in the stratification system has to be investigated in particular societies and periods; the predominance of one or the other would depend upon economic and cultural conditions that affect the extent to which men become class conscious or status conscious. Moreover, political power is relatively independent of both class and status; it is not regarded, as in Marx's theory, as being merely a consequence or reflection of class position.

The introduction of the notion of social status (which was developed from the feudal idea of legal status) led in due course to the formulation of a distinctive theory of social stratification. In contrast to the Marxist conception—according to which society was divided into major classes engaged in economic, political, and cultural conflict over the basic structure of society itself—the new theory viewed society as a hierarchy of status levels, with each level determined by a social evaluation of the importance or worth of the activities carried on by individuals at that level. The emphasis was upon the placement of individuals in the social hierarchy, not upon the formation of groups and their interrelations. The fact that numbers of individuals occupied the same rank position was considered (if at all) as a secondary phenomenon, and those occupying the same position tended to be treated as forming statistical categories rather than social groups. Thus, in many recent studies of social mobility the various social levels have been defined in terms of occupational prestige, and the scheme of ranking that has been constructed in this way is acknowledged to be a partly arbitrary set of categories, not a system of significant social groups. One researcher in 1954, for instance, distinguished seven occupational prestige categories for the purpose of studying movement between lower and higher social levels in Great Britain, while a pair of researchers in 1967 used a prestige scale of 17 types of occupation for their study of mobility in the United States.

Conceived in this way, stratification is very closely connected with social differentiation in the form of the division of labour, since it is the differentiation of occupations and their evaluation and ranking that are taken to be the source of stratification. It is clear, however, that this does not provide a complete picture of the social hierarchy. Besides the ranking of occupations, and the construction of categories of occupations that are judged to have equal prestige, there is a process of formation of social groups that, through their pursuit of particular interests, play a distinctive part in the system of stratification. Most of what are called the liberal professions, and even the professions as a whole, have generally been regarded as constituting groups of this kind, and it has also been widely recognized that there are various types of elites, not all based upon specific occupations, that have a major influence in shaping the institutions and culture of a society. Furthermore, the various stratified groups and prestige categories are themselves closely connected with social classes, as may be seen from the fact that occupations are thought of, and referred to, as being upper class, middle class, or working class in character. As Simmel observed, the process of occupational specialization is accompanied by an opposite process of incorporation of the occupations into social classes.

There is still another aspect from which the association between status, class, and power should be observed. The prestige of occupations, which is often taken to be the crucial element in establishing a hierarchy of status, at least in modern societies, is determined by social evaluations, but one needs to ask how these evaluations themselves come to be made and accepted. The importance or worth of particular occupations is not something established by nature, and it may be argued that it is those persons who are already wealthy or powerful (that is to say, dominant) in society who decide the scale of prestige. In this sense, therefore, the existence of a status system presupposes the existence of dominant and subordinate classes or of governing elites and subject groups. Nevertheless, other influences upon the prestige of occupations also seem to arise directly from economic and cultural conditions—for example, the changes brought about by the rise of modern science and technology, by the growing secularization of culture, and by the development of mass leisure that has given higher incomes and greater prestige to entertainers of all kinds.

The relations between class, status, and power, as different forms of social stratification, are complex and variable. In some times and places the formation of classes and the struggles between them for political power may be of preeminent importance, as was the case in western Europe during the period of the rise of the bourgeoisie in feudal society, and again in the 19th century when the organized working class movement began to develop. Under other conditions, however, though class divisions exist, they may be less significant than differences of status arising from the diversity of occupations, ethnic origins, and styles of life, influenced by a variety of cultural factors including religious doctrines. It has been argued, for instance, that class differences in the Western industrial societies have begun to lose some of their cultural and political importance in recent decades as a consequence of such changes as the more extensive division of labour leading to a greater variety of status positions, greater social mobility, the expansion of education and of middle class occupations based upon educational qualifications, and higher levels of living. It remains a matter of controversy, however, whether or not social classes have yet ceased to be the principal source of political ideologies and conflicts in these societies.

Definitions of class, status, and power

Social mobility. In the earliest human societies, differences of class and status as one of the principal types of social differentiation may have emerged in various ways—partly as a result of the accumulation of property or the monopolization of leadership positions by particular individuals and families and partly by means of conquest and the enslavement of other social groups. Once established, these differences are maintained by inheritance and strengthened by the elaboration of cultural distinctions in dress, education (initially a distinction between literacy and nonliteracy, and, later, differences in quality and duration of education), and style of life as a whole; they are also consolidated by legal and religious codes. Occupations may become hereditary (either in a formal way, as in the Indian caste system, or on a customary basis); property is transmitted through the family; and political power is inherited by the members of royal and noble families. Of course, the inheritance of social position is never complete in the sense that no new men can attain to wealth, prestige, or power. Pareto exaggerated somewhat when he wrote that "history is a graveyard of aristocracies," but it is certainly true that in all societies there has been some circulation of individuals and families between different social levels. Even in the Indian caste system, which was a particularly closed form of social stratification, it was possible for individuals, and especially for caste groups, to rise in the social hierarchy by various means; and in the European feudal societies there were considerable opportunities for mobility, particularly after commerce and the towns began to develop.

It would seem that modern industrial societies are the ones in which social mobility has generally attained its highest degree and in which inheritance of social position has diminished most. But, in truth, it is difficult, and in many cases impossible, to make rigorous comparisons of the degree of social mobility in one society at different times or in different societies; for the judgments that are made about the relative openness of different types of society tend to be based more upon a consideration of formal opportunities and restrictions than upon an examination of actual movement between social levels. Thus there is controversy about whether or not social mobility in industrial societies has increased during the present century; and, moreover, it is quite clear that there is still a very significant inheritance of class position. The distribution of wealth is obviously highly unequal, and it is primarily through the transmission of property that families in the upper strata maintain their social position and exclude those from lower strata.

Ways in which stratification develops

There are some conditions under which social mobility undoubtedly increases. After a revolution, for example, there is likely to be a general reshuffling of the dominant groups in society; thus, in the U.S.S.R. after November 1917, and in the countries of eastern Europe after 1945, many individuals from subordinate groups in society rose to positions of considerable prestige and power, while others who had previously been dominant fell in the social hierarchy. But at a later stage there may be a new consolidation of status positions. This is particularly noticeable in the sphere of higher education; immediately after the revolutions in eastern Europe, the universities were thrown open to children of working class and peasant families, but as the new upper and middle strata of the population became established they began to acquire privileges for their children, in education as in other spheres, and the recruitment of university students from the working class and peasantry declined again.

Another, more gradual, process affecting social mobility has been the changing structure of occupations in the advanced industrial countries. There has been a steady decline in the numbers employed in agriculture and extractive industries and, more recently, the beginnings of a decline in the numbers of manual workers in manufacturing industry; whereas there has been a continual expansion of clerical, technical, and professional occupations. This process has gone farthest in the United States, where one-half of the employed population is now engaged in middle-class occupations, but it is taking place, also, in the European industrial countries and Japan. Its result is to move a proportion of each new generation from low status manual occupations to higher status white-collar occupations and, at some point, to create, as has been claimed for the United States, a predominantly middle-class society. But this process has further implications for social differentiation; insofar as it establishes, over a large area of society, a middle-class style of life, depending upon similarity of income and occupation, it diminishes differentiation and produces a greater uniformity. This is one aspect of the phenomenon mentioned earlier that sociologists have characterized as mass society. The extent of such uniformity is limited, however, even in the most advanced countries. There is still a clear distinction between the small group of wealthy property owners and the rest of the population; and there are many groups—such as certain ethnic groups, workers in backward or declining industries, and retired workers—who live in conditions of relative poverty far below the standards of middle class society.

THE INDIVIDUAL AND SOCIETY

The phenomenon of social differentiation throws some new light upon the relations between the individual and society. It is not the case that each individual, with his array of personal qualities, confronts the whole society to which he belongs. He is born into particular, differentiated social groups, and his life experiences in these groups do something to form his character and determine his outlook, aims, and achievements. Nevertheless, social differentiation may be regarded as one of the main factors responsible for modern individualism. In broad outline, Durkheim's contrast between early, relatively simple societies, in which the individual is molded to a single pattern of behaviour, and modern complex societies, in which the individual is more differentiated, seems valid if one takes into account not only the increasing division of labour but also the increasing cultural diversity (which arose partly from the criticism of established religious and moral ideas), the development of towns and the consequent opportunities for competition between ideas, and the greater mobility of individuals. Durkheim thought, in fact, that individualism had gone too far in France and was endangering the cohesion of the society. Simmel, on the contrary, writing at about the same time, toward the end of the 19th century in Germany, suggested that modern societies had achieved a better proportion of freedom and restraint than had earlier societies; the individual was freer to choose the social circles to which he would belong, but these numerous and diverse circles were more fully integrated into larger groupings and into the whole society. The proportion between freedom and restraint, between differentiation and individualism, on one side, and the overall regulation of social life, on the other, is clearly not established once and for all in a particular type of society. The industrial societies of the late 20th century seem to many observers to have moved toward greater restraint in many areas of life; the sheer size of organizations, the more powerful means of control in the hands of governments, and the emphasis upon national unity in a century characterized by large-scale international conflicts have all contributed to a loss of individuality and widespread sentiments of individual powerlessness. These tendencies seem to have been countered only in a small degree, in limited areas of life outside the economic and political sphere, by the cultural "permissiveness" that has developed during the past two decades. It remains to be seen how a new balance between individualism and the general regulation of social life (which is needed increasingly on a world scale) will be achieved in the rapidly changing conditions of the late 20th century; but there seems no reason to doubt that it will have to take account of the duality that Simmel observed—the human need, on one side, for differentiation, diversity of life, and individual expression and, on the other side, for cooperation with other men within an ordered social framework in which similarities rather than differences are emphasized. (T.B.B.)

Special categories of social differentiation in modern society

SEX DIFFERENTIATION: WOMEN

There is ample anthropological and historical evidence to indicate that cultures vary widely with respect to the roles that they assign to one sex or the other. Building their homes or shelters, making clothes, planting and cultivating the land, trading, and many other occupations may be regarded as women's jobs in one society, men's in another. With one exception, to be mentioned below, the criteria for the division of labour between the sexes—and consequently the feminine role—seem arbitrary, based on local customs and traditions, rather than on differences between men and women in terms of physical or mental capacities. Women's lesser strength and slighter build than men's, for instance, have not prevented them from being the carriers of burdens or from doing heavy agricultural labour and other strenuous and protracted physical labour. The one constant factor determining the division of labour in all preindustrial societies is the fact that women, because of an almost incessant preoccupation with childbearing and rearing throughout the greater part of their lives, are less mobile than men and are, therefore, allotted tasks nearer their homes. As a consequence, there are some occupations that in most simple societies belong to women's domain: all kinds of housework (such as weaving, basketry, cooking meals, as well as producing the receptacles in which to prepare them and for storing the ingredients), some specific agricultural tasks, and midwifery. Medicine was almost universally a female vocation in prescientific ages and cultures—partly because it has been a domestic craft, handed down from mother to daughter, and partly because in more primitive cultures women have often been believed to possess special magical powers. Healing the sick and wounded accords well with creating new life.

The biological and social roles of women

Popular conceptions of masculinity and femininity. One might be inclined to believe that tenderness toward babies, the love and care for young children, and the desire to help the weak and sick are innate feminine personality traits—predisposing women by nature, as it were, to become the infant-school teachers, nurses, and social workers of today—but it is a fact that there are wide cultural variations in child-care practices, and producing babies in large numbers is not the same as bestowing on each of them the individual loving care that modern pediatricians and child psychologists prescribe. In the headhunting Mundugumor tribe, described by the American anthropologist Margaret Mead, mothers reject their babies and produce a people as aggressive as themselves; in the Samoan villages studied by the same author, mothers hand their babies over, as soon as they are weaned, to an older child, itself often not more than six or seven years of age, who takes charge of its care and basic training—much as the Victorian upper class mother left the upbringing of her offspring to a series of nannies and governesses until they were ready to be sent to boarding school. Abortion and infanticide have been well-known means of population control in many parts of the world and could not have been practiced without the active cooperation of women. In short, it appears to be a fallacy to generalize from the present state of civilization even about as basic and apparently universal a psychological phenomenon as maternal love, because this, too, may take different forms of expression in different cultures and subcultures. Whatever its manifestations, child care is, at any rate, sex-linked in the sense that only women can bear and suckle babies; but fondness for children and tender care for the young are not female monopolies, nor are they characteristic of all women.

In other respects, the connection between the social and biological roles of men and women is much more tenuous and may even be purely coincidental. In some instances, roles have been assigned to one sex rather than another on no more rational a basis than are many social conventions governing the relationship between the sexes. Who greets whom first and whether men walk in front of their wives or vice versa are details of no consequence in themselves. What matters is that conventions exist and are obeyed and regulate even such minutiae of everyday life. That in Western culture one sex fastens their coats on the left and the other on the right side or that baby boys are dressed in blue and girls in pink may seem irrelevant trifles. But all serve the function of underlining, almost from the moment of birth, the contrasts between males and females, rather than their common humanity. The cumulative effect of these practices is by no means negligible. They accentuate the bipolarity of the sexes, and they give a clearer, if more rigid, definition of the roles assigned to either of them. This is particularly important in cultures simple and static enough to base their social organization on status assigned to their members by the accidents of birth, such as the order of seniority, the membership of a kinship group, a caste, or of one rather than the other sex. In that kind of society everybody knows his or her place; but it allows little room for change or adjustment to new situations.

Effects of modernization and democratization

As societies become more complex and technologically advanced, they depend increasingly for their further development on individual talent and initiative. The old system of social stratification, with its rigidly drawn lines between categories of people, is ill adapted to meet the challenges of expansion and social change and unsuited for the best utilization of the available human resources. Hence, it has to be gradually modified and to give way to a new system in which social status is based on personal achievement. The old criteria—age, sex, family membership, and so on—one by one lose their weight in determining a person's place in society.

In political terms, this means that modern societies are becoming more democratic, in the sense that they draw their elites from ever-wider catchment areas. The corresponding sociological process involves the gradual transition from a social organization based on assigned roles to one based on achieved roles. This is of particular interest in the present context; for, although women's growing equality of status, of educational and employment opportunities, and, now, among the younger generation, also of manners and mores may each be traced back to a number of more immediate and specific causes, the overall emancipation of women is a part, and at the same time an example, of the general trend toward the substitution of achievement for assignment as the determinant of social roles. To put it simply, what a person is capable of, has accomplished, or is interested in are becoming matters of greater importance in advanced societies than his or her sex, age, or family connections. Ultimately, this should amount to greater freedom of the individual and his or her assessment on personal merit—or, in the words of the United Nations Charter, "without distinction as to race, sex, language or religion." Why this consequence does not necessarily follow is a question outside the scope of the present article.

Changing fashions regarding masculinity and femininity

The statement that cultures differ from one another in the ways they attribute social roles to men and women implies that in a changing society such differences will occur regardless of local or national traditions. During one period, fashion and a man's status may demand that he adorn himself, wear decorative clothes, shoulder-length hair, or a wig; during another, it would be thought effeminate for him to do so. Similarly, there are fashions in psychological traits: during the Romantic era of the late 1700s and early 1800s, men freely expressed emotions of joy, sorrow, and tenderness; they shed tears without inhibitions and committed suicide for unrequited love. At other times, such demeanour would have been thought unmanly. During the 19th century, European women were regarded as "frail vessels," and, indeed, many suffered from "the vapours" and were liable to faint at the sight of a mouse—an image of femininity that, needless to say, did not apply to the women working in factories and coal mines or as domestic servants. Today, with an eye on population statistics, people are rather inclined to think of women as the tougher sex, because their expectation of life is longer and they show at all ages greater resistance to a variety of serious illnesses. In actual fact, women have become explorers, mountaineers, racing drivers, and pilots and have taken an active part in wars, resistance movements, and revolutions, not only in nurturing roles, as protectors of fugitives, as nurses, and as sutlers, but also as combatants. These activities may not have become part of the accepted feminine role, but neither are they rare enough to be considered unnatural or shocking. Women are actually recruited for compulsory military service, though usually not for combat, in some countries (as in present-day Israel or in Great Britain during World War II), which is a sure enough sign of the widespread acceptance of their changed social role; and the majority of women have lived up to the new image as naturally as their grandmothers did to the 19th-century idea of feminine frailty.

Arbitrary as the definition of male and female roles may appear from a comparative point of view, to the members of a stable society the various facets of a role form an integrated whole that is not only coherent in itself but also closely entwined with complementary roles and supported by a network of social norms that the individual contravenes at his peril. Moreover, they appear to be inevitable and the natural order of things, or they did until anthropological and sociological studies put the absolute nature of that order in doubt.

The concepts of masculinity and femininity, then, exist, above all, in the realm of abstract ideas. At different times, diverse and sharply contrasting qualities were attributed to men and women, thus keeping the idea of bipolarity intact. Rationality, for example, was thought characteristic of men, emotionality of women. Greater practical sense was believed to be typical at times of one, at other times of the other sex. Sometimes the same quality was given a different name according to whether it was manifested in men or women: whereas men's thinking was said to be logical, women were credited with intuition; to have a hunch—presumably a bit of intuition—was not unusual or discreditable for a man but a necessary starting point of scientific discovery and invention.

These and similar convictions, though firmly held by rational men and women, can today be consigned to the sphere of folklore. In reality it has proved impossible to isolate among the wealth of behaviour patterns and personality traits exhibited by women even within one society, let alone in different cultures, those characteristics that could be considered typically feminine in a general sense.

Scientific studies of male–female differences. On a commonsense level, it has always been assumed that psychological differences between the sexes must correspond to their differences in physiological structure and biological function—a view epitomized in Sigmund Freud's famous dictum, "anatomy is destiny." The ordinary layman reflecting on the nature of woman has been joined by psychologists, philosophers, biologists, students of animal behaviour, and other scientific experts in the search for the elusive essence of femininity. The history of scientific investigations into the psychology of women is too long and their findings too diverse and uneven in value to be discussed here at length; instead, some findings obtained by a variety of psychometric tests will be summarized, and then the main current theories of male–female differences briefly considered.

Two important qualifications must be made before any data are reported from the wealth of information collected in numerous, painstaking, and often ingenious studies measuring sex differences in abilities, interests, and personality traits. First, individual differences within any class of people outweigh the mean differences between one class and another. Second, all psychological measuring instruments are culturally relative; that is to say, their findings apply to the status quo in one particular culture (or, possibly, subculture) at a given time. Scales designed for the investigation of one kind of population may not be valid in another. Nor do they, of course, provide explanations as to the origin of the observed traits, although, in combination with other psychological studies, they often are useful indicators of causal connections between certain traits and other psychological phenomena (*e.g.,* motivation, cross-parent identification).

Sex differences in abilities and personality traits

Test surveys have shown significant psychological sex differences in contemporary Western society. As far as measured abilities are concerned, men have been found superior in speed and coordination of gross bodily movements, as opposed to manual dexterity, in which women have a higher mean score. Men are better in spatial orientation, mechanical comprehension, and arithmetical reasoning. Women, on the other hand, excel in speed and accuracy of perception and in memory, numerical computation, verbal fluency, and linguistic abilities generally.

Girls, on the whole, develop faster toward adulthood than do boys, both physically and mentally. This applies at all stages, from infancy, when they are earlier in learning to speak and to count, to adolescence and beyond. This fact is in line with the observation that during their school years girls at first equal if they do not, in fact, excel boys, even in arithmetical reasoning, spatial skill, analytical thinking, and some other predominantly masculine abilities. That is to say, girls reach their peak at an earlier age; boys forge ahead later. Developmental studies, testing in greater detail various intellectual abilities in preschool and school children of different grades, have on the whole supported these findings.

With respect to personality traits, men are characterized by greater aggressiveness, dominance, and achievement motivation; women by greater dependency, a stronger social orientation, and the tendency to be more easily discouraged by failure than men. As the miscellaneous nature of this summary indicates, there is a vast variety of personality traits that might be made the object of comparisons between the sexes. This has, in fact, been done, and the number of studies testing the differential distribution of various traits and types of behaviour between men and women is legion.

Experimental studies have led to the conclusion that masculinity and femininity are not mutually exclusive alternatives but are combinations of traits, unevenly distributed among individuals of either sex. Physically as well as psychologically, each person is a unique balance of characteristics that, if one so wishes, may be placed on a continuum, extending from maleness at one end to femaleness at the other, with a wide area of overlap at the centre, which one might, therefore, with some justification consider sexually neutral.

Masculinity–femininity (M–F) scales

"M–F scales" have been designed on which the degree of "maleness" and "femaleness" of individuals, or of classes of people, such as occupational or age groups, can be measured. Scores on the M–F scale can be and, indeed, have been correlated with other psychological characteristics, such as general intelligence, creativity, self-assertion, dependence, and passivity. These studies have led to further interesting findings; for example, people with a high intelligence quotient and individuals who score high on originality, creativity, and analytical thinking show other traits and interests more typically associated with the sex opposite to their own. In other words, men and boys who excel in these intellectual features are nearer the feminine, and women and girls nearer the masculine, end of the M–F scale than their less gifted peers of the same sex.

This does not mean that intellectual or artistic men are effeminate or that women in high-level professions are masculinized, nor that either type is ill-adapted to their appropriate sex roles. These findings simply confirm in quantitative terms what many a layman has sensed and such writers as Virginia Woolf have expressed when she said, speaking about women and fiction, "for anyone who writes . . . it is fatal to be a man or woman pure and simple; one must be woman-manly or man-womanly" (*A Room of One's Own,* 1929).

The fact that intellectuals, creative artists, original thinkers, and writers of either sex deviate from the norm as measured on the M–F scale, moreover, reflects a bias inherent in the nature of the measuring instrument. Because the scale has been based on a comparison between the mean performances of large numbers of people in a wide variety of tests, such qualities as mechanical ability and interest in engineering or sport have been placed toward the masculine end, whereas aesthetic interests, dependence, and desire for security have been placed toward the feminine end of the scale. What, in fact, are only indices of the relative distribution between men and women of certain qualities and interests have thus become criteria of masculinity and femininity. (The relativity of these standards is highlighted by the fact that engineering, which is practiced as a profession by an infinitesimal number of women in Western countries and has been found in test surveys, mostly administered in the United States, to be one of the most unequivocally masculine interests, is studied by a very high proportion of women in the U.S.S.R.)

The general conclusion from a great variety of investigations into the subject is that masculinity and femininity are matters of degree rather than contrasting concepts at two opposite poles, as they used to be thought of in the past. This view is now supported also by biologists and geneticists, who have found that human beings are ambisexual, in the sense that both men and women possess male and female sex hormones, albeit in different distribution, and the sex organs of either sex contain in rudimentary form those of the other; they consist, so to speak, of the same ingredients, mixed in different proportions.

The effect of sex hormones on human behaviour is not yet fully understood. It is known, however, that sex hormones enter the brain and affect its activity and that, vice versa, the brain exerts a powerful influence over the organs secreting hormones. For details of this process, see the articles REPRODUCTION AND REPRODUCTIVE SYSTEMS: *Human reproduction* and BIOCHEMICAL COMPONENTS OF ORGANISMS: *Hormones*. There is evidence from clinical observation of human beings and from experimental studies of infrahuman mammals that an increase in the amount of male or female sex hormones (as by injections) on members of either sex is followed by behaviour changes. The administration of androgen (a male hormone), for instance, appears to produce more aggression and greater motivation to initiate sexual relations (which in females is, nevertheless, appropriate to their own sex); increased levels of estrogen (a female hormone) appear to lead, among other things, to more passive behaviour and emotional disturbances.

The effect of learning on sex roles

In creatures as dependent on learning as human beings, these hormonal influences produce predispositions to behaviour of one kind or another rather than long-term behavioral changes in the face of many other factors at work, such as existing social norms and past experiences. (For this reason, the developmental stage at which the amount of sex hormones is artificially increased makes a great difference as to their effect.)

To test the relative importance of constitutional and environmental factors in determining the psychosexual orientation of individuals, one group of researchers carried out a series of studies of pseudohermaphrodites—that is, people whose physical sex did not, for a variety of reasons, coincide with the sex assigned to them at birth. With very few exceptions, these persons were, as adults, psychologically fully adjusted to their incongruously assigned sex role. The conclusion of the researchers was that a gender role is entirely the product of learning during the first three years of life and is independent of hormonal, gonadal, or chromosomal sex.

Explanations of how the process of learning a sex role takes place vary and are given by psychologists, according to their own theoretical orientation, in terms of conditioning or imprinting, reinforcement by reward and punishment (which includes subtler forms of social control such as praise, on the one hand, and ridicule, on the other), and the imitation of models. In the latter case, it is usually assumed that a child or adolescent as a rule identifies with the parent of the same sex. Cross-parent identifications, however, do occur and are often thought to be responsible for the development of abilities and personality traits considered more characteristic of the opposite sex. Identification with one rather than the other parent may be due to temperamental affinity between parent and child or to the greater affection, strength of personality, or power of one parent. It has been argued that, in societies in which power is unevenly distributed between the sexes, identifications will more frequently occur with the parent of the more powerful sex. Matters get more complicated—indeed, almost impossible to unravel—however, when, for instance, a daughter identifies with her father, who—himself the son of a widow, or for other reasons lacking a male model during his formative years—had identified with his mother. In such a case (which one might call double-cross-identification), a girl may, in fact, acquire her most feminine qualities through imitation of her father.

In conclusion, no statement can with any certainty be made about the origins of feminine or masculine personality traits beyond saying that psychosexual orientation appears to be the outcome of complex interactions among genetic, hormonal, and environmental factors whose relative importance in the whole process of character formation is impossible to ascertain. This view is now almost universally accepted, by biologically oriented scholars no less than by psychologists and sociologists.

Contemporary roles and status of women. The most decisive and, it would appear, irreversible changes in the status of women were initiated in the 19th century; and the circumstances which got them slowly and falteringly under way were the technical, economic, and social upheavals generally known as the Industrial Revolution. Its impact on the lives of women was profound and manifold and can here only very briefly be indicated. It is important to remember, however, that British Prime Minister Benjamin Disraeli's much-quoted phrase of the "Two Nations" ("the Privileged and the People") applied to women to an even more marked degree than to men.

Effect of the factory system

The transfer of production from home to factory, while sharply increasing the productivity of labour, destroyed the family as an economic unit. Henceforth many thousands of men, women, and even children left their homes to work in factories and mines, living in industrial slums or crowded tenements in big cities. Many women of the new industrial proletariat worked in appalling conditions of near-starvation and unlimited working hours under a system of exploitation that became notorious as "sweated labour"—while they gave birth to one child after another. The conditions of the labouring poor, however, eventually aroused public concern, expressed in social investigations, philanthropic work, and, finally, legislation mitigating bit by bit the worst evils of the factory system through state intervention.

The front of social prejudice was breached by the daughters of wealthy merchants and professional men by way of philanthropic work. The social evils of their day were crying out for remedy and increasingly stirred the public conscience. Helping the poor was both virtuous and unpaid and, therefore, a suitable activity for young ladies—especially in an era that had ceased to trust Providence to put things right and had sufficient self-reliance to take it upon itself to redress the grievances.

Soon there were women of outstanding abilities pioneering in many fields of social service: working for prison reform; joining in hospital administration; organizing nursing as a profession; improving the standards of working class housing; fighting alcoholism, cruelty to children, and prostitution; campaigning for the abolition of slavery; and engaging in social investigations and polemics for law reforms. Their names are too many to bear repeating; they mark milestones in the social history of the 19th century. Some, but by no means all, of them worked for specifically feminine causes, such as a mother's right to the custody of her children, reforms of marriage and divorce laws, more equitable property rights. *The* cause, of course, was votes for women.

Women in the labour market

Somewhere between the upper class social reformer and the female proletariat there was a third category of women, who slowly but in ever-increasing numbers entered the labour market—namely, clerical workers. These were of mixed social origin but came mostly from the classes that hitherto had provided the governesses and the "distressed needlewomen." They did not oust men from their jobs but entered a new type of employment, created by such technical inventions as the telephone, the telegraph, and the typewriter, by the vast expansion of business, and by such new institutions as the general post office, the savings bank, and the like.

The obstacles put up by Victorian prudery against the employment of women in any of these fields seem preposterous in retrospect. Women clerks in the civil service, for instance, had to work behind locked doors that no male—not even the tea-carrying messenger—was allowed to cross. When it was suggested that girls might suitably be employed in the letter-sorting section of the post office, the objection was raised that letters might sometimes have to be opened to discover the sender of nondeliverable mail; in this way, words or phrases might come to the notice of ladies which would offend their sensibilities. These and similar matters were earnestly debated in legislatures and fully reported. Fortunately for the emancipation of women, the argument that usually carried the day was that at the same or lower wages one could engage young women of much superior quality (in every sense of the word) in comparison with men.

The changes that in modern times are taking place in the status and role of women everywhere have been viewed from two angles: as part of a world revolution of family patterns or else as a development toward greater equality of the sexes in political rights, access to education, and employment opportunities—women's emancipation, in short. The difference is one of emphasis: in the former case, the focus is on woman's familial role; in the latter, the perspective includes her position in the social, political, and economic structure of society at large.

The two aspects, women's roles in family and in community, are closely interrelated: changes in one sphere normally produce changes in the other—not surprisingly, since each individual is simultaneously a member of both social systems. Women's status, both in the family and in wider society, has been and is being radically altered by a combination of scientific and technological advances, growing industrialization, ensuing changes in the economic and social structure, and the spread of new ideologies.

Changing family patterns. The revolutions in family patterns and in the social and political position of women are not equally far advanced, just as industrialization and urbanization are not equally developed, in all parts of the world. Even among the industrially highly developed countries, there are some notable variations in the status of women, owing to different historical traditions, social customs, class distinctions, religious affiliations, the "generation gap," and other factors. Bearing all these reservations in mind, it still does not seem unreasonable to assume that Western models of family patterns and of women's roles in and outside the family will gradually spread to other parts of the globe in the wake of economic and technological development.

The example of Japan

Japan is a good, if somewhat extreme, example illustrating this process. Starting from a rigidly patriarchal form of the extended family system in which women occupied completely subordinate positions, changes that in western Europe and the United States took over a century to accomplish were telescoped into less than a quarter of that time. In either case, a world war, with its manpower shortages and the need to recruit women for war work, effected changes of public opinion, which led to constitutional and practical reforms. In the West it was World War I, in Japan World War II that triggered the political and economic emancipation of women. The Constitution of 1946 guaranteed Japanese women equality before the law and abolished political, economic, and social discrimination on the ground of sex. Subsequently, a number of legislative measures were taken to reform the family system and secure for women legal equality in the home, in the sphere of education, and in the labour market. A Women and Minors' Bureau was set up at the Ministry of Labour to help and encourage women to take advantage of their newly won status.

Coeducation was adopted in Japanese high schools in 1950 and later in most universities as a means of furthering the emancipation of women. Although a nine-years course of elementary education is compulsory for both sexes, the proportion of girls advancing to high school level was 63 percent and of those proceeding further to a four-year university course was 17.6 percent, in 1964. As in many other countries, women students outnumber men in literary subjects.

More than 60 percent of Japanese women leaving senior high schools, colleges, and universities hold paid positions, mostly clerical and professional. (Considering that within the lifetime of the present generation some firms would not allow their female employees to answer the telephone lest it be thought impolite for customers to be put in a position of having to discuss business affairs with a woman, this is progress indeed.) More than half the female population aged 15 and over are economically active. Women form 40 percent of Japan's total working population, compared with 36 percent in the United States, Great Britain, or West Germany. The Labour Standards Law (1947) established equal pay for equal work while at the same time providing protective measures, such as permission to take menstruation leave and prohibition against night work in industry and against employment in certain hazardous occupations, such as mining.

It would be a fallacy to deduce that, as far as employment is concerned, the status of women in Japan differs little from that in other highly industrialized countries. Quite apart from the fact that the time span allowed for the adjustment of attitudes to the new situation has been very short, the Japanese economy itself, despite its rapid industrial growth, is so much at variance with the contemporary Western model that it presents an entirely different employment pattern. Only about one-half of all economically active women are wage earners or salaried employees. One-third are unpaid family workers, employed in small family businesses or farmsteads without pay. In agriculture, which not only is a declining industry but also is increasingly becoming feminized (in 1968 53 percent of all persons working on the land were women, compared with 50 percent ten years earlier), more than eight out of ten women work without fixed financial reward.

The rapid expansion of industry and the resulting rising wage levels have attracted men into the industrial sector of the economy, leaving the family small holdings, which are the mainstay of Japanese agriculture, for the women and old men to run. But the flight of young women from the land—at the same time, an escape also from the traditional, patriarchal type of family in which two or three generations, now separated by ideas as well as by age, live under one roof—has become even greater. It is getting more and more difficult for a farmer to find a wife—yet another push in the direction of rural depopulation, a phenomenon that is, of course, not unfamiliar in industrial countries elsewhere.

Political rights. It has now been accepted as axiomatic that equal rights to vote and to be elected to national office are fundamental to women's status. Equality of franchise with men was fought for ardently and for a long time by a dedicated minority against heavy resistance on the part of the "establishment." One of the functions of the United Nations Commission on the Status of Women, set up in June 1946, was to further the cause of women's political rights in all countries. In December 1952 the General Assembly adopted the UN Convention on the Political Rights of Women, which was the first instrument of international law aimed at the granting and protection of women's rights on a worldwide basis.

Rights to vote and hold office

By 1971, of the 129 countries that were members of the UN or the specialized agencies or were parties to the Statute of the International Court of Justice, all but eight allowed women to vote in all elections and to be eligible for election on the same basis with men. In three of these eight (Portugal, Syria, and San Marino), women's suffrage or eligibility for office was subject to certain limitations that were not on a par with men's. In five other countries, women had not by that date been enfranchised; namely, Jordan, Kuwait, Saudi Arabia, Yemen (Ṣanʿāʾ), and Liechtenstein.

The first countries to grant women electoral equality with men were New Zealand (1893), Finland (1906), Norway (1913), and Denmark and Iceland (1915). With the exception of a few individual states of the Australian Federation (1902) and the United States (where Wyoming was the first to grant women the suffrage, in 1869), equality of votes for women came in most leading countries at the end of or soon after World War I: The Netherlands and the Soviet Union in 1917; Germany and Luxembourg in 1919; Austria, Czechoslovakia, Poland, and Sweden in 1918; the United States in 1920; and Great Britain in 1928 (though subject to a number of limitations, age and marital status among others, British women were first enfranchised by the Representation of the People Act of 1918). During the interwar period, electoral equality was extended to women in the Union of South Africa (1930), Spain (1931), Brazil and Siam (now Thailand; 1932), Ceylon, Cuba, Turkey, and Uruguay (1934), Burma and Romania (1935), and the Philippines (1937); in the other states it came in the wake of World War II or else of the replacement of the old colonial empires with new sovereign states.

The extent of women's political power

The right to vote, though an essential means of influencing the distribution of political power in a democracy, does not by itself carry political power. The proportion of women elected to be members of Parliament is in the region of 3–4 percent in the British House of Commons, around 2 percent in the U.S. House of Representatives, and between 7 percent and 9 percent in postwar Federal German parliaments. In the Soviet Union in 1970, women deputies comprised more than 30 percent of the Supreme Soviet, the highest legislative body, but were much less well represented at the centres of real power—that is, the highest levels of the Communist Party. After half a century of women's suffrage, however, the number of women in high positions of political power and influence is still small enough for them to be known by name.

Leaving political power aside, women's vote seems also to bear little relation to their social position in other respects. In France, to give one example, women have always enjoyed high status not only in the family but also socially and culturally. The country has produced a number of eminent women in literature and politics (incidentally, even two ministers of state in the Popular Front Government of 1936, before Frenchwomen had the vote) and includes among its national heroes a woman soldier (Joan of Arc). Frenchwomen, nevertheless, were not enfranchised until 1944.

Switzerland, one of the oldest democracies, did not give women the vote until 1971. This is all the more curious because not only have Swiss women played a considerable part in the economy of their country but Switzerland was among the first to extend university education to women on a basis of complete equality with men. The University of Zürich admitted the first woman student in 1864 and subsequently became the Mecca to which women eager to get a university qualification flocked from all over the world. The University of Zürich set the model for the rest of continental Europe in allowing men and women students to work side by side in the same classrooms and medical schools, to take part in the same lectures and seminars, and to sit examinations on equal terms.

The connection between female suffrage and the general status of women is symbolical rather than real. True, there can be no equality that does not also include political rights. The reverse is, however, not the case. The granting of full citizenship is not necessarily a forerunner, and even less a guarantee, of an improvement in women's status.

Education. Education has a much more direct and powerful bearing on the social position of women (as, indeed, it also has on that of men). Hence, equal educational opportunities are of the greatest importance in raising their status. Equal access to education is one of the basic human rights laid down in the United Nations Charter and has been accepted, at least nominally, by most states.

Rates of women's educational enrollments

Nevertheless, even where equal access to education exists de jure, there is de facto a great difference of opportunity between the sexes. In countries where illiteracy is still widespread, it is at a much higher level among women than among men. (In India, for example, the literacy rate is nearly 34 percent among men, 13 percent among women.)

At the other end, even where university education is financed by public grants, as in Britain, or by public loans, as in the Scandinavian countries, irrespective of sex, parents will, in case of doubt or financial stringency, send their sons to university rather than their daughters. The dropout actually begins at secondary school level. According to an official British report, about one-half of the boys but only one-third of the girls capable of advanced-level work (that is, preparatory to higher education) stay at school after the age of 16. The corresponding watershed in the United States, if less marked, is at the end of high school. To take one scholastic year, 1961–62, as an example, somewhat more girls (51 percent) than boys graduated from high school, but among those enrolling in college the following term, 58 percent were men and 42 percent women. During the same year, slightly more than two out of three bachelor's and first professional degrees, more than two out of three master's degrees, and nearly nine out of ten doctor's degrees were conferred on men. The proportion of women in institutions of higher education is still considerably lower than that of men almost everywhere. (The term higher education includes teachers' colleges, technological institutes, and fine-arts colleges, as well as colleges and universities.)

The prospects of a working class girl getting a higher education are considerably lower than those of a working class boy. In Britain, according to one report, the chances of obtaining full-time higher education are four and a half times better for the son and eight times better for the daughter of a nonmanual worker than of a manual worker. As for full-time courses of degree level, the disproportion between the chances of men and women coming from these two types of family background is much greater still. If the father has a nonmanual occupation, his son's prospects are six times better and his daughter's prospects are 18 times better than the chances of boys and girls, respectively, from manual workers' homes.

Curricular differences between men's and women's education

There are also social pressures—exerted partly by tradition and public opinion, partly by the operation of market forces, and partly by vested interests on the part of older, established professions—to limit the professional and, hence, also the educational choices of women. Where the number of places is limited, as in many medical schools, the chances of women to be admitted are very much lower than men's. There is a tendency for women students to concentrate in large numbers on a limited number of subjects—languages, literature, education, usually social sciences (the list differs somewhat from country to country)—to the near-exclusion of most others. This seems particularly true of the United States, where as many as 40 percent of women students graduating in one year (1960–61) qualified in education, another 24 percent in social sciences, including psychology, and 14 percent in humanities, a substantial proportion of the latter probably also heading for the teaching profession. This leaves very small numbers graduating in medicine, the natural sciences, the fine arts, law, engineering, and agriculture.

In many other countries women have made an impact—to varying extent, and in diverse academic subjects—in traditionally masculine spheres of study. For instance, although law, natural sciences, and medicine together accounted for 13 percent of women students graduating in the U.S. in 1960–61, 40 percent of women students in Italy and 53 percent in France graduated in these three disciplines during the same period. In the Soviet Union at one time well over two-thirds—later reduced to more than one-half—of all medical students and 36 percent of students of engineering in higher educational institutes were women.

The rationale for denying many a gifted girl equal access with men to university education is usually that (1) she is not likely to become the breadwinner of a family; (2) her future socioeconomic status will most probably not depend on her professional occupation or earning capacity; and (3) her career will presumably be only of short duration, thus not warranting the investment of time, money, and energy. Some of these assumptions, as the following section will suggest, have been disproved by the facts. In the present context, under the heading of education, two points deserve mentioning. First, in all countries for which data are at hand (the United States, Great Britain, France,

Values of education in employment and marriage

West Germany, and the Scandinavian countries), there is a strong positive correlation between the educational level and the employment rate of women. Irrespective of their husbands' social status and income (both of which are as a rule higher and make the financial need for a wife's contributory earnings less pressing than among the average population), the more highly educated women are, the more strongly motivated they are to continue in or return to their careers. This development may be fairly recent, but it would seem to go a long way toward disposing of the argument that the higher education of women is wasted.

Second, in contrast to the earlier stages of women's emancipation, when it was feared that a university education might diminish, if not ruin, a girl's marital chances (a fear not entirely without substance at the time), education is today held in such high esteem, in developing as well as in developed countries, that it has gained prestige value. Husbands pride themselves on having a wife with a university qualification or professional training, irrespective of whether or not they are putting it to practical use. A degree is therefore an asset on the marriage market; and, since universities are places where young women are likely to meet congenial young men of the appropriate social class, a daughter's college education may be a better investment than a dowry. This may, in part, explain the rapid increase in the percentage of women among university students in countries as far apart as Argentina, India, Italy, and Japan, among others.

Contrasts between preindustrial and industrial societies

Employment. Before outlining the main features that characterize the employment situation of women in modern industrial societies, it is necessary to comment on the widely held belief that the employment of women, especially of married women, is a recent phenomenon. Women have, on the contrary, at all times and in all types of economy made a substantial contribution to the production and distribution of their communities' resources. In preindustrial societies, as, indeed, in agricultural communities today, the family is an economic unit in which all members do their share of work, as they also share in its proceeds. As far as the employment of women is concerned, the difference between preindustrial societies and highly industrialized modern economies is not that in the latter women represent an increasingly large proportion of the labour force but that most working women are employed outside their homes, as independent individuals, and receive a monetary reward in return for their labour.

In terms of labour statistics, the most highly developed countries do not always show the highest proportion of women among their civilian labour forces; but they have the lowest number of so-called unpaid family workers of either sex. If it is sometimes said that in advanced industrial societies more and more women go out to work, the emphasis is on going out rather than on work; but, even so, the statement is not entirely true. In many countries the sex employment ratio has remained remarkably constant, regardless of temporary fluctuations, ever since the first employment census was taken at roughly one woman to two men. (An exception to this generalization exists in the United States, where women formed only 15 percent of the labour force in 1870 and increased their participation to 36 percent in 1970.)

The present employment situation of women in Western industrial countries may be summarized thus. Because manpower—particularly skilled and highly qualified manpower—has become one of the scarce resources of modern times and because the reserves of rural populations that could be recruited into industry have been shrinking, modern societies increasingly depend on the work of women. Contemporary industries, highly mechanized, provide a large variety of skilled, semiskilled, and unskilled work that no longer requires physical force but depends on manual dexterity and speed—abilities in which many women excel.

Modern economies have developed a large apparatus of distribution and administration, and both the clerical and sales sectors employ vast armies of women. Moreover, the steady growth of the so-called tertiary, or service, sector is characteristic of all modern economies: from the expansion of the communications system to beauty parlours, from tourism and catering to the manifold government services, there are growing fields of employment largely depending on female labour. To these newer fields may be added those that are traditionally considered more largely women's domain: schoolteaching, librarianship, nursing, physiotherapy, social work, and so on.

The increase of married women in the labour force

Another striking characteristic of the present employment situation is the ever-increasing number of married women who go out to work. In industrialized countries, well over one-half of all working women (nearer 60 percent in the U.S., Great Britain, Canada, Japan, and Sweden) are married. In part, this is due to demographic factors. The extended life span, increased marriage rates, lower average age of marriage, relatively smaller families, and generally improved standards of health have led to a situation in which women no longer spend the greater part of their adult lives bearing children and looking after infants. Another result of some of the demographic trends is that well over one-third of the people in most industrial countries are either under the age of 15 or over the age of 65; that is, a growing percentage are below or above the working ages. In addition, rapid technological change and other complexities of modern societies seem to necessitate a steady extension of the period of education for a growing section of the population—thus further reducing the proportion of those available for work. It thus becomes clear why economic incentives as well as social and moral pressures operate to induce those of working age who are capable of doing so to join the labour force. Since virtually all men and single women, if they are not still under training, are gainfully employed, the only groups to which the choice is still open are married women, especially those who either have not yet had or no longer have young children.

The early marriages that have become popular today would in most instances not be possible without both partners contributing to the capital outlay needed for setting up a home and to the cost of running it. This is one of the reasons why it is now customary among young women of all social classes to carry on in employment after marriage until soon before the birth of the first child.

The increase in the number of married women working (more than one-third of the married women in Great Britain, France, West Germany, and the U.S.) is only partly due to the practice of continuing employment during the early years of marriage. It is due to a larger extent to the growing number of women who return to paid employment in their mid-30s and 40s; that is, after bringing up their families. The growth in the female labour force has primarily been the result of the increasing employment of married women between the ages of 35 and 55. A graph showing the employment rates of women by age groups in all industrial countries typically consists of a two-humped curve: one high peak at the ages of 18–20, then a steep decline until roughly the age of 35, after which there is a second, lower, and flatter peak (whose level has in recent years been progressively rising) for the age group 35–55.

The reasons for this development are manifold and can here be summarized only briefly. First, there is the reduction of the amount and physical strain of housework, resulting both from the smaller size of families and from the increasing mechanization of modern households. Second, in addition to simplifying the tasks that inevitably have to be carried out at home, industry has altogether taken over many processes that previously were part of the housewife's job. By modern research and mass-production methods, it has become possible to reduce the cost and improve the quality of a large variety of household products. While doing this, industry has at the same time created innumerable new jobs for women to fill.

In contrast to earlier stages of industrialization, which concentrated primarily on the production of industrial equipment, the latest phase of economic development in all advanced countries is marked by mass production of consumer goods—goods that ordinary men and women wish to possess and that may substantially alter the quality of their lives. In order to create a large enough market, industry relies heavily on advertising. The growth of the economy thus exerts a constant pressure on the public to buy more and more, and purchases on the installment plan make it possible to widen the circle of people to whom these goods become accessible. The inducement for married women to make an extra effort in order to increase the family income by going out to work is, therefore, considerable.

Among various other motives for married women to take up employment outside their homes, the most frequent are the need for personal contacts, resulting from the social isolation of housewives in dormitory suburbs and new housing estates, and the desire to exercise their abilities and put their previous training to good use.

The question of equal rates of pay for women

Despite the dependence of the economy on women's work and in spite of a growing egalitarian philosophy, women are almost universally either paid at a lower rate than men or employed in the lower grades or both. Equal-pay conventions have been passed by such organizations as the International Labour Organization, and many countries have enacted equal-rights legislation. Even though the United States is one of these countries (having passed its equal-pay act in 1963), the median earnings of year-round, full-time working American women was 58 percent that of American men in 1966. Although it is argued that the principle of "equal pay for work of equal value" is very difficult to establish in industry (By what standards are the respective values of particular processes to be assessed? Do men and women ever do the same kind of work?), professions that have traditionally been staffed by women—such as teaching, nursing, and social work—have a low pay structure that is only beginning to be improved by the growing infiltration of men into their ranks. It is true that in most of the high professions—including the executives, administrators, and scientific experts in industry and government, journalists and editors, physicians, lawyers, university teachers, legislators, and government officials—remuneration is related to the post, irrespective of sex. But it is much more difficult for women than for men to reach high-level positions, and it is a moot point whether this disparity of achievement between the sexes is primarily due to the entrenched male prejudice (as is often asserted), to the discontinuity of most women's careers, or to the reluctance of many women to assume positions of responsibility because their families come first in their order of priorities. The new employment patterns, outlined earlier, in which women's main career period coincides with their years of maturity, when they may well be able to give another 30 years' continued service, is too recent a development to have yet changed employers' attitudes. In the present climate of opinion it seems reasonable to assume that, as the number of women acquitting themselves with distinction in high office goes on increasing as it has done in recent years, the prospects of promotion will improve for all women of ability.

Marriage and sexual behaviour. The ideal of marriage generally accepted in contemporary Western societies and among the westernized strata of developing countries is that of partnership—that is, a sharing of interests and responsibilities between husband and wife on as nearly equal a basis as possible. As with so many other ideals, practice often falls short of precept, especially if, as in this instance, social class and family traditions sometimes combine with personal temperaments to put up an emotional resistance against the implementation of principles to which one is intellectually committed. There are, however, many factors, both material and ideological, which support a trend toward an increasing approach to the model, quite apart from the fact that it largely emanates from and is actually put into practice by the younger generation, especially the more highly educated among them, and is therefore likely to spread as more adolescents grow up to marriage age.

The ideal of marriage as a partnership is in line with the prevailing ideology of equality between the sexes. It is backed up by similar educational levels between husband and wife, by women's experience of personal independence and occupational as well as other extradomestic interests before marriage. Sharing of interests is, together with sexual attraction, a major factor in mate selection; and since joint interests—be they at work or at leisure—usually provide the focal points for people to meet, it is not surprising that they assume a central place in their relationship.

Roles of men and women in the home

If women's interests have increasingly stretched out beyond the home, men, in return, have become more home centred, partly under the influence of television and partly as a result of suburbanization. The mechanization of modern homes has made many domestic chores less dependent on traditional skills and less feminine in character. With many married women employed away from home and sharing in the breadwinning activities that previously were the husband's responsibility alone, a sense of fair play has perforce induced many men to share in some, at least, of the household tasks. Increased leisure hours allow this male participation in the home, and the do-it-yourself type of hobby in fashion encourages it. Psychologists, moreover, increasingly emphasize the importance of the paternal role in child rearing; in the present age of the mass media, these and similar theories and their practical implications are brought home to the widest possible public.

All these factors combine to make family life, in and out of the home, a more cooperative venture. Who, in these circumstances, is the dominant partner of a couple is a matter of personality characteristics rather than of social conventions. The times have passed when the father was the generally recognized authority in the family. Many of the more important family decisions, as well as the assertion of parental authority, have to be taken by the mother. Apart from the fact that a family, as a rule, moves wherever a husband's work takes him for any length of time, there are hardly any generally accepted conventions about male or female dominance, either in general or in any particular sphere.

This is one of the factors that have made compatibility between marriage partners an issue of crucial importance and that have made the choice—now left entirely to the couple concerned, without parental intervention—so difficult. In these circumstances, it would seem reasonable that a certain amount of experimentation in human relations goes on among young men and women before they make what, for all intents and purposes, is a permanent decision. This experimentation increasingly also includes premarital sexual relations.

The Kinsey female

When the Alfred Kinsey report *Sexual Behavior in the Human Female* (1953) brought to light the fact that 50 percent of the women in his sample had had premarital sex experience, the American public seemed surprised. Yet it is difficult to see how the peer-group culture that discourages any kind of chaperonage or control by adults and that accepts as natural facts such practices as mixed holiday camps, weekend trips, travel abroad, and so on could have had any other effect. The double standard of morality between men and women was repugnant to younger people long before the equality of the sexes in other respects had been widely accepted and long before safe contraceptives made equal standards a practical proposition.

Since the data were collected by Kinsey's group, further changes in the direction of greater permissiveness have no doubt taken place. Sex mores are in a state of flux. This does not mean to say that there is a trend toward more promiscuity. It may well be that new norms are evolving that have been called "permissiveness with affection" and that are not only more in line with the egalitarian standards accepted in other fields of activity but also less hypocritical than the bourgeois morality they seem to be replacing. Among the working classes and rural populations, premarital intercourse has previously been an accepted part of their mores. If, for whatever reason and in whatever guise, a similar pattern is spreading among the middle classes, this seems to be a sociologically unusual case of a social practice moving upward on the social scale. (V.K./Ed.)

AGE DIFFERENTIATION: YOUTH

Until comparatively recent years, attention was focussed primarily on the biological and physical aspects of youth, or adolescence. Physiological changes were visibly so striking that other, more subtle social factors were either ignored or undervalued. It was thought that the onset of puberty set up strong emotional states of mind characterized by periods of moodiness and storm and stress. This was a view that greatly appealed to romantic novelists and poets, but one that studies of different cultures later discredited.

The physical changes that characterize the postpubertal stage are indeed striking, but it has to be remembered that growth and change have already become familiar to children and of themselves need not seem threatening. It is the social attitudes that are built up around these body changes that determine young people's own attitudes and subsequent behaviour in relation to them. Physical growth, although subject to much variation among individuals, tends to accelerate at this stage. There are changes in glandular balance, and the development of secondary sexual characteristics is conspicuous. For girls, the onset of menstruation seems to have a critical significance, while the less clearly marked signs of puberty in boys, including seminal emissions, proceed, it seems, at only a slightly slower rate.

Physical maturation. A strong body of professional opinion suggests that during the last 50 to 100 years children have been maturing in the physiological sense at progressively earlier ages. It is not certain that this is a worldwide phenomenon, but it is agreed that potentially it is likely to become one as the more technically backward and underdeveloped nations become more prosperous. In some societies it seems that puberty is reached by as much as five years earlier than a century ago, and evidence derived from records kept over a number of years in schools and clinics suggests that children since about 1900 have on average increased in height, at age five to seven, by one to two centimetres (0.4 to 0.8 inch) each decade and, at 10 to 14, by two to three centimetres (0.8 to 1.2 inches) each decade. Body weight has also gone up proportionately. Only temporary disruptions of normal life caused by war, famine, or economic crisis seem to have halted this trend.

As far as earlier sexual maturing is concerned, the records largely rely on statistics that have been sporadically kept during the last century relating to the age of menarche (first menstrual period). Information for boys, concerned, for example, with the appearance of pubic hair, seems much less reliable.

National comparisons of the onset of puberty

Comparative data deriving from different sources over a long period of time are not entirely satisfactory, but there is a degree of consistency that suggests that the age of menarche in Europe has been getting progressively lower. Data from Norway, Germany, Finland, Sweden, Denmark, Great Britain, and the United States follow an almost identical pattern. At the present time Cuban girls have the earliest recorded menarche age, with a mean of 12.3 years for Negroes and 12.4 for whites. Chinese girls in Hong Kong are almost as early and well in advance of the European mean of 13.3. East Europeans appear to lead the west in physiological maturity, while the more prosperous American girls are also a step or two ahead. In Africa, the less well-off, such as the South African Bantu, lag seriously behind, with a mean age of 15.4. Only the Bundi of New Guinea, with an average menarche age of 18.8, are comparable to the Europeans of a century ago. Some experts consider that this trend toward earlier maturity will continue and that the mean age throughout the world will drop even lower. The reasons are probably both genetic and environmental. A better diet and higher standard of nutrition are clearly the most important, and this is supported by the finding that there is a significant relationship between the age of menarche and the number of children in the family—the larger the sibling group, the later the menarche. The same seems to hold true for children's height and weight. By implication, social-class differences and parental means seem to be important, since diet and medical care depend upon them to a considerable extent. There is some evidence that the onset of menarche and the development of secondary sexual characteristics do occur earlier in the more privileged and better off social groups, but the relationship between social class and maturation is not a simple one.

Earlier physical maturity has obvious social implications. Clearly, the lowering of the mean age for marriage in some societies is closely connected with it. Together with the later age of menopause, it has expanded the period of the individual's potential fertility at a time when world population is seen to be approaching a crisis point. Earlier maturity also has important consequences in the educational sphere, especially in more advanced societies, in which the years of formal training are necessarily becoming longer. Pedagogical practice is beginning to alter, sometimes radically, to meet the new levels of maturity of pupils. Parents, too, are finding that older disciplinary sanctions are becoming less and less effective, that simple veto or moral exhortation no longer go unchallenged.

Psychological maturation. Freudian theory tended to direct attention to the first few years of life as being the most crucial. It posited a latency period that seemed to operate until around the age of 11, followed by the necessary resolution of the Oedipal conflict (libidinal attraction of a child to the parent of the opposite sex) in the early teens that had to take place amid the rising tide of overt sexuality. Such a thesis tended to confirm the storm and stress notion of adolescence and to a very large extent depended on the idea of basic instinctual drives impelling the individual in more or less predetermined directions. Latter-day Freudians imply that it is always necessary to accept adolescents on their own terms because they can behave in no other way. Not to do so would produce frustration and be likely to result in various kinds of deviance, delinquency, and psychic sickness. Adolescence, in this view, is itself almost a kind of temporary mental illness that the young person simply has to endure, a disturbing but inevitable feature of the growing-up process that must be tolerated and cannot be either cured or altered. But both psychoanalytical and physiological theories have been considerably modified during recent years. Increasingly there has been a deeper realization of the social and cultural influences on the psychic development of the individual and a concomitant exploration of the sociocultural determinants of behaviour.

The growing-up process has come to be seen as a search for identity. Identity implies self-realization within a particular social environment and period. The individual, innermost self may be unique, but it must always be realized within specific and, historically speaking, widely different social milieus. There is a dynamic interplay between the inner self and the surrounding culture, and it is out of this tension that individual self-identity finally emerges.

Identities may be partly given or ascribed by the community in which the young person finds himself. They may also be achieved to some extent by hard work and competitive success. In some milieus the pressures to conform to a particular behaviour pattern are severe indeed. Some youngsters react by choosing a deviant identity, as this seems the only way open to them to become themselves and to actualize their inner potentialities.

In most cultures the acceptance and realization of a sexual identity is a fundamental social necessity. Male and female roles, even in "primitive" societies, are very different. Boys and girls during the developmental period are hence called on to solve different problems and to learn to behave in dissimilar ways. Boys early become conscious of their future tasks as breadwinner, father, and citizen and tend to focus their identity around work, vocation, and earning a living. Girls, by contrast, are less oriented to the working world and more concerned with the search for their future mate. They see themselves more as wives and mothers than as employees or professionals.

But the key orientation for all adolescents is the future. They know that they are passing through a transitional, preparatory stage for a fuller and longer period of adult life that is not far away. Boys, because of their vocational and breadwinning role, need to develop greater independence of thought and action than their sisters so that they can challenge the environment successfully. Girls, by contrast, become more involved in interpersonal relationships and are much less group minded. Boys, too, may be called on to defend the state with their lives or at least to protect their family and kin group. They need therefore to develop an external stance of greater courage and toughness than girls and are much more susceptible to accusations of cowardice. By the same token, they are under considerable social pressure to prove their masculinity in purely sexual terms—not to have had sexual experience at a time when their coevals claim to have had it is felt to be denigrating. Feelings of sexual incompetence, fears of impotence, and doubts about homosexual tendencies in boys are matched in girls by anxieties about attractiveness and lesbianism.

Varying physiological and psychological maturity and the demands of society

Socialization. The physiological and psychological development of the individual during adolescence is always closely associated with social influences and societal demands. These demands can at times pull in conflicting directions, thus imperilling the individual's inner harmony and self-image. Society requires the growing boy and girl to develop socially, as well as physically and psychologically, through a series of stages that are more closely associated with chronological age than general maturity. Differences between individuals composing the same status group can thus be considerable. Variations in the age of puberty, of physical growth and emotional stability, of family background, and of innate intelligence lead to further confusion and uncertainty. At this stage young people are acutely conscious of themselves and of differences between themselves and their contemporaries. They are inclined to be highly introspective, self-critical, hypersensitive to criticism, anxious to be accepted, yet impelled to assert their own individuality. Clashes with parents and teachers are common, largely because most of the adults with whom they come into contact are in a superior position and tend to misread every kind of youthful nonconformity as rebellion or conceit.

The fact that society expects its young people to develop along certain lines and to remain involved in either full-time or further education long after physiological maturity has been reached and passed has been called social adolescence. There is so much factual information and technical know-how to be transmitted that the span of formal education must necessarily be lengthened every generation. For the most highly specialized and theoretically demanding work, the period of continuous training could well last into the mid-20s. This means that more and more young people remain in economic dependence either on their parents or on society during their prolonged educational phase. It is not surprising that at times this lack of financial independence proves irksome and that some students react by excessive sensitivity to what they feel is authoritarianism and overregimentation by those who are in control of their studies and their destinies.

While individuals are arriving at their various maturational stages at differing times, the principal social and personal developmental tasks, as they have been called, must be tackled and, if possible, accomplished before final maturity and individuality are attained. A developmental task is one that arises at a particular stage of life, successful achievement of which leads to happiness and success with later tasks, while failure leads to personal unhappiness, social disapproval, and subsequent difficulty with the other tasks that lie ahead. Such tasks are closely interrelated, and failure at one level will almost inevitably produce stress and strain at another level. The young person who succeeds in mastering the earlier problems is more likely to be successful with the later ones.

The goal of these developmental tasks may be termed adult competence. In conventional social groups this implies doing well at work, succeeding in sexual activity and family life, and becoming a law-abiding, responsible citizen sharing the economic and political burdens of the community as a whole. Less orthodox groups might well scorn the notion of good citizenship. Others might regard economic success and material rewards as distasteful and even perhaps positively evil, while others, yet again, might prize sexuality for its own sake apart from either marriage or family life. But the norms that may be said to characterize society as a whole are those that the respectable middle class uphold. Almost by definition, the values of stable family life, conscientious work, and responsible citizenship comprise the normative order in most modern industrial urban societies.

The main developmental tasks that each young person in modern society is obliged to tackle may be summarized under three chief heads: (1) coming to terms with society, (2) coming to terms with self, and (3) coming to terms with life as a whole. Coming to terms with society consists of being educated, getting a job, participating in political affairs, and bringing up a family. Coming to terms with self involves attaining a valid sexual self-image and the realization of individual talents, especially artistic and creative, of every kind. Coming to terms with life as a whole is a more synthetic experience and refers to the acquisition of a general moral code, a set of values, and, in some cases, the acceptance and internalization of religious beliefs and attitudes.

The three main obligatory developmental tasks

These three are, of course, related aspects of a general process of self-realization and the acquisition of a valid identity. They comprise in fact the basic issues that face all people in all epochs. But adolescence is the stage when they confront individuals with the greatest urgency and emotional force. This is precisely why adolescence is traditionally regarded as a preeminent time of artistic and spiritual flowering. It is the reason why it is thought of as a time when the intensity of individual loneliness is most keenly suffered, when there is leisure to contemplate the great issues of life and death, love and rejection, faith and despair, before the commitments and responsibilities of adult existence force the mind to operate on a lower and more pragmatic level. The loneliness and at times the despair of modern life are tragically demonstrated by the fact that the suicide rate for this younger age group in Great Britain has nearly doubled since the end of World War II.

Functions and roles of youth. In primitive and stable cultures the function of youth is to continue the traditional way of life even when at the same time there is a tension between youths and adults for status and power. Adults maintain a jealous social distance between themselves and the young, who aspire to their prestige.

Thus, the Nuer of the south Sudan maintain social distance between the age grades by careful segregation at sacrificial feasts and by a formal deference to elders that precludes open contest. Relations between peers are accordingly all the more affective and friendly, and adolescents regard one another as brothers who can offer each other mutual support and comradeship.

In traditional peasant societies, delayed maturity is institutionalized. Rural Irish society, for example, was characterized by late marriage, a low fertility rate, and long-withheld status, since the basis of economic life was the family farm, which a son inherited only when his father was too old to work it himself. Peer groups consisting of middle-aged men as well as younger men existed to cope with the problem of delayed status and intergenerational tension. Similar peasant societies in Wales, traditional China, French Canada, and central Europe presented a comparable social structure.

In modern advanced industrial societies, the adolescent period reveals parallels with peasant and traditional cultures. But the function of youth is much more dynamic and creative. Duty to parents and submission to authority have to be learned and maintained, and once again the peer group acts as a cushion against excessive reaction to frustration and provides a realistic social world of equals in which the adolescent learns how to cooperate with contemporaries in a completely egalitarian way. In the adolescent period, peer groups are at the height of their influence. Young people with weak parental attachments seem at this stage to regard the valuations and approval of their peers as much more important than the judgments of parents, teachers, and the adult world in general.

Adolescent reactions to looming demands of adulthood

Young people react to their experiences at this developmental stage in two very different ways. They tend to be defensive and overreactive to the adult world, which they perceive as both frustrating and threatening. At the same time, rising anxieties about their own competence may drive them to regress to childhood for temporary comfort and security. Both reactions can bring them into conflict with the representatives of authority and the adult world. Boisterous high spirits, rough playfulness, and the flouting of prohibitions are all in evidence during this stage, and thumbing a nose at older people seems to be something of a psychological necessity. At such times and, more especially, during skirmishes with authority figures, the very fact of being an adolescent provides a symbol of collective identification, enabling youth to see itself as good in and for itself over and against other age groups in the community. The organizers of Fascist youth in Italy and of Nazi youth in Germany before World War II exploited this aspect of adolescence and endeavoured to alienate children from parents—with devastating success.

Youth groups, whether formal or spontaneous in origin, whether officially sponsored or made up of groups of friends and street mates, tend to die away and lose their appeal after the transitional phase has passed. Youth cultures with their own style of clothing, behavioral standards, and characteristic drugs and music are obviously only temporary substitutes for maturer and delayed satisfactions.

The creative and dynamic aspects of the role of youth in modern industrial societies—and, indeed, in almost all societies that are in a state of conscious transition from older to more westernized patterns and life-styles—are becoming more and more apparent during the closing decades of the 20th century. In broad terms this relates to ways in which, during the teenage period, growth is deliberately encouraged, experimentation fostered, and criticism stimulated. The role of the university is here of paramount importance, especially new universities that have been founded on modern liberal ideas and that encourage the study of politics and sociology.

Political impact of youth activists

At a time of accelerated social and technological change this creative aspect of the adolescent role is of great social significance. It was, for example, the zeal and enthusiasm of the young Red Guards in China, in which country youths probably outnumber adults, that made Mao Tse-tung's cultural revolution possible. Further, the rebellion of left-wing students in Ceylon during 1971, although put down, seems to have forced the government into more obviously Socialist policies than they had hitherto been pursuing. In the early 1970s it was estimated that half the population of Africa was below the age of 20. Thus, it is clear that in such developing countries as the new African states political influence and power to shape the future destiny of their country will, by sheer weight of numbers, lie mainly with the young.

Student demonstrations and campus rebellions that have erupted in the United States, France, Germany, Great Britain, and other countries reveal that even in very affluent societies young people still possess revolutionary and reformist zeal and are prepared to insist on the right to have a say in the organization of their own societies. At the same time, however, it is necessary to point out that such politically active young people are probably still in a minority, certainly in the Western world, in which the pressures for social and economic conformity are very persuasive.

Youth and the family. The family is clearly a most powerful socializing institution, and, because of its primacy, it probably has a greater long-term influence on a child's personality and future life than any other group. The quality of emotional relationships within the family and the values and attitudes that are either encouraged or discouraged by parents seem to differ according to social class and status. In Western societies, the children of middle class parents are expected to behave in a way that is appropriate to the family's status and their own future roles. They are required to work hard at school, to control their aggressiveness, to organize their lives in terms of long-term goals and deferred rewards, to be cooperative with authority, and to work and play harmoniously with their contemporaries—without, however, sacrificing individual achievement or ultimate competitiveness.

Working class youngsters, on the other hand, are more likely, except for the socially ambitious, to be brought up in a much less controlled manner. Discipline will be more erratic and less consistent. Future ambitions do not go much beyond the desire to obtain a steady job with good pay, following, if possible, an apprenticeship. Spontaneity and aggressiveness of a physical nature are widely permitted and expected. All this is in contrast to the more rationally ordered, emotionally inhibited, and occupationally oriented upbringing of middle class culture. This is one reason why casual delinquency is a common feature of lower working class life and why the destructiveness and aggression of middle class pupils and students is frequently channelled into organized pranks.

Girls in lower class social strata are more home centred and less free ranging than their brothers. Domesticity comes to them at an early age, and there is seldom any idea of a lifetime's occupation outside the combined role of wife and mother. Middle class girls, however, who have usually enjoyed as good an education as their brothers, are more likely to think in terms of a job or a profession in which even during marriage they may retain an interest and to which they may return when their children are no longer dependent.

Working class children leave school much earlier than middle class children, although the gap is being narrowed in more affluent societies, which can afford a longer period of formal education and dependency than can poorer countries. It has further been noted that access to higher education is easier for middle class than for working class young people. In Great Britain, of the 10 or 11 percent of the school population who go on to university education, rather less than 30 percent are working class in origin, out of a total population that probably contains about 70 percent who are working class in socioeconomic terms. West Germany appears to do much worse, with only about 5 percent of its graduates coming from other than middle class backgrounds. In the United States, there is a much greater degree of equality of opportunity for members of all social classes to obtain admission to higher education than is generally the case in most European countries.

Contrasting parental roles in the socialization process

Parental involvement in the training of children varies widely. In more simple communities the association is often very close, the child self-consciously accompanying the appropriate parent as he goes about his daily business so as to be able to make a complete identification with his own future role. A deeply traditional society such as India contrasts with modern American culture in the degree of emphasis given to parental significance. In the United States a child is taught to relate to contemporaries and peers from quite an early age and to base behaviour as much upon their views and attitudes as upon what parents say and advise. In India, on the contrary, children seem to be brought up to be highly receptive to parental sway, the father particularly serving as a pious model for filial emulation and as a symbol of traditional authority. Such a family structure makes for great social stability and for continuity of behaviour between the generations. It looks backward as much as forward and is highly resistant to social change.

A wide variety of relationships between parents and children exists throughout the world. The link, however, is obviously still of enormous importance, even in communities where peer groups are most vital. Postwar Japan is a good example of the tenacious way in which family life, although shaken, is not completely disrupted even by national catastrophe. In spite of the gigantic emotional and social upheaval following military defeat, national humiliation, and the drastic diminution of the authority of the semi-divine emperor, the stability of family life seems to have remained unbroken at the personal and emotional level, even though growing political and ideological differences have weakened the traditional cultural continuity of the generations.

Youth and education. It is the function of the school and of organized education to carry on and extend the socialization process begun in the family. In advanced Western societies, education is concerned mainly with maintaining the existing culture, supporting the status quo, and promoting a limited amount of social change and social mobility for its more successful participants. In younger countries, however, the role of education is often to stimulate creative dissatisfaction among more intelligent youth, thus making the emergence of a new elite possible. Such new elites are usually urban rather than rural in origin and, as in Latin America, tend to be recruited from the offspring of already existing bureaucratic and governmental elites.

It is clearly understood in all societies that education holds the key to individual future success, and in societies that are already stratified along social class lines it is inevitable that schools will reflect existing differentials. In the United States, for example, professional employees have on average twice as long a formal education as farm labourers and considerably more years at school than factory workers. By and large, although the middle class is permeable and capable of increasing its numbers by recruitment upward from lower status groups, it retains its elitist position by its strategic grasp on educational institutions.

The elites thus fostered by the educational system and the competitive strains experienced in school make a substantial contribution to the creation of many and varied youthful maladjustments. The failure of traditional educational institutions to care for the whole child, the inability to educate for emotional as well as for intellectual growth, the emphasis on self-interest and paper qualifications, all have contributed to the emergence of many of the antisocial types of behaviour to be discussed later in this article.

Youth and work. Work is obviously one of the most important formative elements in the developing self-image of the adolescent. The occupational role links the individual closely to people of other ages and ranks and, at the same time, anchors him firmly in the economic life of the community. It gives him deeper social roots, enhances future prospects and opportunities, and, in theory at least, should nourish feelings of dignity and usefulness. It is through employment, moreover, that the young person achieves financial independence and is thereby enabled to marry and set up his own home. As already indicated, male and female attitudes to employment are different in most cultures because of the differential roles males and females are called on to play in regard to homemaking and child rearing. But in all societies a failure at the work level for a young man means a failure at the economic level, too, and this in turn almost inevitably produces feelings of inadequacy and insecurity that impede the individual's growth to full maturity and satisfactory citizenship. Many kinds of psychological problems derive from a boy's inability to achieve his occupational and work role, and many social problems, such as vandalism, stealing, and gang warfare, are directly attributable to youths' feelings of worthlessness and resentment at being excluded from the material achievements and rewards of ordinary life. Much juvenile crime in modern industrial societies has been traced to economic frustration and consequent status frustration among socially disadvantaged youngsters who have been obliged to live all their lives in substandard conditions and attend low-grade schools.

Consequences of failure in the world of work

Since work is so important for the economic health of both society and individual, it is remarkable how little thought and care has been given to this stage of a youth's life. Even in the most advanced industrial societies there has often been a serious lack of continuity between school and work. Vocational education as compared with bookish learning has often been despised by teachers. There has also been considerable disdain for manual work. Communist countries have done their best to redress this adverse bias and have deliberately extolled the virtues of physical work. Students in Communist states have often been obliged, for example, to do a stint of manual work as a necessary part of their preparation for citizenship, and in modern China pupils spend part of their time in workshops making serviceable articles. Leisure time, also, is partly devoted to "socially useful labour" of various kinds. In this way every child is made community minded and encouraged to see the virtue of manual work.

In the non-Communist Western countries much more is left to chance, to market demand, and to personal and family choice. Middle class children are on the whole very carefully trained and initiated into white-collar jobs and the professions. But the majority of children are still obliged to leave school at a time when their physical, mental, and emotional growth is unfinished, and they are plunged headfirst into adult-dominated working life with little or no skilled help to enable them to adjust to the new and sometimes disturbing experience. The divorce between school and work is much too complete. Attempts to provide vocational guidance undoubtedly help, as does the allocation of certain teachers to careers advice at the school level.

No amount of personal counselling, however, can overcome the major problem, which, in the advanced industrial capitalist societies, is primarily one of a growing lack of employment. Periods of economic difficulty and policies of deflation make the young unskilled worker, especially school leavers, exposed to the humiliation of unemployment and reliance on public welfare. Even those lucky enough to have secured apprenticeships cannot be guaranteed continuing employment during times of economic stress. As with other elements in the labour force, moreover, the young are increasingly exposed to the risk of redundancy—in spite of their training and skill—as technological changes cause the labour force in particular industries to shrink. In Great Britain in 1971, with the total of unemployed running near the 1,000,000 mark, thousands of school leavers and even hundreds of university graduates found themselves unable to get work. The government was forced to contemplate taking the unprecedented step of recruiting as many of them as possible for community projects designed to improve the environment. Such measures, however, can be regarded only as temporary palliatives to a widespread problem that has hit even the richest of all industrial societies, the United States. Despite many millions of dollars being invested in various urban aid programs and ameliorative schemes designed to promote better attitudes to education and to prevent juvenile delinquency, teenagers, black teenagers most especially, seem to be chronically exposed to long-term unemployment. The community simply has no use for their services, and serious social commentators have suggested that as many as a third of all young blacks leaving schools in American cities are facing a hopeless future.

The burgeoning problem of youth unemployment

Yet, while the most disadvantaged youngsters face the prospect of long-term uselessness, other working class youths have in recent years, in some manufacturing trades especially, been enjoying a higher wage increase than any other section of the working population. Such contrasts are hard to reconcile with the ideology of the welfare state and the caring community.

Youth and the economy. Most modern industrial societies have enjoyed increasing affluence as scientific, technological, and organizational skills have been further developed and exploited. In advanced societies the young, whether as workers or dependents, have received a generous share of the general prosperity. Their economic importance, however, as contrasted with the 19th century, has been probably as much as consumers as producers of goods and services. The United States economy since the 1960s has been able for the first time to sustain a large leisure class of youngsters not in the labour force but yet consumers on a considerable scale or, if employed, free to spend their earnings as they please. In western Europe, too, a similarly rising tide of affluence has swept teenagers into prominence as a group well worth the serious attention of manufacturers. There are now a whole variety of goods and articles that are especially designed for the teenage market and for that age group alone. The clothing trade, the soft-drink industry, and much of the entertainment industry rely very heavily on the patronage of the young. Teenagers also spend a fair proportion of their cash on food, alcohol and tobacco, cosmetics, holidays, and gasoline.

Growth of the youth market

The rise of the teenage market was originally a direct result of the better job and pay prospects of young workers in the period following the end of World War II. It had a clear proletarian base and was firmly linked to the 15-to-25 age group. Middle class youth, for the most part still held back in full-time education, had to rely on parental generosity, but figures for Great Britain show that middle class boys tended in the event to have almost as much spending power as their working class counterparts. Most of the money was spent in similar ways; boys and girls from middle class backgrounds attending selective and privileged schools tended to buy the same records, wear the same style clothes in their leisure hours, and enjoy the same amusements. A growing classlessness among youth arising from sharing a common teenage culture has been observed and commented on. At college and university level, students seem to want to look like young workers, and the almost universal popularity of jeans among all kinds of young people, male and female, is a sociological phenomenon of great interest, witnessing not only to egalitarianism but also to the unisexual trends in contemporary youth culture. The old distinctions between the behaviour and attitudes of the sexes have almost everywhere become blurred, and girls, perhaps for the first time in history, have openly become pacesetters in certain consumption spheres.

Teenage culture, with its emphasis on youthfulness, novelty, daring inventiveness, sexuality, tolerance, and gregariousness, does not exist either at its more commercial level or in its psychological depths apart from the wider society. It satisfies the young who subscribe to its values, and in so doing it feeds the adult economy with profits. Some social commentators see youth as typifying a postulated future leisure-orientated culture—the culture of a time when working hours will be drastically reduced and when more and more people will have to find deep satisfactions in leisure-time activities.

Youth and sexual behaviour. The growth of permissiveness and the rejection of older, more puritanical attitudes toward work and sexual behaviour is an aspect of a supposed widespread movement toward the creation of a much freer and more creative way of life that certain self-conscious sections of the youth believe themselves to be pioneering. The growth of student power on campuses and the movement toward "pupil power" at secondary schools, as well as the advent of what has been described as "playpower"—a general philosophy of pleasure-seeking and self-indulgence which seeks to overthrow established authorities—are now almost worldwide phenomena.

In the years following the conclusion of World War II great changes have affected sexual behaviour. What has characterized Western societies is an increasing permissiveness among the young, especially among girls. More widely available and improved contraceptive techniques have combined to reinforce the pleasure ethic, which defines chastity as unnatural and regards full sexual experience outside marriage as the norm. Many young people clearly accept such a norm and openly practice what in an earlier generation would have been condemned as promiscuity. But there are great variations between individuals and social groups, and in this respect, although the general approach to sexuality is much franker and more guilt-free, it is still possible to observe differences related to social class, religious affiliation, and educational level. Young people from strong religious backgrounds, for example, are still more likely to value virginity very highly. It is very clear, however, that in the sphere of intersexual relations there is widespread acceptance of equality and that it is no longer solely the male prerogative to initiate activity nor is it a question of the girl tamely submitting to the masculine urge. Both the pleasure and the affection of sexual relationships tend more to be based on complete mutuality and partnership.

Findings of sex surveys

It is exceptionally difficult to obtain reliable information in so sensitive an area as sexual behaviour. It is necessary to rely to a large extent on research findings that may be out of date or, as was the case in the Kinsey inquiry conducted in the United States in the early 1950s, based on a statistically imperfect sample. The Kinsey report indicated that sexual activity during adolescence was at its height among American boys between the ages of 17 and 20 and that lower class boys were much more active than their college-going contemporaries. Girls, by contrast, showed much less activity up to the age of 15, when almost all the boys had achieved orgasm with or without a partner. Masturbation was widely practiced and homosexual experience by no means rare.

A later English study, based on evidence provided by 1,873 young people between the ages of 15 and 19, presents a more balanced picture, which, while confirming that the momentous years for sexual growth are indeed the middle and late teens, tends to undermine the view that teenagers' sexual ethics and behaviour have radically changed in recent years. The study estimated that 34 percent of the boys in the sample and 17 percent of the girls had had sexual experiences by the age of 18, although this in many cases fell short of full intercourse. The girls in this study seemed to start their sexually active lives before their brothers but at a later date were overtaken and passed by them. The boys appeared to be rather more promiscuous and sexually adventurous than the girls. The great majority of both sexes, however, seemed to have survived the difficulties of adolescence without serious sexual problems or breakdowns arising. On the other hand, the casualties, although in a minority, cannot be lightly dismissed. Illegitimate birth rates for girls between the ages of 11 and 16 have risen during recent years. So, too, has the incidence of youthful abortion in those countries that have legalized it and that keep statistical records. Growing permissiveness is also reflected in the rising incidence of venereal disease among the teenage population of Europe and other areas. In spite of widespread sex education—in Sweden, for example—an increasing number of youngsters contract diseases that seem to have increasing resistance to antibiotic treatment.

There is, however, no reason to think that the adolescent age group is behaving in any society very differently from older sections of the population. Marriage and family life, furthermore, are still regarded by most youngsters with extreme favour. But there is a shift of emphasis, and a new pattern of relationships has emerged. This may be summarized as an acceptance of individuality, greater tolerance for other people's behaviour, and genuine desire for equality and mutuality in all things. Tolerance and unwillingness to condemn those who are different in colour, race, or culture forms the foundation for a new ethic among young people almost everywhere. As allegiance to formal religion and acceptance of established dogmas weaken, ethical sensitivity and commitment reappear in the rejection of militarism, the campaign against all kinds of discrimination, and a repudiation of the grosser manifestations of modern materialism.

Factors behind changing attitudes to sexual relations

Such changes as have been outlined above arise from a variety of causes. One is the growing numerical and economic strength of youth in modern society generally. Much, too, must be attributed to new patterns of child rearing and new ideas in the child-care field that have encouraged the more child-centred home and school. Emphasis in modern education is increasingly on the emotional and creative aspects of the curriculum and less and less on the hard slog of rote learning and reliance upon discipline imposed in an arbitrary fashion by adults from above. Principles of participation, discussion, and rational explanation and individual decision making, which formerly were only to be found operating in experimental schools, are becoming much more widely accepted in general education at both state and private levels.

Varieties of youth. Much of what has been said above can now be summarized in a tentative typology of modern youth. The foregoing analysis seems to lead to four principal illustrative profiles that highlight the trends and regularities in youthful behaviour in many different countries.

The upper class conformist. The apparently well-adjusted, mainly conformist, upper and middle class youth may be said to typify success within a specific cultural system. They have been pupils at leading schools. They have had the support of parents throughout the growing-up process. On the whole they have identified with their parents' outlook and way of life, obeyed the codes of their teachers, passed the necessary examinations, acquired the diplomas and certificates that will admit them into the business and commercial world at a reasonably well-paid level or that will give them access to high-status professions. Some have come from lower class families and have risen in the socioeconomic hierarchy by their hard work and natural ability, but the great majority have been born into the middle and upper classes and have been socialized in the appropriate norms. Their political views are more likely than not to be conservative or mildly liberal, since these are the values they have been brought up to respect. Their future prosperity, moreover, depends very largely upon the stability and perpetuation of the existing system. In some countries, this class of youth will have been educated in elitist institutions such as the French *lycées,* which set out to inculcate qualities of leadership in the professions, in the military sphere, or in political and public life. As a group, these young people are a blend of established privilege with meritocratic achievement. They tend to deprecate excessive emotions of any kind and to be against histrionic gestures. The males are inclined to treat women as idealized inferiors and on the whole are well-content with a social, economic, and political world that gives them a monopoly of opportunity, authority, and autonomy. The psychological stresses to which they will be exposed will arise from anxiety about living up to family expectations and doing at least as well in life as their parents. Some of the more sensitive ones may also feel themselves constantly challenged by the ideals of political democracy to defend themselves against the accusation of elitism and class privilege.

Emotional attitudes of the upper class conformist

The working class conformist. The second group is lower in social status than the group delineated above. It comprises the well-adjusted, conformist, lower middle class and upper working class youth who constitute a very large proportion of the age group in modern industrial societies. They will have been brought up with a comparatively good health service and in a bland social climate that stresses the community's responsibility for its children. At school they will have enjoyed a fair general standard of education but will probably have been particularly interested in practical and technical subjects with a strong vocational usefulness. When they leave school they are likely to seek employment in skilled manual work so that, financially at least, they will receive high wages at an early age and so come to have a considerable stake in the political stability of modern industrial society. They work hard and expect good pay. They see money as a means to enjoy their immediate leisure in the form of stylish clothes, cars, drink, music, sex, and all the sensuous good things of life. When they get a little older and consider it time to settle down, they will want a modern house or flat in a new town or in a suburb. They enjoy and in the future will expect to go on enjoying a high standard of physical comfort, and their lives will become increasingly home centred. They are often politically inert and, in Great Britain, unlike their Socialist grandparents, do not see any need to go on fighting the old class war. Their links with the old working class culture and traditions are hence growing weaker. They are the new, almost classless, self-confident young people of an affluent age, who are unlikely to show radical political attitudes. On the contrary, they are as likely as not to regard those members of the same social class who are unemployed or in much lower status and less skilled work as constituting, in time of economic depression, a threat to their own recently won high standard of living.

Political uninterest of affluent working class youth

Nonconformists. Most societies other than primitive cultures have a quota of nonconformists, some of whom may be regarded as moral deviants but others of whom simply refuse or are unable to accept the norms of the majority of the community. Modern European and North American societies have been producing an abundance of such types since World War II. Even the Soviet Union, during the 1950s, produced its own version of rebellious unconventional youth, the *stilyagi,* who, like the British Teddy boys and the German leather-jacketed *Halbstarken,* were in their day conspicuous for both their clothing and behaviour.

It is useful to distinguish lower class anti-authority delinquents from their more morally committed and more intellectually oriented contemporaries. Underprivileged minority groups of working class youth are in the main merely reactive against their lower status and depressed opportunities. In a sense, they are hitting back at those who have denigrated them and, by limiting their opportunity for development and mobility, have condemned them to second-class citizenship. Dropouts from the school system are the victims of chronic underprivilege, grinding cultural poverty, and social disadvantage. In former days they would have perhaps submitted to the dead-end jobs that were all that would have been left for them. But amid surrounding affluence they have reacted more aggressively. In a study of American youth made in the late 1950s, the American sociologist Robert Havighurst and his colleagues estimated that something like 29 percent of the rising generation was composed of drifters and socially alienated youngsters coming from all social classes, although predominantly from the lower middle and working class group. The actual proportion of such disaffected youth in contemporary Western society is difficult to estimate. All that is certain is that the number is large and perhaps increasing. The violence of such groups as the Hell's Angels, who are to be found in many Western countries, is deeply disturbing. Whether in Copenhagen or California, their deportment is that of aggressive dropouts from conventional society. Their groups, or chapters, are growing, and in England they appear to be the natural descendants of the motor-cycling Rockers who in the mid-1960s fought the Mods (fashionably dressed teenagers) in various public places. Their gang warfare, remotely suggestive of classical Chicago and New York models, with female followers emotionally identifying with them, presents a strangely melodramatic image of confused adolescents in a society that has no place for them and in which they presumably feel psychologically obliged to hate and to fight back against everyone who is not a member of their own group. (J.B.M./Ed.)

AGE DIFFERENTIATION: THE AGED

The position of older people in modern industrial society must be interpreted in light of the fact that many of them survive beyond the age at which they would perform or be rewarded in major social roles: worker, parent of dependent child, or spouse. By contrast, in most primitive and agrarian societies, old age (at least up to senility) has been associated with some special characteristic regarded as an asset in the respective culture, such as usefulness in the performance of chores, skill in dancing or storytelling, control of property rights, power in the family, seasoned experience, or (especially in preliterate societies) extensive knowledge.

The status of the aged in the economically advanced nations today derives from broad changes over the past century in the entire society and its culture. Demographic change has altered the age composition of the overall population, and for the individual, the expectation of life has been increasing as mortality rates decline. Economic development has been accompanied by decreasing proportions of older men in the labour force, posing problems of income maintenance for those no longer economically active. Levels of educational attainment have been rising. Urbanization has transformed the residential setting. The structure of the family has altered. Thus the social context of the aged today differs markedly from the earlier situation when today's older people were reared and almost certainly from the projected future when today's younger people will have grown old.

The social context. *The demographic base.* In Western nations generally, there has been a marked rise in the number of older people, paralleling the expansion of the population as a whole. Over the first half of the 20th century the annual rate of increase in the age category of 65 and over has ranged from about 1 percent in France or Sweden to nearly 3 percent in Canada or the United States.

Numbers of older people have been increasing even faster than the total population. Thus, in Sweden and France the proportion of the total population aged 65 and over had already reached 8 percent by the turn of the century; in the United States between 1850 and 1960, this proportion rose from 3 percent to 9 percent. This long-term tendency for the older age groups not only to grow larger but to grow faster than the younger segments of the population is characteristic of industrialized societies generally; by contrast, in most of the less developed countries of Asia, Africa, and Latin America, the proportions of persons 65 and over still fall below 5 percent. These changing proportions of old people within the population have resulted primarily from declines in fertility rates, which have reduced the accretions to the younger age categories, and not, as commonly supposed, from the declines in mortality. The improvements in reducing mortality have, in the past, had their effect most directly on infants and small children (though any future benefits in Western countries are likely to accrue at the later ages). Although the increasing proportions of survivors may eventually swell the ranks of old people, they have in the meantime added to the numbers at the young ages, not only initially when they were young themselves but later on when they became parents.

Increases in females

Not only have the number and proportion of older people been changing in the West but among older people the proportion of females has been rising. Except in some underdeveloped areas, male mortality rates are generally higher at all ages than corresponding rates for females (even for some animal species). Hence there is typically an excess of females over males among the aged. In the category of people age 75 and over in the United States, there are only about 73 men for every 100 women, a discrepancy that has been widening.

For individuals within a society, the secular declines in mortality (and the consequent increase in the expectation of life) have greatly improved the chances of surviving into old age. The change is not in the *span* of human life (the ultimate length of life attainable by a member of the human species) but in the proportion of people enduring into the higher ages of this span. Thus the average lifetime in ancient Rome or medieval Europe is estimated at some 20 to 30 years. In several Western countries, a man born in the middle of the last century could look forward to four decades of life; today a man can look forward to living well into his 60s, a woman to living into her 70s. Such demographic changes characterizing Western civilization appear to be unparalleled in human history. They have profound societal implications—for example, for government involvement in programs for older people, for the increasing supply of mature individuals available for the labour force, or for the structure of the family because all the family members, but particularly the females, tend to live longer.

Economic status. The age structure of society is closely linked to the economy. In this realm older people tend, on the average, to be comparatively disadvantaged, although individual variations are great. Their low status is indexed by their comparatively low rates of participation in the labour force and by their low average incomes.

Decline in employment

There has been a steady decline (apart from short-term fluctuations attendant upon mobilization for war) over several decades in the proportion of men 65 and over who are in the labour force in most industrialized countries (in contrast to agricultural countries).

No firm explanations for the long-term decline in the employment of older men have been satisfactorily demonstrated. Extensive research on past trends and on age patterns today point, however, to a variety of associated factors and facilitating mechanisms, such as the possibility that technology may have advanced to the point of a sufficiency of manpower to meet existing demands for goods and services without the participation of older workers; the decline of agriculture and of self-employment, in which the aged are free to continue work or to taper off from it gradually; the steady rise in the educational level of the labour force, which appears to favour the better educated and the more recently trained younger workers; and the extension of public and private pension plans, which afford alternative sources of income to increasing proportions of older workers and which often specify the age for mandatory retirement.

The striking discrepancy in rates of labour-force participation between older and younger men (most younger men are in the labour force) has differential consequences for the age strata, affecting not only their respective involvement in the goal-directed activities of the society but also their financial status and the time available for their leisure pursuits.

Decline in income

In regard to financial status, people over 60, and even those in their 50s, have markedly lower median incomes than people in their middle years, even after various adjustments are made (for example, for family size). The substantial data available for the United States, for instance, show that, over time, there have been absolute increases in the income and the purchasing power of older people, with a declining share from earnings and a rising share from social security and other retirement benefits (often paid in fixed dollars); however, older people continue to lag behind the young.

To be sure, older people on the average, with their years of asset accumulation mainly behind them, own more and owe less than younger people. Among elderly married couples, substantial proportions (the majority in the United States) own their homes. But assets, which are very unequally distributed, tend to be correlated with income, so that those families with the lowest incomes are also the least likely to have any assets.

What kind of life does the older person's income support? Here again data available for the United States show that he tends to live closely within the limits of his income, spending less than younger age groups for most budgetary items, including leisure items, but spending a larger share for health expenditures when his government has no comprehensive program of medical insurance. His spending is restricted both by age-associated factors (such as reductions in family size after the children leave home) and by his comparatively low income; the income of many older people, when judged by arbitrary standards set by economists, is not adequate to meet their needs.

Education. The age strata of the society are distinguished from one another not only in labour-force participation and in income but also in educational background. Older people have less formal education and less recent education than younger people. Among persons 75 and over in the United States, where the century-long spread of formal schooling accentuates the age differences, more than 70 percent have had only eight years of schooling or less, compared with less than 20 percent of those aged 25 to 29. Moreover, although the average educational attainment of older people is rising, the lag of old behind young persists.

Lower levels of education

These striking age differences in education have widespread ramifications throughout society. Since comparatively few older people are well educated, few of them possess those characteristics, typically valued in Western culture, that are associated with high education. Thus, the less educated majority of older people are less likely than the few who are well educated to remain active; they are more likely to retire. They are also less likely to belong to voluntary associations or to read or to want to learn more. They have lower incomes. They are often less happy and generally less optimistic about the future, but at the same time, they are less introspective and less ready to doubt their own adequacy as spouses or parents. They are more negative in their view of death and think about it more, although fewer of them are disposed to plan for it.

Age categories in the population differ not only in level but also in recency of education. Thus, the content of knowledge and attitudes acquired vary by cohort. The information imparted to doctors or engineers during their training, for instance, differentiates sharply between the backgrounds of old and young. Yet within each cohort, there is a certain age homogeneity in values and beliefs (about what is good, beautiful, or true) among individuals who were educated at the same point in history.

It is interesting to note that one reason education can make such enormous differences among the age strata in modern society is that formal schooling, rather than being spread over the life course, is almost completely concentrated in the early years of life. A person enters adulthood with a fixed educational background that functions sociologically in almost the same way as such inborn characteristics as skin colour or sex. Adult education (a channel that might close part of the age gap) is developing, but it still reaches only small proportions of people in their middle or later years.

Residential setting. Historic changes in population distribution, in community planning, and in the designing and marketing of houses affect the physical and social environment of the older person; influencing the range of his human contacts, his day-to-day activities, and the community services and facilities available to him. In the modernized nations persons who are now old have, to a great extent, been swept along in the massive movements of the population from agricultural to industrial areas, from farms to cities or suburbs. Many of the elderly, however, have been left behind, as age-mates die off and younger generations move away. Moreover, the houses in which most older people live were built in an earlier era. As their houses have aged, the inhabitants themselves have moved along the life course toward increased leisure in retirement, diminished families, and reduced incomes.

In general, however, most older people do not avail themselves of the freedom from occupational and family commitments to move away from their former homes, a fact that gives some indication of the meaning of the older person's home and the factors mediating his relation to his environment. Rates of moving are generally highest among people in the youngest age groups. Moreover, when older people do move they are more likely to move within the confines of their immediate locale than to change their community setting. It thus appears that, basically, the individual's ties to his place of residence become stronger as he grows older. And older people who own their homes, or those who have strong social connections with the neighbourhood, are the least willing to move.

Family status. The social position of the aged is importantly tied to the structure of the family, which has been gradually changing in Western society over the past century. Although this change is usually described as the isolation of the nuclear family (*i.e.,* parents and their children), perhaps it may be more accurately described as the subdivision into two or even three generations of distinct nuclear families: the young couple with their dependent children, the middle-aged parents, and the aged generation of grandparents.

Independence of the aged

As life expectancy has risen, husbands and wives have become increasingly likely to survive together into old age; and widowhood, even if not curtailed in length, has been postponed. Today, in such countries as the United States, Great Britain, or Denmark, in the 65-and-over age range, most men (about two-thirds), but a minority of women (about one-third), are living with their spouses. Moreover, most of those 65 and over, the married and even the widowed, maintain independent households. In the United States this way of living has been increasing, as decreasing proportions of older people live in the same household with their children. Even when households are shared today, the older person is, more often than not, himself the head of the house or the host, not a guest or subordinate in the home of his child. Living entirely alone is a frequent pattern among older people. This is especially true among women, more than one-fourth of whom (mainly widows) live alone. Moreover, the norms appear to favour separate households for older people, except for those who are ill.

Thus a new type of nuclear family, the independent, older family (often a single-person "family") is developing, with little-explored consequences for the older person or for the society as a whole.

The aged individual. The social aspects of old age cannot be understood without reference to the physiological processes that distinguish old people from the young. The age structure of the society is fundamentally affected by biology, not only because there is the succession of births and deaths but also because individuals at different stages of life show important differences in both physical structure and function.

With age, there are declines in the number and quality of vital cells and a decreased ability of the organism to adapt to changes in the environment. Associated with such age-related differences is the fact that the health of older people is generally poorer than that of the young. While older people have fewer acute illnesses, they are more subject to chronic conditions: such impairments as failing vision or hearing or such ailments as rheumatism and arthritis, heart disease, and high blood pressure. In the United States, for example, some four out of five persons 65 and over have at least one chronic condition. Thus there are increases by age in physician visits, in the number and duration of hospital stays, and in the number of days spent in restricted activity or in bed. Nevertheless, only a small proportion of old people are so severely handicapped that they cannot carry on their major activity, whether it be a job outside the home or housework.

Behavioral changes

Physical changes associated with age are related not only to health but also to behaviour, though the causal linkages seem far from clear. When compared with younger people, older people are more likely to show deficits in sensory and perceptual skills; in muscular strength; in the ability to react quickly; in complex sensorimotor coordination (such as that required in driving a car); and, most important of all, in certain forms of memory, in learning and in various aspects of intellectual functioning. There are also age-related decrements for both men and women in sexual activity, though sexual capacity persists at least into the 70s or 80s.

Of course, not all the differences between old and young are necessarily reflective of a biological, or even a social, process of aging. It would be fallacious, for example, to attribute entirely to old age the comparatively poor performance of older people on intelligence tests. In cross-section studies (conducted among several age groups at a single time), intelligence appears to reach a peak among people in their late teens or early 20s and then to decline with age in the older strata. But to infer that intelligence falls off so early in the life course of particular individuals because of age-associated changes (as in general health) is open to serious question, as a few longitudinal studies (which trace individuals over the life course) have begun to suggest. In this instance, the interpretative difficulty seems associated with social change, with the long-term upward trend in education. Intelligence test performance is highly correlated with educational level, and education sets older cohorts sharply below younger cohorts.

Unfortunately, there is little information about trends over time in health, physical functioning, or the associated changes in behaviour. For the population as a whole, to be sure, the rising proportions of older people point to the likelihood of increasing prevalence of the chronic conditions associated with senescence; indeed, there have been increases over the past century in death rates resulting from such illnesses of later life as heart disease or cancer. But it is uncertain whether older individuals today are healthier than older individuals in the past. Medical advance may have improved their physical condition, or it may, by interfering with the principle of survival of the fittest, have yielded a larger but less hardy population.

Whatever the trends may be, the inferior physical state of the older segments of society, in contrast to the younger segments, has both societal antecedents and societal consequences. Health is conditioned by such social factors as standard of living, education, and advances in medicine and public health. In turn, the physical state of older people sets limits to their social adjustment and to the contributions that they can make to society.

How, then, does the older individual respond to a life situation in which he tends to be comparatively disadvantaged both in health and physical fitness and as a member of society? Does he in fact conform to the tragic stereotype of the old person as destitute, ill, facing irreparable losses, no longer integrated into society, and no longer subject to society's controls and sanctions? Is it true, as often supposed, that his feelings are fully reflective of the relatively deprived status of the aged within the society and that he is characterized by a loss of self-esteem, a deprecatory view of his low education, a sense of dejection and despair over his losses, and anxiety about his health, finances, and death? A glance at the available clues shows a picture that is, at least in certain respects, at sharp variance with such a stereotype.

Personality. Old people resemble the rest of the population in many of their personality characteristics, despite certain distinctive types of emotional expressions and modes of adaptation. Thus the aged, in comparison with the young, appear to be more rigid, passive, and introverted, more restrained and cautious, and less oriented to achievement. Age differences in personality and the individual's disposition to organize his behaviour in particular ways cannot be separated, however, from his biological condition and the social roles that he plays. And there is little evidence to indicate how many of the unique aspects of the personality of the older individual are a result of accumulated experience, relative lack of education, inexperience with personality tests, weakened sensory and motor skills, or limited opportunities to test out his ideas on other people.

Degree of neurosis, psychosis, and deviance

Disturbances of thought, affect, or behaviour afflict, in varying degrees, somewhat fewer than 10 percent of the people 65 and over in several localities studied in Europe and the United States. The milder mental disorders (neuroses) do not appear to increase by age. But psychoses (and various psychiatric symptoms experienced as bodily illness) are apparently more prevalent among older than among younger people, both because in later life there are certain conditions connected with the degeneration of the central nervous system and because the older the individual is, the greater his chances are of having any chronic or irreversible disease that does not result in early mortality. Although the burden of mentally ill older people upon families and treatment facilities has undoubtedly been rising over time, the rise is due partly, perhaps entirely, to changes in longevity—to the increasing numbers of people who survive into old age.

There are differences by age in various types of deviant behaviour. For example, crime rates, as observed in disparate times and places, show a consistent decline as age advances beyond adolescence. Moreover, life-course analyses demonstrate a curtailment in criminal behaviour with age even among those who have previously engaged in such behaviour. Thus, in the comparatively rare instances when older people commit criminal acts, these are ordinarily relatively harmless offenses rather than aggressive outbursts against society or acts of violence against other persons. In contrast, suicide is most pronounced in the older age categories, a pattern observed in many countries in both the 19th and the 20th centuries.

Social roles. The older the individual is (beyond the middle years), the fewer social roles he plays on the average (as he retires from work, his children leave home, and his age peers die off) and the fewer and less varied are his contacts with other people and with the environment generally. Individuals are not all alike, however, and social activity is greatest among those aged who are in good health or who come from higher rather than lower socioeconomic backgrounds. Moreover, much depends upon the particular exigencies that permit some older individuals, but not others, to retain their major familial and occupational roles.

Work and retirement

The aging individual faces the culmination of his occupational role and a final period of retirement. For older men today, the modal role is retirement, not work. If they grew up in an era when retirement was not widely institutionalized, they now confront a world increasingly populated by age peers who are no longer full-time members of the labour force.

Those men and women who do remain in the labour force during their later years are not making generally inferior contributions, despite their frequently poorer performance under laboratory conditions. Studies under actual working conditions show older workers performing as well as younger workers, if not better, by most, though not all, measures. (Of course, such age patterns of actual performance are traceable in part to labour-market conditions and to selective processes that allow the retention of the more competent older workers in their jobs.) That there is no necessary decline in creative productivity with aging is evidenced through analyses of the published biographies of contributors to numerous scholarly, scientific, and artistic fields (though widely quoted earlier studies had propounded somewhat misleading contrary findings).

In addition to maintaining their productivity, many older people also take a generally positive view of their occupational role. As compared with the young, older people still in the labour force tend to be more strongly committed to the kind of work they do, to adapt better to the job, and to express greater satisfaction with it. Moreover, despite their inferior education and the lowered aspirations attendant upon age, older workers do not fall very far short of the young in their sense of occupational adequacy or in the emphasis they place on the intrinsic satisfactions of the work itself.

Once older people have retired, they seem on the whole to accept the new role of retired person, although many say that they regret losing their work associates, the feeling of usefulness, the work itself, and the associated earnings. Retirement per se does not appear to have a deleterious effect on health nor does it affect social participation in the other roles of the retired person. To be sure, adjustment to life is generally poorer among retired old people than among age peers who are still working, but this results in part from the selective process whereby the healthier and more advantaged oldsters tend to continue working.

Political roles

Older people, even those with little education, are significantly involved in the political system. They are, for example, at least as likely as younger people to vote. They are better represented among the political elite and in leadership positions in many types of decision making. They are more disposed to keep up with the news and with public affairs through the mass media.

In general, today's generation of older people are more conservative than younger people in their political attitudes. There are important exceptions, however, as on issues affecting their own economic self-interest. In the United States, for example, the old have often tended to support extensions of social security or government health insurance but have frequently voted against school bonds. Yet there is little indication whether future generations of old people, subjected to shifting political climates and to rising levels of education, will perhaps adopt a less conservative stance and a generally greater flexibility of attitude.

Religious and other associations

Religion assumes greater importance among older than younger segments of the society, although research (which is less complete than research on the polity) does not show whether religious values become intensified with aging or whether the older cohorts reflect a stronger religious emphasis in their early training. The aged are more caught up than the young in such personal observances as prayer and reading scriptures, and elderly members of religious organizations benefit from a variety of special programs, such as home visits and nursing homes. Church attendance, however, declines in the oldest years.

In addition to religious and political affiliations, modern pluralistic societies offer a variety of voluntary associations as potential links between older people and the larger community—associations such as clubs, lodges, auxiliaries, and the many other formal organizations to which people belong part-time and without pay. To be sure, membership in such associations is less widespread among older people generally than membership in a church or voting in a national election; and average membership rates show a drop in later life after a rise in the middle years, though individuals vary greatly around this average—depending upon their education and income. The special clubs (such as Golden Age Clubs or Senior Centres) established in many Western countries to provide older people with recreational, educational, health, or welfare services seem to attract comparatively few participants.

Family roles

Old people play a variety of family roles, and they seem no less likely than younger people to feel satisfied and adequate both as spouse and as parent. For those older people who have a living spouse, the marital relationship is of central importance, and older couples share many joint activities; but many of the aged, especially women, must adjust to widowhood.

Most old people play the enduring role of parent. While their children in turn form families of their own, new linkages develop with children-in-law; and grandchildren, even great-grandchildren, proliferate. Living separately from children does not mean complete isolation from them. More than 80 percent of older people who have living children live less than an hour away from the nearest child, and a similar proportion see one of their children at least every week. Bonds with siblings or other relatives are often maintained, and a few old people continue in the role of offspring to parents who have survived into very old age.

The ties of the elderly to their relatives and to their children in particular are by no means limited to visiting. There are also widespread exchanges of material support of various amounts and kinds, ranging from financial contributions and care in illness to baby-sitting and help with housework and home repairs. Contrary to the often-held theory of a one-way flow of contributions to older people, the flow of support between aged parents and their adult offspring appears to be two-directional, from parent to child or from child to parent as need and opportunity dictate.

Although close parent–child ties seem important to both older people and their offspring, the offspring feel closer to their new family of procreation than to their family of orientation. Thus the norms of intergenerational independence seem to operate to reduce the emotional dependence of the elderly upon their adult children.

Leisure

Friends and neighbours play an important part in the lives of many older people, often providing help and services as well as informal contact with the world outside the home. They are, however, generally less important to the aged individual than children and other relatives, serving more as a complement than as a substitute for kinship association. Friendships and neighbourly relations, which tend to be maintained well into later life, are most widespread among older people of comparatively high socioeconomic status and among those who have resided for a long time in the same neighbourhood. Older people tend, though by no means exclusively, to have friends who are similar to themselves in status characteristics (notably age) that reflect common experiences or values.

Maturity yields new increments of leisure time, as commitments relax to parental and later to occupational roles. Many and varied pursuits fill these hours of leisure. Visiting with friends and relatives is an important activity. Much time is devoted, for purposes of information as well as entertainment, to watching television and reading the newspaper. Only small minorities, however, spend their leisure in crafts, hobbies, or intellectual and artistic activities; and vacations and outings beyond the home, particularly foreign travel, are restricted to the few. Indeed, some older people, especially the very old and the disadvantaged, often have difficulty finding anything at all to do.

When the leisure of old and young is compared, varied age patterns emerge for different sets of activities. In the United States, for example, certain types of activities (gardening or walking) seem characteristic of older people. Other kinds of activities (swimming, museum attendance, or playing a musical instrument) are widespread among the young but largely unfamiliar to today's older generation—a possible portent of social changes to come. Thus a "mass" culture may exist at the younger age levels that will permeate society only as these cohorts (generations) reach maturity.

Attitudes and satisfactions. Age is related not only to psychological characteristics and to performance in roles but also to the individual's estimate of himself, his attitudes toward the environment, and his feelings of gratification or deprivation.

Relatively positive self-evaluations

Underlying the several dimensions of his personality, a clear sense of his own identity is experienced by the typical older person. Although conceptions of old age are often assumed to be largely pejorative, the modal older person appears to evaluate quite positively such aspects of the self as his moral virtues or the adequacy of his role performances. At the same time, he minimizes in his self-image many of those aspects that are negative, such as his failing health, his personal appearance, his relative lack of education, or the fact that he is old.

If many older people, even though objectively disadvantaged, fail to take deprecatory views of themselves, the reasons are not entirely clear. By virtue of their long years of life, these people have acquired experience, possible wisdom, and the opportunity to come to terms with themselves. Perhaps they judge themselves less with reference to younger people or to themselves at an earlier age than with reference to others who are at least as old as they.

Older people differ sharply from the young in many of their opinions, feelings, and dispositions toward such central aspects of life as health, personal problems, or death. To a greater degree than younger people, they take aches and pains for granted, put relatively little faith in medical science, and feel they understand their own health best—though the frequency of thinking about health, visiting doctors, and taking medicines rises with age. Certain elements in their conception of physical illness appear also in their outlook on mental illness: the sense of inevitability and incurability and the stricture that the individual should look after himself. They are likely to define their personal problems generally as unchangeable and to stress the individual's responsibility, although they seem ready enough to seek assistance in those instances when assistance is deemed appropriate. Death is typically confronted openly as an inexorable fact. Most older people report that they think about death, are willing to discuss it, and have made preparation for it. Few stress otherworldly aspects, either hope of heaven or fear of a last judgment. Indeed, few show marked fear of any sort, expressing the view rather that death is more tragic for the survivor.

Fatalism combined with stress on individual responsibility

Certain recurrent themes are discernible beneath the myriad specific differences, as old and young define and assess their life situations differently and hence are inclined toward differing courses of action. First, older people tend to have less "sense of mastery" over the conditions of their lives than do younger people, considering the world potentially less changeable. Second, older people tend (paradoxically) to stress the "responsibility of the individual" for his own destiny; whereas younger people are more likely to stress environmental influences. Third, in line with such differences in definitions of the situation, old and young tend to favour different "types of approaches" to life situations. Thus older people, defining many of life's ills as inevitable, may be more disposed to seek palliative rather than corrective or preventive treatment. Or again, if older people attribute to the individual the responsibility for cause and cure of his problems, they may be more disposed to lay exhortation or blame upon him than to attack his social or his somatic conditions. Fourth, when older people do become committed to a particular approach, they may be typically as willing and able as the young to *implement* it. Thus much of the often noted lower activity levels of the aged may perhaps be explained through their differing definitions and evaluations of the situation rather than through any age-related tendency toward generalized passivity.

Such differences in the life attitudes of old and young are probably traceable in part to the differences in educational level and to the impact of the aging process itself. In addition, the differences are undoubtedly reflective of the long-term social and cultural change in which each new cohort (generation) has been socialized to new understandings, norms, and habit patterns.

In some general sense, satisfaction with life (happiness, morale, adjustment) seems to diminish with age. This decline, already apparent in early adulthood, is not peculiar to senescence; but it becomes intensified (as both longitudinal and cross-section studies suggest) with age-related deterioration in health, loss of key roles, or reduction of activity. Thus age appears to be associated with a general diminution of the opportunities for happiness.

Nevertheless, when research is focussed, not upon overall satisfaction but upon more specific reactions to particular areas of life, older people appear to differ from younger people in the kinds of gratifications and anxieties experienced and regarded as salient. The typical older person is as likely as the younger person to seem content with his occupational and familial roles and to encounter no greater problems in them. And despite the objective difficulties confronting him, he appears even less likely to worry about his health, for example, or his finances. Thus, the older a person is, the more nearly he seems to have come to terms with many of the specific conditions of his life.

Many studies of life satisfaction have challenged the provocative notion of "disengagement," conceived by some researchers as a mutually satisfying process of withdrawal between the individual and society in preparation for the incapacitating diseases of old age and eventual death. In the main, the evidence lends little support to the theory, although continuing activity is clearly not a *necessary* condition of satisfaction for all older people. For the majority, however, the sense of well-being is associated with high rates of interaction and contact with the environment—just as adjustment is associated also with good health, with favourable occupational and familial circumstances, and with a high sense of self-esteem.

Programs and policies. The modern stereotypical view of old age as a situation of utter and inevitable disadvantage has often overemphasized the problems and weaknesses of the elderly without paying proper attention to their strengths and their potential contributions to society. In fact, there are many areas of independence and an absence of serious physical disability among the majority of today's aged. Moreover, the relative deprivations of the older population as a whole are offset by numerous exceptions, since there are segments of this population who enjoy adequate education or income or who exhibit high levels of adjustment and interaction with their fellows. Such exceptions serve both to deny the intractability of the current problems of the elderly and to suggest areas in which ameliorative effort is likely to produce needed solutions.

Nevertheless, the deprivations remain. The central fact is that a great many older people nowadays are not filling socially valued roles in the economically developed countries. This fact can be punishing to the individual, and for the society it can be a costly waste of human resources. Concern for this fact—a concern that has become intensified with the multiplying proportions of old people in the population and with the recently emerging power of the medical profession to prolong the lives of many dying patients—has led to the formulation of various policies and programs.

Types of programs. The problems of the aged, like those of the poor or the disabled, have been handled through a wide variety of institutions and arrangements at different times and in different places. In most agrarian societies, the aged have found security in a familistic social organization in which food is shared, oldsters often marry younger mates, or progenitors hold sway over descendants. The family has been bulwarked by such other structures as the church or the guilds in medieval Europe, by private philanthropy and the giving of alms as broadly established in many cultures, or by the building of institutions to house the aged (which existed in Europe as early as the 3rd and 4th centuries).

Comparisons between agrarian and industrial societies

Modern industrial societies, with their complex structures and their large proportions of older people, have developed many special forms of adaptation.

The family is still of central importance to the older person's security (statutory responsibility of adult children for their needy parents continues to exist in a number of countries); today's aged couple, however, with their increased years of joint survival and their tendency to live apart from their children, are likely to depend primarily upon each other, turning to children or relatives only under conditions of special need.

Following the breakup of medieval institutions in Europe, the aged are variously assisted by state intervention, private and corporate philanthropy, religious and fraternal organizations, or labour unions; and private enterprise is often engaged in building housing and residential communities for the aged or in arranging for the delivery of health care or other services to older people. In the United States, for instance, many older people have secured financial protection above the government-supported minimum through private pensions, voluntary health insurance, or life insurance benefits to the bereaved spouse. Only small fractions of those 65 and over (less than 5 percent in the United States and less than 4 percent in England and Wales) are residents of old age homes and other institutions.

Public programs and governmental acts to meet the needs of the aged, from the bread and circuses of ancient Rome to the modern welfare state, have had varying histories in different countries throughout the world—depending upon particular social philosophies; upon the crises created by war, plague, or famine; or upon the maladjustments and dislocations created by basic changes in economic, religious, military, and other social institutions. In England, for example, a series of 16th-century enactments, culminating in the Poor Law of 1601, recognized a degree of public responsibility for the aged, along with the sick and the poor, although the administration was left to the local parish. In subsequent years, various legislative and administrative modifications were made in the law, including the reversion (under the Poor Law Amendment Act of 1834) to an austere system of workhouses in which a uniform, stringent discipline was imposed upon the aged, the sick, and the able-bodied poor. Not until 1925 did England introduce social insurance for the aged (followed in the 1940s by the legislation underpinning the full welfare state system), in which contributions are regulated by law and benefits disbursed without a means test. Meanwhile in Germany Bismarck instituted as early as the 1880s a full program of social legislation, including old-age pensions. Austria, the Scandinavian countries, the Low Countries, France, and Italy followed the German model; and by the mid-1960s, well over 100 countries had social security programs. The United States, slow to yield its philosophy of individualism, long clung to vestiges of Elizabethan-type poor laws, not enacting an old-age pension program until 1935, when, as a consequence of the Depression, the Social Security Act was signed. The basic plan has been supplemented by other federal, state, and municipal programs, including Medicare, Medicaid, the Older Americans Act, income-tax exemptions, subsidized housing and liberal financing of mortgages, and old-age assistance as the last line of income defense. These and many other aids and services to older people have alleviated some of the difficulties and set a floor to older people's income, but they have not brought this income up to the level of younger people.

Poor laws and social insurance

Development of the professions. The extension of programs to ameliorate the condition of older people has enlarged the demands upon the practicing professions. The long-term intrusion of old age and its concomitants into the established professions, though often unnoticed, has increased the likelihood that the physician or nurse must treat the chronic ailments and disabilities associated with senescence, that the architect must provide housing for older couples or widows living alone, that the social worker or minister must deal with persons no longer absorbed in occupations or parenthood, and that the lawyer must advise on the estates of persons who can expect to live well beyond the age of retirement.

Professional aid to the aged

In addition to increasing in numbers within the clientele of the professions, older people have created demands that often require new kinds of services or shifts in professional emphasis. Thus educators may be expected to provide continued education or retraining to adults, mass communicators to forge links between old people and society, or public-health experts to expand the delivery of health care to this segment of the community. Architects and city planners must arrange for nearby facilities (such as stores, buses, medical services, churches, or recreational activities) on which older people in failing health and vigour are especially dependent; and, since the elderly seem loath to move away from familiar surroundings, homes may be designed to allow expansion and contraction, for example, or ranges of housing suitable for all stages of the life course may be planned within a single neighbourhood.

Furthermore, certain developing needs of older individuals, notably their need for help in planning for retirement and for income maintenance, point to gaps in the structures of existing professions. To be sure, advice is proffered by a number of specialists on selected aspects of the older worker's participation in the labour force or on financial management, and thousands of recreation workers offer leisure-time programs for the aged. Yet there is typically no single source of unified advice for the person who is considering retirement, who is seeking to optimize his use of leisure in retirement, or who is attempting to understand the intricacies of social security, pension benefits, insurance annuities, real property, and other assets. This gap becomes increasingly apparent as the mass of today's workers confront the necessity of planning for retirement, suggesting possible future changes in the structure of existing professions to meet developing needs.

The basis in research. The planning of programs and policies and the training of professionals to implement them rests increasingly upon research on old age and on the related processes of individual aging and social change. Early forerunners of research leading to social policy include the monographs of Frédéric Le Play (*Les Ouvriers européens,* 1855), which were used in propaganda for state reform in France entailing patriarchal control in the family; or the detailed house-to-house surveys of Charles Booth (*Labour and Life of the People,* 1889), which led eventually to the Old Age Pensions Act of 1908 in Britain. More general studies of age as an explanatory factor in social behaviour were conducted as early as the mid-19th century, when Adolphe Quetelet made quantitative analyses by age categories of crime rates and suicide rates (*Du système social et des lois qui le régissent,* 1848).

Recent social research on old age has benefitted from the advances of the social sciences generally, from the compilation of basic demographic data for many countries, and from the proliferation of empirical studies in the Western world, particularly in the United States. Research has been stimulated by interdisciplinary groups formed to develop gerontology as a special field (in particular, the International Association of Gerontology and its branches in many countries). The focus of such groups, initially centred on biology and clinical medicine, was extended in the 1950s to include psychology and the social sciences as well. The accumulation of knowledge has been fostered by universities, by foundations, and by governmental and international agencies and has begun to yield not only more rational solutions to practical problems but also a more complete understanding of old age in terms of the dynamic processes of aging and the flow of cohorts through a changing society. (M.W.R./A.F./Ed.)

RACE DIFFERENTIATION

Social relevance of physical traits

Race, in the sense relevant to racism, refers to a human group that defines itself or is defined by others as culturally different by virtue of innate and immutable physical characteristics. Thus, under racism a race is defined *socially* but on the basis of *physical* characteristics. Such physical characteristics have no inherent significance, but only such significance as is socially attributed to them in a given society. If a group in a given society is defined in terms of its skin colour, its hair texture, its facial features, its body build, and so on, then it is a "race," in racist terms. Presumably, if a group were socially defined in terms of sharing a common language, a common set of religious beliefs, or some other cultural characteristic—*without* physical considerations—then it would be an "ethnic group." In the case of both race and ethnic group, the defining criterion is relative to a given society, and the same person or group may be differently defined in different societies. Thus a black North American may be a "Negro" in his own country but a "Yankee" in Mexico. The first label is racial, the second ethnic or cultural.

In practice, the distinction between a "race" and an "ethnic group" is not always clear-cut, and many groups are socially defined in terms of both physical and cultural criteria. Jews, for instance, have frequently been regarded as both a racial and a cultural group. In such equivocal cases, it is necessary to try to determine which criterion is paramount. In Nazi Germany, Jews were primarily regarded as a race, whereas in the Middle East they are more an ethnic group. In any case, the phenomena of prejudice and discrimination that are commonly linked with differences between human groups are not markedly different whether one deals with a race or with an ethnic group. The one essential distinction is that culturally based groups generally tend to be more open and flexible in composition than physically based ones. One can convert to another faith or learn another language, whereas what one can do to one's physical appearance is decidedly limited. It should be remembered, however, that ancestry can mark a person in racist terms even though he may outwardly have no distinguishing physical trait. "Aryans" in Nazi Germany, for instance, were summarily reclassified as "Jews" when search into their ancestry discovered a Jewish forebear, and "whites" in South Africa have been similarly reclassified as "coloureds" after investigation had revealed a non-European ancestor.

Parallel to the problem of distinguishing between race and ethnic group is the problem of distinguishing between racism and ethnocentrism. Presumably, the distinction is that in ethnocentrism the alleged inferiority, disabilities, and negative traits of the outgroup are thought to be culturally determined (and *only* culturally determined), whereas in racism there is a belief that the disabilities are inborn. In practice, however, the cause of the alleged disability is not always made explicit, and it is not always easy to distinguish between racism and ethnocentrism. This is the case, for example, with anti-Semitism, which sometimes takes a cultural emphasis and sometimes a racial one. Ethnocentrism is in reality a much more widespread phenomenon than racism. Indeed, it may be said to be almost universal. Members of nearly all the world's cultures regard their own way of life as superior to that of even closely related neighbours. The exceptions to the universality of ethnocentrism are some of the societies, often small nonliterate ones, that have been conquered by more powerful ones and reduced to the status of a subject people. In some cases, after many years of oppression, some peoples have accepted their conqueror's derogatory views of themselves and flattering view of their masters. Under slavery, great numbers of American blacks, for instance, were psychologically conditioned to consider themselves inherently inferior. Unlike ethnocentrism, racism is far from universal. In fact, as a fully explicit theory of the biological causation of cultural behaviour, it is the exception rather than the rule. Ethnocentrism and racism are not mutually exclusive, of course: racist societies are almost invariably ethnocentric as well; people often feel superior to others on both racial and cultural grounds. Nevertheless, many highly ethnocentric societies are not racist.

Ethnocentric and racial aspects of anti-Semitism

The term racial discrimination denotes all forms of differential behaviour based on race. The most notable form of racial discrimination is, of course, physical *segregation* by race, but there are many others, such as rules of *etiquette* defining forms of address between racial "superiors" and "inferiors," or choice of friends or spouses. *Racial endogamy* (that is, marrying within one's own racial group) is frequently required and almost always preferred in racially stratified societies. *Commensality* rules (rules determining with whom one may or may not eat) are also a very common manifestation of racial discrimination. The term racial discrimination, then, always refers to behaviour and indeed to social behaviour.

A second aspect of racism is *racial prejudice.* Prejudice is a psychological phenomenon; it is defined as an attitude, usually emotional, acquired without, or prior to, adequate evidence or experience. It can be favourable or unfavourable and can develop in a person through suggestion, belief, or emulation. Racial prejudice consists of negative attitudes directed in blanket fashion against socially defined races (not necessarily coincident with biologically defined races).

Although it is important to distinguish between the psychological and social aspects of racism—that is, between discrimination and prejudice—it is equally essential to understand that the two are inextricably related. Prejudice against a certain category of persons could not develop in the minds of individuals unless their society had already set groups apart and subjected some of them to discrimination. Conversely, it is difficult to conceive of a society in which a system of discrimination could exist in the absence of widespread prejudice among members of the discriminating group. Discrimination and prejudice are mutually reinforcing. Prejudice is a rationalization for discrimination, and discrimination often brings forth in the victims those behaviour patterns that seem to validate the prejudice. A white bigot, for example, can easily rationalize the existence of inferior schools for blacks if he believes that blacks are not capable of benefitting from equal schooling. The same bigot, when he compares the competence of graduates of the inferior black schools and

Reinforcing roles of discrimination and prejudice white schools, will find confirmation for his belief that blacks are intellectually inferior.

If there were always a one-to-one relationship between discrimination and prejudice, the distinction would serve no useful purpose. However, this is not the case. There are wide differences in individual prejudice, tolerance, and open-mindedness within a given society. One finds relatively open-minded or unprejudiced persons who live in virulently racist societies and pathological bigots who live in nonracist societies. It is in such cases that one can expect to find discrepancies between the level of prejudice and discrimination. The open-minded white person who lives in South Africa, for example, finds it practically impossible not to be racially discriminatory in his behaviour. The law compels him to be so; and even if he chooses to go to jail in protest, he will be sent to a superior white jail. The reciprocal is also true. The bigot who lives in a nonracist society will often hide his prejudice and refrain from open acts of discrimination. In both cases, the person conforms to social norms in his outward behaviour; that is, he behaves in a manner consistent with the behaviour of the majority but inconsistent with his own attitudes.

A number of studies have demonstrated, however, that such situations are transitory for most people because they are so difficult to live with. Open-minded persons tend to become prejudiced after moving to a racist environment, or at least to become more prejudiced than they would otherwise have been; conversely, prejudiced persons become more tolerant in a nonracist society. These changes in attitudes can be quite rapid, taking no more than a few days or weeks, and changes in behaviour can be nearly instantaneous. Generally, however, there is an appreciable time lag between change in behaviour and change in attitudes, and it is this lag that makes the imperfect correspondence between discrimination and prejudice.

The history of racism. In nearly all the world's societies, men have apparently developed pride in the cultural accomplishments of their own groups and a corresponding derogation of those of their neighbours. Notably, however, the idea that certain groups of people are superior to others because of their genetic makeup does not appear to have been widespread. Where it now exists, it is mostly an outgrowth of the rationalizations of slavery and colonial expansion in the vast territories dominated by European settlers.

Hindu racism. Some scholars have argued that the Hindu caste system originated in a physical difference between the conquering Aryans and the conquered Dravidians. The top three caste levels, or *varṇa*—Brahmin, Kṣatriya, and Vaiśya—were in all probability recruited from the ranks of the Aryans, whereas the fourth *varṇa,* the Śūdra—as well as the "untouchables" who fall outside the scope of the four *varṇa*—were perhaps originally composed mostly of Dravidians. The word *varṇa* means "colour," even though the colours traditionally associated with the *varṇa*s show no relation to skin pigmentation. It is true that in Hindu culture, the colour dualism of white and black has associations with good and evil as it does in the Judeo-Christian tradition, but the link with skin pigmentation is not established. In modern India, there is a weak association between light skin and high caste, and perhaps a slight overrepresentation of lower castes among the darker skinned southern Indians, but skin colour in India is anything but a reliable index of social status. Krishna, the most cherished character in the Hindu mythology, is often represented as a darkly handsome young man frolicking with a bevy of fair girls. Many Indians express a marked aesthetic preference for light skin, especially in women, but, on balance, the evidence that the Indian caste system is racial in origin and that India is or was a racist society is unconvincing. The basic caste dichotomy between "once-born" and "twice-born" was probably related to the *cultural* distinction between Aryan conquerors and Dravidian conquered. The latter were probably darker skinned than the former, but it is not established that this physical distinction was the socially significant one.

Skin colour and Hindu caste

Middle Eastern racial attitudes. The Bible, one of the oldest sets of written documents in the Judeo-Christian tradition, contains no positive suggestion that the ancient Semites were racists. The ambiguous reference in the Song of Solomon ("I am dark, but comely, O daughters of Jerusalem, like the tents of Kedar, like the curtains of Solomon. Do not gaze at me because I am swarthy, because the sun has scorched me. My mother's sons were angry with me, they made me keeper of the vineyard;") might be interpreted as expressing a prejudice against blackness. Several other passages in the Bible, notably the curse of Noah on Canaan in Gen. 9:22–27, have been interpreted by later racists in a racial light, but in themselves they do not necessarily suggest anything of the kind, and, on balance, the evidence is negative.

The same is true of the Qur'ān and the Islāmic tradition in which religious conversion has always been the test of membership in the ingroup. Although it is true that black Africa was the source of many of the slaves in Muslim countries and that, to this day, in some Arab countries a dark skin is a presumptive sign of slave descent and hence of low social status, none of the Muslim countries ever developed a racial caste system. Since the spread of Islām across the Sahara in the early 11th century, it has become by far the most important religion in black as well as in white Africa, and by now approximately one-fifth of the world's and one-half of Africa's Muslims are black. From medieval times, black African kings went on pilgrimage to Mecca with thousands of their subjects and were well received, and Arab writers on black Africa reveal no racism in their writings. They often express admiration for the black Muslim kingdoms that they encountered, and whatever negative value judgments they make are religious or moralistic in nature (such as the disapproval of the nakedness of women). Even the devastation brought about by the Arab slave trade in East Africa in the middle of the 19th century does not appear to have been rationalized on racial terms as European slavery was.

East Asian racial attitudes. Chinese and Japanese cultures have traditionally been conscious of physical differences, especially from an aesthetic point of view, but neither society ever developed a racial caste structure. Even in pre-Tokugawa Japan, where there existed a fairly rigid "estate" system similar to that of medieval Europe, the distinctions were apparently not racial, and the Burakumin caste—an outcast, "untouchable" group that still exists today—is not physically distinguishable (even though one theory, popular in Japan, holds that the ancestors of the Burakumin were anciently Korean war prisoners and slaves). In any event, both Chinese and Japanese express a sometimes marked aesthetic dislike for physical types that are at variance with their own well-established canons of beauty. A 16th-century Chinese mandarin, for example, likened the hairiness of Portuguese sailors to that of monkeys and pronounced the Portuguese totally lacking in the social graces of civilized life, though surprisingly clever and capable of learning. Many Japanese find black skin and other Negroid characteristics unattractive and, since the recent wave of American influence, appear to have adopted some Western canons of beauty, as witnessed by the eyelid operations undergone by some women. The generally inferior status of Koreans in Japan seems to be based more on socioeconomic and cultural than on racial criteria; and the distinctive-looking Ainu minority on Hokkaido Island do not appear to be the target of racial prejudice. Thus, despite narcissistic canons of physical beauty and highly ethnocentric judgments of other cultures, East Asian civilizations do not exhibit what might properly be called racism.

Japanese outcasts

Racism in nonliterate societies. The evidence from the world's many nonliterate societies is fragmentary on the topic of interethnic and interracial relations. There are, however, a few documented cases of indigenous systems of racism not attributable to contact with Western societies. One of the most notable cases is that of the traditional kingdoms of Rwanda and Burundi in the interlacustrine area of central Africa. Until the overthrow of the *ancien régime* shortly after independence in 1961, the population was stratified into three racial castes. The tall, semitic looking Tutsi constituted some 15 percent of the population and, as pastoralist conquerors, ruled over the Hutu horticulturalists. The Hutu, of average stature and distinctly Negroid in physical appearance, constituted a servile peasantry of approximately 80 to 85 percent of the population. At the bottom of the hierarchy was a small group (approximately 1 percent of the total) of pygmoid Twa. The close relation between physical stature and social status made for a truly racist system. The Tutsi claim to superiority was based to a considerable extent on their physical characteristics.

European racism. Far and away the most widespread, enduring, and virulent form of racism and the costliest in terms of human suffering has been that which developed in western Europe and its colonial extensions in Africa, Asia, Australia, and the Western Hemisphere. Western racism is of relatively recent origin. None of the main waves of influence on European civilization seems to have brought racism with it. In ancient Greece and Rome, the status criteria were cultural and not racial. The barbarian was the person who spoke another language. Slavery was a juridical and economic condition unrelated to racial or ethnic origin. The statuses of patrician or plebeian, patron or client, slave or freeman, citizen or noncitizen were all based on strictly nonracial criteria. Although Roman literature is often quite ethnocentric in its judgment of barbarians, differences in physical appearance are generally mentioned neutrally as interesting *exotica.* A number of black Africans must have reached Rome in the last centuries of the empire, and blacks appear on frescoes in a number of Roman sites in Europe and North Africa, but there is no evidence that they were regarded as inherently inferior.

In the Middle Ages, the religious criterion of membership in the ingroup became paramount. Anti-Semitism was clearly religious and not racial in nature and continued so through the Renaissance, the Reformation, and the Wars of Religion of the 16th and 17th centuries. The important thing was whether one was Christian, Moor, or Jew, Catholic or Protestant. Even at the height of religious persecution during the Spanish Inquisition and the Thirty Years' War, acceptance was typically granted in return for a simple public profession of faith. "Race," in any case, was irrelevant.

The anti-Semitic wave that swept Germany in the 1930s and ended in the crematoria of the death camps in the 1940s will hopefully remain the most heinous manifestation of racism in human history. Although Nazi anti-Semitism grew out of a long tradition of religious intolerance in Europe and more especially of victimization of Jews in eastern and central Europe, Hitler's theory of the master race gave it a hitherto unknown genocidal virulence.

Even when Europeans first came into more extensive and direct contact with large numbers of dark-skinned peoples, as a result of their colonial expansion starting in the late 15th century, racism took some time to develop.

The Spanish

The Spanish conquest of the New World was more than averagely brutal, and the economic exploitation of the Indians was thorough; but the Spanish crown quickly settled the issue of the Indians' humanity by declaring that they did have souls worthy of salvation and that they should not be enslaved. It is true that in the Spanish colonies of the Americas, there developed a caste system that was at least partly racial: a fundamental distinction was made between an Indian, a *mestizo* (or mixed-blood), and a Spaniard; but an equally important difference was made between a Spaniard born in Spain and a criollo, a Spaniard born in the colonies. Initially, the Spaniards expressed wonder and admiration for the accomplishments of the Indian civilizations that they destroyed. This did not prevent them from condemning what they regarded as barbarism and idolatry, but there is no suggestion that the Spaniards regarded the Indians as genetically inferior when they first met them. Only after the Indians had been reduced by epidemics, wars, tribute exactions, and countless acts of brutality to the status of an impoverished servile peasantry did negative stereotypes about them develop. The Spanish colonial caste system established in the 16th century was continuously undermined by a dual process of hispanization and miscegenation, which blurred both cultural and racial distinctions between groups. By the time that the criollos gained their independence from Spain in the second decade of the 19th century, the caste system had lost its social significance except in the countries that, like Guatemala, Peru, and Bolivia, were still predominantly Indian.

Muted racism in Spanish America

Today, racism, though not totally absent in Spanish America, is certainly much less prevalent than in other parts of the continent. Cultural criteria are far more important than physical ones in most of Spanish America, even in the heavily Indian countries. In some Latin American countries, the term *mestizo* no longer generally denotes a person of mixed Spanish-Indian descent but instead often designates a person who speaks Spanish as his mother tongue even though he may be of pure Indian ancestry. Many derogatory stereotypes about Indians do continue to exist, and in the heavily Indian countries, Indians continue to be subjected to various forms of social discrimination; but, by and large, an Indian ceases to be regarded as such when he becomes hispanicized in his language, customs, and religion. Racism, as such, is minimal.

The Portuguese

The Portuguese case is somewhat different from the Spanish one. In Africa the initial contacts between Portuguese and Africans were relatively free of racism and relatively peaceful and friendly, except in east Africa, where Portugal came in conflict with Arabs and Persians. There is, for example, a record of a friendly correspondence between the kings of Portugal and Kongo and of an exchange of ambassadors. Starting in the middle of the 16th century, these auspicious beginnings were increasingly compromised by Portuguese military incursions, the sacking of cities, and incessant demands for slaves for the Brazilian plantations. Brazil soon became, with the West Indies and the Southern English colonies of North America, one of the world's three major consumers of black slave labour, and it remained so until the middle of the 19th century. Generally, the Portuguese claim that its colonialism in Africa has been nonracial is correct, at least by comparison with the British, Belgian, and Dutch. This is not to say that the Portuguese regime in Africa has been any less oppressive and exploitative than the regimes of the other colonial powers, but whereas the latter have frequently applied racial tests of discrimination, the Portuguese have been ethnocentric rather than racist.

Complex race relations in Brazil

In Brazil, race relations are quite complex and vary greatly from one region to another. Brazil's reputation as a "racial paradise" appears to most observers to be undeserved; on the whole there is considerably more racism in Brazil than in Spanish America, though much less than in the United States. Consciousness of physical differences is highly developed, and Brazilians use a complex racial nomenclature to describe a score or more of combinations of facial features, hair texture, and skin pigmentation resulting from the intermixture of Afro-, Euro-, and Indo-Brazilians. Certain stereotypes are attached to these physical types as well as an order of aesthetic preference, but the very complexity of the nomenclature itself has made impossible the establishment of any system of racial castes, and members of the same family may frequently fall into different "racial" categories. Thus, Brazil might be described as a highly racially conscious country but without a rigid system of racial castes and without well-defined forms of racial discrimination. Such discrimination as exists is usually a subtle combination of racial, ethnic, and social-class factors, with race frequently not the most important one.

The French

The French, like the Portuguese and Spaniards, tended to be more ethnocentric than racist in their colonial policy of "assimilation," but in practice, like the Portuguese, they failed to assimilate more than a tiny minority of their African subjects. Consequently, except for a few black professionals and intellectuals who were totally integrated in French society, French colonial policy was not all that different in practice from that of more avowedly racist powers. It should also be noted that in Algeria, where there was a European settler population of more than 1,000,000, the French exhibited considerable racism vis-à-vis the Arabs.

The Netherlands and Great Britain were responsible for the growth of the most racist colonial societies that the

The Dutch and English

world has ever known—namely, South Africa, the United States, and Australia. In Australia, racism has taken the form of discrimination against the aboriginal population and the exclusion of immigrants of non-European stock. In the United States, virtual genocide against the Indian groups was accompanied by the institutionalization of a slave plantation system and, after 1865, of racial segregation and discrimination against the "emancipated" Afro-Americans. In addition, a number of states adopted racially discriminatory laws against other nonwhite groups, notably Asians, in the field of immigration, marriage, and political rights. As late as World War II, tens of thousands of American citizens were interned for several years in camps, solely on the basis of their Japanese ancestry.

The South African policy of apartheid has become a byword for racial discrimination and, next to the Nazi policy of genocide against Jews, represents the most extreme and systematic form of racism practiced in a modern society. The white government of South Africa is attempting to create four rigid colour-castes (Europeans, Asians, Africans, and Coloureds), to segregate them physically, and to perpetuate the economic and political privileges of the white minority at the expense of 82 percent of the population. In theory, apartheid aims to establish a "separate but equal" system, but in practice the indefinite maintenance of white supremacy is clearly the objective. Countless laws limit the nonwhite South Africans' rights to travel, own and occupy land, hold meetings, seek work, attend universities, enter public places, marry, vote, and indeed be present almost any place without the consent of the white authorities.

Causes of racial prejudice and discrimination. As a well-developed theory, racism is a fairly recent phenomenon, even in Western history. The 18th century was predominantly environmentalist in its outlook; the science of that day tended to attribute social behaviour either to climatic and geographical environment or to sociocultural factors. Racism as a widely accepted "scientific" theory of behaviour did not appear until the 19th century, which was the age of racism *par excellence.* Although Charles Darwin himself was not a racist, his theory of biological evolution was extended to social evolution, giving birth to the theory of social Darwinism. Mankind was regarded as having achieved various levels of evolution, culminating in the white-European civilization. These stages of evolution were thought to be related to the innate genetic capabilities of the various peoples of the world. By the second half of the 19th century, racism was accepted as fact by the vast majority of Western scientists, and various forms of it were popularized through the writings of Joseph-Arthur, comte de Gobineau, Houston Stuart Chamberlain, Rudyard Kipling, Alfred Rosenberg, and Adolf Hitler.

"Scientific" support in 19th century

By the 1930s the intellectual climate had swung clearly away from racism, and racism had lost its apparent scientific respectability. The social sciences began to adhere to a strict theory of the social determination of human behaviour, to the nearly complete exclusion of biological or physical-environmental factors. Man was held to be almost entirely a product of his culture, and each culture was to be evaluated in its own terms. This cultural relativism was popularized by such anthropologists as Ruth Benedict and Margaret Mead, and the study of race relations became distinctly antiracist in orientation. Gordon Allport, Otto Klineberg, Roger Bastide, Gunnar Myrdal, and E. Franklin Frazier, to name but a few of the prominent scholars active in the 1930s and 1940s, all took the position that such "racial" differences as are found between human groups are attributable to the differential social environments in which they find themselves and not to any intrinsic physical properties.

Most recently the pendulum is once more swinging away from the extreme social environmentalism and cultural relativism that dominated the social sciences until the 1950s. Genetic and biological factors in human behaviour and aptitudes are once again becoming accepted as having at least some importance, although the crude racism, evolutionism, and social Darwinism of the 19th century seem to be permanently discredited.

As a rule, complex social phenomena like racism cannot be explained in terms of a single causal factor. Causation is not only multiple but also often reciprocal, in the sense that A generally is a cause of B and B in turn frequently is a cause of A. This is clearly what happens in the relation between discrimination and prejudice. Each is a cause of the other, and this vicious circle is often difficult to break. The reciprocal causation between discrimination and prejudice illustrates the interplay between psychological and sociocultural variables, and both approaches are essential to a comprehensive understanding of race relations.

Specious popular theories. Numerous theories on the causes of racism have been advanced, some patently false, others partly true; at the present stage of knowledge, only an eclectic but selective acceptance of several of the theories can give a comprehensive view of the phenomenon. Perhaps one of the commonest popular theories to account for prejudice and discrimination (both racial and ethnic) is that it is "their own fault." The theory advanced by most dominant groups in racially or ethnically stratified societies is that the allegedly reprehensible behaviour and qualities of the outgroup causes discrimination and prejudice in the ingroup. This theory, even when it is not made explicit, is often implied in the very definition of the situation as the "Negro problem," the "Jewish problem," the "Asian problem," and so on. It should be clear that such an approach is a *symptom* of racism or ethnocentrism and not an explanation thereof.

"Their own fault" theory

Another unviable theory of the cause of racism is that there is an innate or instinctive repulsion between groups of people who look different. Experiments on young infants have shown that infants could easily be conditioned to scream with fear when a white rabbit entered a room and, conversely, to pet snakes. There seem to be no innate dislikes of any animals in man, much less of relatively minor differences in appearance between members of the same species. If this is true, then there is no intrinsic relation between racism and the sheer presence of physical differences. Physical differences are one of the important conditions facilitating the development of racism once it is there, but they are not a cause of racism. Differences in physical appearance are simply visual stimuli to which prejudices may or may not become attached.

Psychological theories. Several psychological theories about the causes of racial and ethnic prejudice have linked the phenomenon with certain personality traits or with certain responses to social situations. The "frustration-aggression" theory holds that frustration frequently leads to aggression and that this aggression becomes "displaced" onto scapegoats that are quite unrelated to the source of the frustration. Outgroups are frequently blamed for one's frustrations and failures, and this displacement of aggression is often accompanied by "projection"—that is, by the attribution to others of one's own undesirable or unavowed traits. Although a number of experiments have shown that experimentally induced frustration can lead to displaced hostility against outgroups, the theory does little to explain why certain persons are more apt to displace their hostility than others and why certain groups are chosen as scapegoats. It also fails to account for the fact that highly privileged groups can be as strongly prejudiced as groups that have suffered much frustration.

The "authoritarian personality" approach holds that persons who exhibit certain attitudes and personality traits such as respect for power, submission toward superiors, aggression toward subordinates, lack of self-insight, superstitiousness, and contempt for weakness are predisposed to be generally prejudiced against all ethnic and racial outgroups. Numerous studies have shown some relation between these factors, at least in Protestant, Anglo-Saxon cultures, but the theory does not easily account for great differences in levels of prejudice between groups of people who show substantially the same amount of "authoritarianism." Two matched groups of American whites, one Southern, one Northern, for example, showed nearly identical levels of authoritarianism, but widely discrepant levels of anti-Negro prejudice.

Authoritarian personality theory

The relation between personality traits and prejudice seems to vary considerably depending on the social climate. Thus, in societies in which racial discrimination is the norm, as in South Africa, even persons who show little authoritarianism will typically be prejudiced. (This has been referred to as "conformity prejudice" to distinguish such attitudes from the more psychopathological forms of bigotry.) It may also be that, in societies that do not openly condone discrimination, the authoritarian personality is more apt than other personalities to be a bigot. Thus, although attempts to apply this psychological theory at the societal level—as, for example, by explaining the rise of Nazism through the authoritarian structure of the German family—are at best unconvincing, the theory seems able to account for individual differences in levels of prejudice within the same cultural group.

Economic influences. Social scientists have stressed a number of social variables that may be causally linked with prejudice and discrimination. Economic factors are of unquestioned importance. Marxian writers have interpreted racism as a rationalization for slavery and colonialism and also as a means of splitting the working class along colour lines and of deflecting attention from the central reality of class conflict to the ancillary problem of "race." It is certainly no accident that racism flourished at the time of the second great wave of European colonial expansion and the scramble for Africa, and the ideology of colonialism and the white man's burden was often expressed in racist terms. It is also true that racism provided ideological justification for slavery. But slavery antedates the development of racism in Western societies, and some slave societies, notably those of Latin America, have been much less racist than those of English or Dutch origin in North America, the Caribbean, and South Africa.

Non-Marxian writers have stressed the importance of other economic factors, such as competition for jobs, as problems aggravating interethnic or interracial conflicts. Racial segregation in urban ghettos and its resultant conflicts have been linked with the pattern of absentee landlordism in slums and the control of the housing market by realtors and investors. In blatantly discriminatory societies like South Africa the link between racist legislation and the economic interests and privileges of the white minority is obvious. A common form of economic determinism in racial and ethnic relations is the "middleman syndrome." A number of scattered minorities, often of alien origin, are subjected to much the same syndrome of prejudice. Indians in east and South Africa, Lebanese in west Africa, Chinese in Southeast Asia, and Jews in various countries, insofar as they have concentrated in retail trade and middle-level white-collar occupations, have often been the victims of an "anti-Semitic" type of prejudice, being accused of clannishness, rapaciousness, underhandedness, dishonesty, stinginess, and exploitation of the indigenous majority. These stereotypes are due in part to the envy felt toward an alien minority whose living standard is often above the average in the host country.

Minority middleman as bias target

Relations of production—that is, the economic and class position of the constituent racial or ethnic groups within a society—are always of crucial importance in determining the type of relations that exist between these groups. Equally important and closely related to the relations of production are the relations of power between groups. For example, a multiracial society in which a minority exerts a clear domination over a majority, such as is characteristic of colonial regimes, makes for a type of race relations different from that of an ostensibly democratic society such as the United States, where a majority discriminates against a minority. An agrarian, patriarchal state such as characterized the slave regimes of the Western Hemisphere or the Boer republics in the 19th century makes for a type of race relations different from that of a highly urbanized and industrialized country like modern South Africa. The techniques of racial domination, for example, are radically different in a slave plantation and in an urban ghetto. Relations of power determine the legal structure of the society, and many studies have demonstrated the importance of legislation in changing a society either in a more egalitarian direction, as in the United States since the 1940s, or in a more racist one as in South Africa under apartheid.

Religion's role. Religion has also been shown to be related to the amount of prejudice and discrimination. There is an undeniable difference between the more racially tolerant Catholic countries of Europe and their colonial extensions and the more racist Protestant countries. The Catholic Church has frequently taken a more universalistic position and rejected racism, whereas many Protestant denominations, especially the more fundamentalistic and puritanical ones, have often interpreted the Scriptures in a racist fashion. The role of the Dutch Reformed churches in South Africa in supporting apartheid as the will of God is well known, and Protestant Fundamentalism in the United States has sometimes also been deeply racist.

Demographic influences. Numerous racial studies have stressed the importance of demographic and ecological factors in intergroup relations. The demographic ratios between groups, the number of distinguishable groups, and the geographical concentration of these groups within a country all affect the system of intergroup relations. Thus situations in which the dominant group is a majority are different from those in which it is a small minority, and different again from those in which two groups of nearly equal size maintain a competitive balance of power. It also matters whether groups are evenly spread throughout a country or heavily concentrated in a given province or packed in urban ghettos. And the balance of power of course varies according to whether there are two, three, four, or many groups.

Social and cultural factors. The degree of cultural differences between groups is also a relevant factor. Ethnic or racial prejudice is often exacerbated by barriers of language or customs. Acute racial conflicts, however, can also exist between groups that are culturally nearly identical—as between white and black Americans.

Situations of rapid change are often marked by an intensification of racial and ethnic conflicts because one aspect of change is frequently an alteration in the traditional relations between groups. Such changes are commonly perceived as a collective threat, and this makes for outbursts of conflict. The rapid immigration of racially distinctive groups, as in post-World War II Britain, is a case in point. The rapid influx of a hitherto rural group into urban areas can have the same effect, as shown in the urbanization of Afro-Americans in the last few decades.

Conflict from sudden social change

International conflict can have domestic repercussions in intergroup relations, as witnessed by the wave of anti-Japanese racism in the United States during World War II or by the wave of anti-Algerian feelings in France during the Algerian war of independence.

The complex interplay of these numerous social and psychological factors on racial and ethnic relations makes predictions and generalizations hazardous. No single factor ever accounts for more than a fraction of the phenomenon to be explained, and the relative importance of various factors is often difficult to establish.

Effects of racial discrimination. The consequences of racial discrimination are as diverse as its causes. It is useful to distinguish here between psychological and social effects and between those effects on the group that discriminates, on the group that is discriminated against, and on the society at large.

Effects on victims. There is little doubt that psychologically racism is harmful to its victims. The most profound effect of racism associated with situations of extreme degradation (such as is found under slavery or in concentration camps or in racist states like South Africa) is the acceptance by the oppressed group of the dominant group's definition of the situation. This is the phenomenon of self-hatred found, for example, in cases of Jewish anti-Semitism or in the acceptance by blacks of white aesthetic criteria such as the desirability of having straight hair or a light skin. Self-hatred is often accompanied by neurotic symptoms of apathy, anxiety, and depression or by forms of self-destructive escapist reactions such as alcoholism or drug-addiction or, in extreme cases, by paranoid, schizophrenic, or manic-depressive psychoses. In such situations of extreme degradation then, the oppressed group frequently reacts in an "intropunitive" fashion; that is, it turns its frustrations inwardly against the self or the ingroup at large. At the social level, this intropunitiveness takes the form of predatory crimes by an organized underground against the oppressed group. Racial ghettos in the U.S. and South Africa, for example, have very high rates of crimes committed by blacks against other blacks, a phenomenon that is encouraged by the disinterest of the police in providing adequate protection.

When victimized groups do not accept their inferior status or when conditions improve to such a degree that they regain self-respect and conceive of the possibility of changing the status quo, frustration turns outward as hostility and aggression against the dominant group. Paradoxical as it seems, situations of open racial conflict, like other forms of revolution, are generally associated with periods of both relative and absolute improvement in the position of the subjugated groups. Hostility toward the dominant group may take many forms, ranging from nonviolent passive resistance to apolitical crime and politically inspired guerrilla warfare. The search for an identity independent of the dominant group's definition of the situation is often only a preliminary step to concerted political action, frequently violent, to try to overthrow the existing racial order. Group cohesiveness and political militancy replace apathy and self-hatred.

Turning against the dominant group

Effects on the dominant group. From the perspective of the dominant group, the effects of racism are more mixed. Psychologically, racism warps the personality of the oppressor as it does that of the oppressed, though probably in a less devastating way. Racially prejudiced persons living in highly racist societies can behave "normally" in situations not involving race; yet racism, by erecting an irrelevant and artificial barrier between people, strains relations and distorts social perception in the dominant group. Social consequences of racism for the dominant group can vary widely. Generally, in colonial-type situations in which the dominant group is a minority and has entrenched itself in an economically and politically privileged position, the benefits it derives from racism can be considerable. Thus the artificially high standard of living of the white South African population probably can only be maintained by the elaborate apparatus of racial laws limiting the freedom of movement, organization, and employment of the nonwhite majority. Some of those material benefits for the whites are diverted to the maintenance of the repressive apparatus; however, the economic balance remains positive for the whites, even though the total economic cost of segregation for the society as a whole is quite high.

In situations in which the dominant group is in large majority, as in the United States and Australia, the economic benefits of racial discrimination to the dominant group as a whole tend to be much more marginal and are often overshadowed by the costs in conflict, violence, and lost productivity. There are also situations in which the economic effects of racism may be opposite for different segments of the racially dominant group; in the antebellum South, for example, slavery benefitted the slave- and land-owning aristocracy but not the white yeomen and workers.

If one attempts to assess the overall effects of racism on an entire society, one must conclude that they are negative insofar as racism erects an artificial barrier to the full use of talents and often generates destructive conflicts.

Three modes of relations

The reduction of racial discrimination. In practical terms, situations of racial discrimination and conflict may be reduced to three broad types:

Segregationist or apartheid societies. In societies like South Africa, it is clearly in the collective interests of the racially dominant group (typically a minority) to maintain the status quo. The reduction of racial discrimination in such cases can come about through a drastic change in the power structure (as it did in the U.S. South as a result of the Civil War) or it can come about in evolutionary fashion (as it did in colonies ruled by Great Britain); but, in any case, the new ruling group no longer feels the need to defend its interests by maintaining a racial caste system.

Pluralistic societies. Those societies in which several racial groups, none of them clearly dominant, compete for power and economic resources (as in some parts of Africa and Asia) probably offer the widest range of alternatives for the reduction of racial tensions. Racial and cultural assimilation may reduce differences and blur old lines of cleavage. Or a more amicable modus vivendi may develop from changes in government policy, the spread of a broader nationalist ideology, or changes in the economic and political structure making for a more democratic society. When conflicts escalate to chronic violence and civil war, political partition and emigration are other possibilities for reducing tensions.

Societies publicly committed to ending racial discrimination. In some societies, as in Australia and the United States, official government policy is often against racial discrimination and has to contend both with the inertia of conservative opinion from the dominant group and the "revolution of rising expectations" from the hitherto subordinated group.

There seem to be two basic approaches to the alleviation of racial tensions in this third type of society, assuming an official desire to do so. One is to attack racial prejudice by educating the public. There is abundant evidence that people's attitudes can be changed by propaganda and that behavioral changes follow changes in attitudes; but there is also evidence that as a method to bring about rapid social change in a conservative population, this approach is relatively expensive and ineffective. In terms of practical consequences, it may be more important to reduce racial discrimination than racial prejudice, and, hence, it follows that the strategy of attacking discrimination is frequently more directly effective.

Dilemma of means and ends

Legislation is undoubtedly one of the main ways of destroying racial discrimination, provided it is followed by forceful implementation. This often entails a dilemma of means and ends: the achievement of a more democratic society may imply the use of force against a majority—a policy that democratically elected governments are naturally loath to adopt. Consequently, attempts to outlaw various aspects of racial discrimination (for example, in housing and employment) have frequently failed to bring about the rapid change that was intended because of a lack of determined implementation. The delays in school integration in the United States can be considered a case in point. By contrast, integration in the U.S. armed services has been more rapid and thorough, largely because it was done autocratically.

At the economic level, it is clear that the profit system of production helps to perpetuate disabilities for racial minorities. High returns on slum property, for example, are one of the factors making for the persistence of racial ghettos in the United States. Racial discrimination in housing creates an artificial scarcity of housing for blacks, who are thus forced to pay higher rents than their white counterparts. This makes for high returns on investments in the black areas that, because of high population density and municipal neglect, quickly become slums. Other economic systems, such as the "consumer credit," are also known to discriminate against the poor and thus to affect racial minorities disproportionately. Reforms in the economy might be effective in reducing racial discrimination, but the realistic prospect of such reforms in the United States is limited.

In the last analysis, the most effective way of reducing racial discrimination may be militancy on the part of the groups that are discriminated against. When such groups are large enough to affect the outcome of elections and when the political system is democratic enough to give the voters some real alternatives, political action can be effective within the constitutional framework. When such conditions are lacking, political action ranging from civil disobedience to guerrilla warfare has been known to bring about important changes. (P.L.v.d.B./Ed.)

SOCIAL STRATIFICATION: CLASS

Although the economic basis of classes has been generally recognized, many disagreements have arisen as a result of attempts to define social class. First, there has been confusion about how widely the concept should be employed. It was used initially to denote the social groups that emerged in postfeudal societies, but it was then applied to the divisions in a wide range of societies. James Madison, for example, said in *The Federalist:*

> Those who hold property and those who are without property have ever formed distinct interests in society. Those who are creditors, and those who are debtors, fall under a like discrimination. A landed interest, a manufacturing interest, grow up of necessity in civilized nations, and divide them into different classes, actuated by different sentiments and views.

Similarly, Marx and Engels in *The Communist Manifesto* spoke of "freeman and slave, patrician and plebeian, lord and serf, guild master and journeyman" as being engaged in class struggles and regarded the distinctive feature of modern societies as being merely that the class antagonisms have been simplified, so that society is more and more splitting up into "two great classes directly facing each other: bourgeoisie and proletariat." In an unfinished chapter on social classes in *Das Kapital,* it seems to have been Marx's intention to confine the use of the term class to the "three great classes of modern society": wage labourers, capitalists, and landowners.

A major problem is that of choosing between a definition of class that would make it possible to apply the concept to all forms of society—ancient city-states, early empires, caste society, and feudal society, as well as to modern capitalist societies—while taking into account the specific characteristics of each of these forms; and a definition that would confine the concept to the social divisions in modern societies, treating other forms of society as having quite different kinds of stratification, arising from other factors besides the ownership of property. It is difficult, for example, to apply the concept of class to the traditional Indian caste system, which seems to depend, to a large extent, upon occupational differentiation and religious distinctions and which also seems to result in a fragmentation of society that is an obstacle to the formation of broad social classes (see below *Social stratification: caste*).

Even when the term class is applied to modern societies differences of view arise about its scope and meaning. Some of these differences result from disagreements concerning the relations between classes and the influence of classes in political life, and they will be more fully considered below. There are also problems, however, in distinguishing classes in a precise way from elites, status groups, or interest groups. The concept of elite, for example, was elaborated by the Italian sociologists Vilfredo Pareto and Gaetano Mosca as an alternative to the Marxist concept of the ruling class, and, although it has not been used consistently in this way, it does still overlap, to some extent, with the notion of class. Among recent writers, an American, C. Wright Mills, used the term power elite instead of ruling class (1956), in order to avoid the assumption that political power is always based upon property ownership; others have made a careful distinction between elites and classes and have been concerned to examine the nature of the connection between them.

Classes and elites

The distinction between classes and status groups was made most clearly by the German sociologist Max Weber, who regarded the former as being based upon economic interests, while the latter were constituted by evaluations of the honour or prestige of an occupation, cultural position, or family descent. But this distinction also has not always been observed, with the result that individuals having roughly the same social prestige, such as those in a particular category of occupations, have sometimes been regarded as forming a class.

It is also necessary to differentiate between social classes and interest groups, even though classes themselves may be regarded as being, in a very broad sense, interest groups. In modern societies, particularly, interest groups are numerous. They include many occupational associations, as well as cultural, recreational, regional, and other groups that pursue their special interests by a variety of means. The number of social classes is, in the view of all writers on the subject, much more limited, and classes may be seen as pursuing the more general economic and political interests of their members and as being involved, to some extent, in shaping and changing the character of society as a whole.

Another major disagreement in the definition of social class exists between those who regard classes as real social groups, possessing a class consciousness and exhibiting a distinctive culture and style of life, and those who treat "class" as an analytical construct, useful for some purposes in distinguishing categories of individuals on the basis of their income, education, or occupation. The first kind of definition is to be found in Marxist theory, which is concerned particularly with the historical development of classes as social groups having a determining influence upon the course of social events. Marx himself sketched a historical account of the rise of the bourgeoisie and, subsequently, of the proletariat (see especially, *The Poverty of Philosophy,* 1847) and made a distinction between the "class in itself" (that is, a collection of individuals who were placed in similar economic conditions) and the "class for itself" (that is, a more or less organized group, pursuing in a conscious fashion definite economic and political ends). Among later Marxist writers, the Hungarian philosopher and critic György Lukács, in several essays collected in *History and Class Consciousness* (1923), emphasized the importance of class consciousness, while introducing a new element in his account of the role of revolutionary intellectuals in the formation of such a consciousness. It is, however, not only in Marxist theory that the idea of class, in Marx's latter sense, has prevailed, since non-Marxist sociologists have made studies of working class, middle class, and upper class as distinct communities; and Weber regarded class as one of the most important bases for communal action.

Marx's two concepts of class

The concept of class as a collection of individuals sharing similar economic circumstances has been used widely in censuses and in studies of social mobility. Thus, the Office of Population Censuses and Surveys in Britain makes use of five broad categories of social class, based on an occupational classification; and studies of social mobility, examined below, have generally employed categories formed in a similar way by grouping together occupations that are regarded as being at approximately the same level in the social hierarchy. Some other sociological studies have made distinctions between categories of individuals in terms of prestige, without paying much attention to questions concerning the formation and development of classes.

Theories of social class. Many disagreements over the definition of class are bound up with differences in theoretical approach, involving divergent conceptions of the importance of classes in the social structure and of the nature of the relationship between classes. Theories of social class were only fully elaborated in the 19th century as the modern social sciences, especially sociology, developed. Earlier writers, although they could not fail to observe the existence of great economic inequalities in their societies, did not attempt to formulate a general theory of the causes and consequences of such inequalities. At the most, they reflected, as did Plato and Aristotle, upon some of the more prominent features of social stratification and proposed ideal systems that would be conducive to political stability and rational government. Plato thus advocated a society divided into three classes—guardians, auxiliaries, and workers—in which the guardians would form a disinterested ruling elite, while Aristotle recognized in society three elements—the very rich, the very poor, and those in the middle—and considered the best political system to be that in which the middle strata had a preponderant influence.

Class relations in the political theory of Plato and Aristotle

Pre-Marxist views. In the 17th and 18th centuries a new discussion of social inequality and stratification began, provoked largely by the challenge to aristocratic rule from the bourgeoisie. In the writings of the English political philosophers Thomas Hobbes and John Locke there was developed an individualistic theory of property, corresponding with the growth of a market economy, which either assumed, in Hobbes' case, an equality among men, a struggle between individuals needing to be regulated by a powerful state, and the absence of cohesive classes; or, in the case of Locke, accepted the existence of a division between property owners and a labouring class. The French political philosopher Jean-Jacques Rousseau, in an essay on the origin of inequality (1755), made a distinction between natural inequalities and those that resulted from social conventions, but he conceived society in terms of the relations between individuals rather than between classes. Madison, writing in a society which had declared the principle of human equality and in which few aristocratic privileges existed, was more aware of the social divisions that might arise from divergent economic interests, and his emphasis upon the distinction between the owners of property and those without property foreshadows one of the important elements in Marx's theory. Madison, however, went on to distinguish a considerable number of interest groups as the basis of political factions, and he was far from conceiving society as a class system in which one class must necessarily dominate others.

Other writers in the later 18th and early 19th century who contributed elements of a theory of social classes include the Scottish philosophers Adam Ferguson and John Millar, in whose works a distinction was established between civil society (*i.e.,* the nonpolitical elements in society, such as the economic system and the family) and the state. The conditions of civil society, especially the property system, were seen as determining, to a large extent, the form of political life. This idea was expounded by the French social theorist Saint-Simon (1817), who argued that government was only a form that corresponded with the character of the underlying system of production and who analyzed the changes in production that led to the decline of the feudal ruling class and the rise of the new class of industrialists. In the writings of Saint-Simon's Socialist followers, a theory of the proletariat was first adumbrated (1824). A later work, by the sociologist and legal theorist Lorenz von Stein (1842), developed a conception of the proletariat as a major political force in modern society. This directly influenced the development of Marx's theory.

Another contribution to the theory of class came from political economy. In Adam Smith's *Wealth of Nations* (1776) the principal emphasis was upon social differentiation and stratification resulting from the division of labour, but the English economist David Ricardo (1817) treated the main factors of production—land, labour, and capital—as constituting economic interest groups or classes, between which there was a profound clash of interests.

Marx's analysis. These diverse elements were brought together in Marx's theory of class, which has dominated all later discussion. In the passage quoted earlier, Marx disclaimed any credit for having discovered the existence of classes or the conflict between them; what he had accomplished, he claimed, was to show: "(1) that the *existence of classes* is only bound up with *particular historical phases in the development of production,* (2) that the class struggle necessarily leads to the *dictatorship of the proletariat,* (3) that this dictatorship itself constitutes only the transition to the *abolition of all classes* and to a *classless* society." The starting point of his theory was an elaboration of the distinction between civil society and the state. The anatomy of civil society, according to Marx, was to be found in political economy. What distinguishes one type of society from another is its mode of production (that is to say, the nature of its technology and division of labour), and each mode of production engenders a distinctive class system in which one class controls and directs the process of production while another class is, or other classes are, the direct producers and the providers of services to the dominant class. The relations between the classes are antagonistic since the classes are in conflict over the appropriation of what is produced; and, in certain periods, when the mode of production itself is changing as a result of developments in technology and in the utilization of labour, such conflict becomes extreme and a new class challenges the dominance of the existing rulers of society. The dominant class at any time controls not only material production but also the production of ideas; it thus establishes a particular cultural style and a dominant political doctrine, and its control over society is consolidated in a particular type of political system. The subject classes and especially a class that is growing in strength and influence as a consequence of changes in the mode of production generate political doctrines and movements in opposition to the ruling class.

Marx's view of the inevitability of class antagonism

The theory of class is at the centre of Marx's whole social theory, for it is the social classes formed within a particular mode of production that are regarded as establishing a specific form of state, animating political conflicts, and bringing about major changes in the structure of society.

Marx distinguished several historical types of society: "In broad outline we can designate the Asiatic, the ancient, the feudal, and the modern bourgeois modes of production as progressive epochs in the economic formation of society" (1859). But Marx and the Marxists who followed him did not undertake any detailed historical studies of earlier forms of society, and he was not able to set out a convincing empirical account of the development of social classes over the whole span of human history. He restricted his analysis of class structure mainly to modern capitalist society, and he was concerned above all with the rise of the proletariat, the progress of its struggles against the bourgeoisie, and the conditions arising out of the capitalist system of production that would make possible a successful working class revolution and the inauguration of a new type of society that would be without classes.

Views of Max Weber

Post-Marxist analyses. Later theories of class have been concerned in the main to revise, refute, or provide an alternative to the Marxist theory and, in a similar way, have been concentrated upon the development of classes in modern capitalist societies. In recent years, however, a growing amount of attention has been given to the formation of new classes in the Socialist societies of eastern Europe and in the developing countries. The German sociologist Max Weber was one of the first to undertake a critical revision of the Marxist theory, mainly in his essay on class, status, and party (1922), but, in more general terms, in many of his writings. Weber, in the first place, questioned the importance attributed to social classes in the political development of modern societies and in the causation of major social changes. In his view, capitalism had been brought into existence by the influence of certain ideas and especially by the influence of Protestantism on economic behaviour, as well as by the economic interests of the bourgeoisie. He also argued that the development of the European countries in the late 19th century had been affected by nationalism as much as it had by class ideologies and movements. There was another factor, too, that limited the significance of class as a basis of political action, namely, the existence of social stratification in terms of social honour or prestige, which cut across class divisions and tended to produce a condition of social harmony in which peaceful competition and emulation prevailed, rather than an increasingly intense class struggle. More generally, Weber contested Marx's view of the future development of capitalist society, for, whereas Marx envisaged the successful outcome of the working class struggle as the creation of a liberated classless society, Weber foresaw a continuation of the increasing organization of social life and a monstrous growth of bureaucratic regulation, which would attain its greatest extent in a Socialist society.

From a different aspect, Pareto and, in a more qualified way, Mosca also rejected Marx's division of society into contending classes and propounded instead a theory based upon a distinction between the "governing elite" and the masses. Pareto praised Marx's conception of the class struggle as a major contribution to social theory but then transformed it into the idea of a conflict for the possession of political power between an established elite and a rising elite. This conflict resulted, throughout history, in the substitution of one ruling elite for another, while domination and exploitation of the masses continued; in Pareto's view there is no progressive movement toward a classless society as Marx thought. This notion of the rise and fall of ruling elites is to be found in much modern writing on the class structure; it appears, for example, in the U.S. economist Thorstein Veblen's account (1921) of the rise of the engineers (though, in other respects, Veblen remained closer to the Marxist theory) and in recent discussions of technocracy.

The extent and significance of class conflict

One of the most important issues that has divided social theorists in their analysis of class structure has been, in fact, the assessment of the extent and significance of class conflict. Those who have opposed the Marxist theory most strongly have focussed attention on the functional interdependence of classes and their harmonious collaboration. The sociological theory known as functionalism, which attained its greatest influence in the United States, has developed this theme. Functionalism stands quite clearly in a general opposition to Marxism by virtue of its unhistorical approach, by its emphasis upon the regulation of social life by values that are more or less universally accepted in a given society, and by the importance that it assigns to social integration as against social conflict. Functionalism does not offer, in the strict sense, a theory of class. Its concern is with stratification as it arises from social differentiation and the evaluation and ranking of mainly occupational roles. The functionalist writers, therefore, take up the problem of social status, in the sense in which Max Weber distinguished this from the problem of social class, and they deal with the placement of individuals in the social hierarchy rather than with the formation of distinctive social groups.

Although the attention given to social status grew considerably in the two decades after 1945, partly under the influence of functionalist ideas, problems of class were far from neglected. Indeed, renewed reflection upon the nature and significance of social classes was stimulated by changes in class structure that seemed to be taking place, especially in Western industrial societies.

The theme of the relation between class and politics has figured prominently in most recent discussions. It is argued that in the new "postindustrial" society, which many believe is now emerging, the ruling class is no longer the property-owning bourgeoisie but the technocrats and bureaucrats who direct the process of technological innovation and economic growth; and the opposition to their rule is led not by the working class but by those groups that feel most keenly their dependent position and their exclusion from genuine political participation. Social conflict has therefore assumed a directly political character, instead of arising from economic antagonism.

The Marxist theory of class has thus been questioned in several respects. First, it has been claimed that classes in capitalist societies have tended to lose their distinctive character, while the antagonism between them has declined to such an extent that it no longer produces serious political conflict. From this point of view it may be argued that the class system has lost much of its social significance and that gradations of social status are now more important in determining the actions of individuals and groups. On the other side, however, it is suggested that although the major social classes of 19th-century capitalism have declined, new social classes are being formed in the advanced industrial societies (both capitalist and Socialist) that are beginning to engage in social struggles of a new kind. Furthermore, there are Marxists who would contend that, despite the uncertainties produced by the economic and social changes of the past few decades, the division between bourgeoisie and proletariat remains the fundamental source of conflict in society and continues to be expressed in the ideologies of vigorous social movements. These theoretical disagreements are far from being resolved, but further light may be shed on the problems that they pose by considering some of the ways in which differences between social classes have been characterized.

Characteristics of different social classes. In spite of controversies over the theory of class, there is a large measure of agreement among social scientists on the characteristics of the principal social classes, especially in modern societies. A distinction is habitually made between the upper class, the middle class, and the working class, but this needs to be refined and supplemented in various ways. Some writers have used the term lower class to refer to the working class, but others have distinguished a lower class, or lumpenproletariat, from the main body of the working class. This lower class may comprise particular ethnic groups, such as black and Mexican workers in the United States, and immigrant workers, such as those recruited from former colonial territories and from southern Europe by the industrially advanced European countries, as well as casual and unskilled workers in backward regions and declining industries, all of whom are characterized by exceptionally low levels of living, economic insecurity, and lack of social rights.

Conventional class divisions

It is also necessary to take account of the peasantry as a distinct social class, though it was neglected for a considerable time by Western social scientists. In many European societies there are still a substantial number of peasant cultivators; and in the Socialist countries of eastern Europe, in particular, a description of social classes cannot ignore the development of the peasantry and its relations with other groups in society. The importance of the peasantry has also been more fully recognized in recent years because of the role of political movements based largely upon peasant support in the developing countries. Marx's view of the peasantry as a predominantly conservative element in society and as a group that was unlikely to evolve an independent political consciousness or organization has been modified or abandoned by later Marxists. Radical social thinkers are now more inclined to attribute a major revolutionary role to the peasantry, or to large sections of it, in the countries of the Third World.

The delineation of the principal social classes does not imply that these classes have exactly the same character in all modern societies nor that they are entirely homogeneous groups. Nevertheless, there are broad differences between classes and common characteristics within them that make the distinction valid and useful.

The upper class. The upper class in Western capitalist societies is distinguished above all by the possession of largely inherited wealth. In Great Britain, for example, some 40 percent of all personal wealth is concentrated in the hands of the top 1 percent of property owners; in the United States, where the distribution of property is rather less unequal than in most European countries, the top 1 percent still owns about 30 percent of all personal property. Since the beginning of the 20th century, this concentration of wealth has tended to diminish in most countries, but the wider diffusion of property ownership is still mainly confined to the upper 5 percent of the population. The ownership of large amounts of property and the income derived from it confer many advantages upon the members of the upper class. They are able to develop a distinctive style of life based upon expensive cultural pursuits and leisure activities, from which the great majority of the population is excluded, to exert a considerable influence upon economic policy and political decisions, and to procure for their children a superior education and economic opportunities that help to perpetuate family wealth.

Whether these characteristics make the upper class a ruling class in the Marxist sense is still vigorously debated. The question involves considering how far major political and economic decisions are influenced by the general interests of the upper class and how far this class is preeminent as a source of recruitment of top decision makers. On one side, it is argued that the upper class of large property owners provides the elites—economic, political, military, and intellectual—that determine the main course of a society's development, while, on the other side, it is claimed by the pluralist political thinkers that decision making results from the activities, often conflicting, of many different groups in society, none of which is clearly or permanently predominant.

The "new class" of Socialist societies

In the Socialist societies of eastern Europe, personal wealth obviously has little significance in the formation of a class, but it has been argued by the Yugoslav writer Milovan Djilas that a new upper class, indeed, a ruling class, has emerged in these societies on the basis of the possession of political and administrative power. The members of this class are not able to accumulate personal property on a large scale but can acquire in other ways all kinds of privileges—in housing, education, travel, and consumption—that enable them to develop an exclusive style of life. One difference from Western capitalist societies is that such advantages are less easy to transmit by inheritance than is private property, and for this reason an upper class in the Socialist societies is likely to be less stable in its membership over time. Nevertheless, some advantages may well be transmitted as a result of the privileged access to higher education and to important jobs that the children of upper class families enjoy.

Although the upper class in most societies shows a considerable degree of continuity in its membership from generation to generation, changes do occur through the rise and fall of individuals and groups as a result of changes in the social structure, the occurrence of exceptional ability or lack of ability in particular individuals, or by sheer chance. Such changes are particularly evident in many developing countries, and there may be considerable variations in the composition of the upper class. In some countries (as, for example, in Latin America or in the Middle East) large landowners may still form an important element in the upper class, though their preeminence is being challenged by the rise of a business class. Elsewhere, in Asia and Africa, nationalist political leaders who rose to prominence in the course of struggles against colonial rule and high officials or military leaders may form the nucleus of a new upper class, while in those countries in which revolutions have occurred, as in China, the new elites may be largely recruited from the working class and peasantry. They may tend, as in the countries of eastern Europe, to draw together in the formation of a new upper class, although some aspects of the Chinese cultural revolution seem to have been directed against such a development.

The working class. The principal contrast with the upper class in the industrial societies is provided by the working class, constituted essentially by manual workers in extractive and manufacturing industry. What characterizes the working class as a whole is lack of property and dependence upon wages. With this condition are associated relatively low levels of living and of education, restricted access to secondary and higher education, limited opportunities for leisure and cultural activities, and exclusion, to a large extent, from the spheres of important decision making. There are, of course, considerable differences within the working class, and a distinction has commonly been made between skilled, semiskilled, and unskilled workers, corresponding broadly with differences in income level. More recently, social scientists have emphasized the difference between the relatively prosperous workers in modern expanding industries and the less prosperous workers in some of the older industries.

There have been significant changes in the condition of the working class over the past few decades. Levels of living have risen, although as a consequence of general economic growth rather than any major redistribution of wealth and income; and access to education has improved, though it is still the case, in most countries, that a very small proportion of working class children enter higher education and hence make their way into the occupational and social elites. At the same time, the economic security of most workers has been increased by policies of full employment and by more extensive welfare services. These changes were widely interpreted during the 1950s as the *embourgeoisement* of the working class—that is, as an erosion of its distinctive characteristics and its gradual assimilation into the middle class. Such accounts have, however, been examined more critically in recent years, and the extent to which the working class remains socially, culturally, and politically distinct from other groups in society has been emphasized. The idea that the advanced industrial societies were becoming "middle class societies" and that this condition had already been more or less attained in the United States, which developed from the *embourgeoisement* thesis, was also sustained by the general shift in the economy from manufacturing to service industries, which involved a relative contraction in the numbers of manual workers and an increase in the numbers of clerical, technical, and professional workers. Thus, although there are still disagreements among social scientists about the size of the working class and the general features of its development, it can scarcely be denied that some important changes have occurred that require, at the least, a reconsideration of the Marxist theory of its political role.

The *embourgeoisement* thesis

It is evident that the position of the working class varies quite widely from one country to another. In some European countries, notably in France and Italy, a large part of the working class is strongly influenced by Marxist ideas and supports political parties that aim to bring about revolutionary changes in society, whereas in other countries—in Great Britain, West Germany, and Sweden, for example—the main working class parties are reformist in their policies. In the United States, on the other hand, there has been no major independent political movement of the working class since World War I, and class consciousness has had much less influence in political life than has been the case in Europe. These differences cannot easily be explained in general terms, but it has been observed that working class consciousness was more intense and took a more revolutionary form in those European societies in which the development of capitalism was slower and more backward than elsewhere; and this observation would apply also to Russia before 1917. In the case of the United States, the failure of a political movement of the working class to develop on anything like the European scale has been accounted for in various ways: by the exceptional opportunities for mobility (or at least the strong and pervasive belief that such opportunities existed), by the fragmentation and disruption resulting from massive immigration, by the existence of a large black lumpenproletariat, and by the divisions in the working class movement that were produced by World War I and the Russian Revolution. At all events, it is clear that the working class in different countries has assumed different characteristics and has espoused very diverse ideologies.

The middle class. While the upper class and the working class can be viewed as relatively homogeneous groups quite sharply differentiated from each other, the middle class often has been treated as a more varied, residual category, comprising those groups, mainly defined in terms of occupation, that do not clearly belong to one of the other classes. Marx, for example, referred to the "intermediate and transitional strata" that obscured the boundaries of the principal classes. Generally speaking, the middle class has been taken to include the various levels of clerical workers, those engaged in technical and professional occupations, supervisors and managers, and self-employed workers, such as small shopkeepers, farmers, and (in some societies) the wealthier peasants. At the top—in the case of wealthy professional men or managers in large corporations, for example—the middle class merges into the upper class, while at the bottom—where routine and poorly paid jobs in sales, distribution, and transport are concerned—it merges into the working class. When a contrast is drawn between middle-class and working class styles of life it is often based implicitly upon "ideal types": the middle-class individual is conceived as a professional or senior clerical worker, the working class individual as a factory worker.

Middle-class divergencies in social and political outlook

To some extent, these distinctions within the middle class are recognized by references to the upper middle class and lower middle class, but there are other differences that do not coincide exactly with this division. There is, for example, considerable diversity of social and political outlook in the middle class: in Western societies the predominant political attitudes may range between liberal and conservative, but in some conditions a part of the middle class, at least, may adopt extreme right-wing views (as in its support for National Socialism in Germany); and, on the other hand, there is a well-known phenomenon of middle-class radicalism, which became more prominent during the 1960s with the development of the largely middle-class student movement. Other distinctions arise from longer term historical changes in the composition and social situation of the middle class. The "old" middle class, with which Marx and other 19th-century writers were largely concerned, consisted chiefly of small independent producers and independent professional men; the "new" middle class of the late 20th century is made up primarily of employees in public services and in large business enterprises at various levels of clerical, technical, and professional work. This transformation seems to have produced a greater diversity, fluidity, and uncertainty in the cultural and political character of the middle class, so that to speak of the advent of middle-class societies, as the numbers of the new middle class increase, may not provide any clear indication of the form that such societies will eventually take.

The industrial countries resemble each other in their occupational structure, and the development of occupations under the influence of technological innovation and economic growth is also similar, but this does not prevent considerable differences in the class structure. There is, as has been noted, an important distinction to be made between the Western capitalist countries and the Socialist countries of eastern Europe in respect to the character of the upper class or elites, and other classes in these two types of society may also show important differences, especially in their social outlook and political involvement.

International comparisons. It is difficult to make rigorous comparisons of the character of different classes in different countries, since comprehensive studies of the class system are lacking, especially in the Socialist countries, though a major study in Czechoslovakia (1969) provides much valuable information. Even within Western societies there is much diversity, and the degree of class consciousness and class conflict, in particular, varies widely from one country to another, as well as fluctuating over time. One observer, for example, wrote in 1914 of the first decade of the 20th century that it was a turning point in American society, when deep class feeling found expression in both political parties and brought to public attention the existence of a profound conflict between great wealth and the lower middle and working classes; but the feeling, and the conflict, subsided again, and in spite of a limited revival of class consciousness in the 1930s the general character of American society has been one of ideological classlessness. In most of the European countries, though, class consciousness has found expression in distinctive parties, and class conflict has been overt.

Classes in the developing countries

In the developing countries, modern social classes exist, for the most part, only in embryonic form. It has been observed already that the upper class in these societies may be far from homogeneous or stable and may include such disparate elements as feudal landowners, capitalist entrepreneurs, nationalist or revolutionary leaders, higher castes, and military chiefs. In those countries in which industrialization is not far advanced, the urban working class will be small, and, if the workers have only recently migrated to the towns, as is the case in many African and Asian countries, there may be little development of class consciousness. Similarly, the middle class is likely to be small, although some sections of it may be particularly privileged and influential by virtue of the educational level of their members, who may form part of a Western educated elite. Numerically, the most important class in African and Asian societies is the peasantry, and, even in the more urban and industrial societies of Latin America, this class is still relatively large. Consequently, peasant organizations and movements are of great political significance, and major social conflicts are likely to occur over issues of landholdings, land reform, and agricultural prices, rather than over industrial conditions. As industrialization proceeds, the occupational structure of the developing countries will come to resemble more closely that of the industrial countries, but the class structure may take various forms, depending upon historical traditions and political changes. Many developing countries aim deliberately to bring about a condition of much greater economic and social equality, and it has still to be seen what kind of class system, or classlessness, will emerge from the profound changes that are occurring in such societies as India, China, Tanzania, and Chile.

Social mobility. The term social mobility, in its widest sense, refers to any movement of individuals, families, or groups between different sectors of society. Thus movement from one occupation to another, or from one region of a country to another, is a kind of mobility. On an international scale, migration is a very important type of mobility.

Types of mobility. A distinction is usually made between "horizontal" and "vertical" social mobility, the former involving no change in the position of the individual or group in the social hierarchy, while the latter does involve a change of social level. When an industrial worker moves from one factory to another or a manager takes a position in another company, there is no significant change in his social status or class membership. But, if an industrial worker or the child of an industrial worker becomes a wealthy businessman or lawyer, he has changed quite radically his position in the class system. Vertical mobility may, of course, involve either upward or downward movement. A ruined aristocrat, or a member of an upper class dispossessed of his wealth in a revolution and obliged to enter a manual occupation, has experienced a change of social level that may affect his whole life or particular aspects of it. Modern sociologists have concentrated their attention mainly on upward social mobility, because they have been preoccupied with the question of equality of opportunity in the relatively open and democratic industrial societies; but it has been suggested more recently that the degree of downward mobility might be a better indicator of the "openness" of a society, since it would show the extent to which it is relatively easy or difficult for privileged individuals and groups to maintain and transmit to their descendants the advantages that they enjoy.

There has also been a marked preoccupation with the mobility of individuals, although it is evident that whole families, groups, and even classes may at certain times change their position in the social structure. One of the few writers to bring out the diverse aspects of vertical mobility was the economist Joseph Schumpeter (1927), who analyzed and illustrated what he termed the "rise and fall" of individuals, families, and whole classes within the class system. But Pareto, too, in his examination of the "circulation of elites" (1915–19), paid attention both to the movement of individuals into and out of the elites, and to the rise and fall of whole elite groups.

Intragenerational and intergenerational mobility

In studying the movement of individuals, sociologists have distinguished between intragenerational and intergenerational mobility; that is to say, between the mobility of an individual within his own adult lifetime and the mobility represented by a change in social level from the parental to the filial generation (almost always from father to son). Comprehensive national studies of mobility that have been undertaken in recent years have dealt almost entirely with intergenerational mobility, examining changes in occupation between father and son in terms of a scale of occupational prestige. At the same time, such studies have concentrated heavily upon educational opportunity as a major factor influencing upward mobility.

The rise and fall of families, groups, and classes is less easy to study in a quantitative manner, but such changes can be documented in other ways. The emergence of new ruling dynasties is one example of the upward mobility of families, and other examples can be found in the rise and fall of family businesses, of politically influential families, or of intellectual families. The mobility of particular social groups may occur as a result of economic, political, or cultural influences.

There have been important changes in the position of various groups in the modern industrial societies: the prestige and influence of scientists has increased, while that of clergymen has diminished; bureaucrats have become a much more important group, as have the officials of a dominant political party in some societies. Finally, the rise and fall of whole classes is a phenomenon that occurs when there are substantial changes in the whole structure of society. In the Marxist theory, such changes take the form of a social revolution, in which the established ruling class is overthrown by a new class, as in the rise to power of the bourgeoisie with the development of modern capitalism. But a change in the relative position of classes may also occur in a more gradual way, and an example may perhaps be found in the growth of the new middle class in present-day industrial societies.

Vertical social mobility is a feature of all societies, except perhaps the most simple and primitive. In European feudal societies, there were opportunities for upward mobility, one avenue being through the church, and these opportunities increased with the revival of the towns and the growth of trade. One study of the circulation of elites in France between the 11th and 18th centuries gives examples of the rise and fall of particular individuals and families, as well as tracing the fortunes of such groups as lawyers, financiers, and traders.

Inadequacy of historical studies of mobility

The problem posed by such historical studies is that they largely provide illustrations and, for lack of data, do not show the degree of mobility; that is to say, the numbers of individuals or families who actually changed their position in the social hierarchy over a given period. Historical comparisons of mobility are thus difficult and largely speculative, and there are many problems even in examining the trends of mobility in the industrial societies in recent times. It is nevertheless generally considered that vertical social mobility (especially of individuals) is greater in modern industrial societies than in earlier societies. Modern studies suggest, however, that, even in industrial societies, mobility is limited and that there are no substantial differences between industrial societies in the amount of mobility, even though, for example, it has been generally believed that social mobility is greater in the United States than in European countries. In particular, the movement of individuals from the working class into the upper class is rare in all societies. Sociologists have, therefore, introduced another distinction between short-range and long-range vertical mobility; and comparisons between industrial societies show that the greater part of mobility is short-range.

These modern studies suggest a cautious interpretation of historical accounts of social mobility, which may give undue prominence to exceptional cases of upward or downward movement. It is likely that in all societies, at most times, there is a considerable stability of class membership that, even when not maintained by any formal sanctions, is ensured by the inheritance of property, educational advantages, or political influences.

The process of mobility. The nature and degree of vertical mobility in a society is affected by a variety of influences. One universal factor is the occurrence, in any population, of individuals who are exceptionally endowed—with intelligence, physical strength, beauty, skill in warfare or business, or the quality of charisma that enables them to become religious or political leaders. Beautiful women have risen to social eminence as the mistresses of kings and nobles and as film stars, just as men (and to a much lesser extent women) have risen by the accumulation of wealth, the attainment of political or military power, and intellectual or artistic achievement. But the manifestation of such personal qualities and the advantages that they bring are limited and controlled by many social factors.

In the first place, the open or closed nature of the class system has a powerful influence. It is not only that the individual in a closed system will encounter great practical obstacles if he seeks to escape from his situation as a slave, serf, or member of a lower caste; the ideology that upholds such a system and emphasizes the importance of everyone "knowing his place" tends to inhibit the development of personal talent and ambition at the lower levels of society. In the more open class systems of modern societies, on the contrary, there are no formal restrictions upon vertical mobility, and the ideas of equality and reward for merit, which have gradually entered the general social consciousness since they were formulated in the 18th-century revolutions, encourage an individualistic striving for success throughout large sections of society. Even in these societies, however, talented individuals in a lower social class have to surmount many obstacles, arising mainly from poverty and the difficulties of access to education. On the other side, less talented individuals are able to maintain their position in the upper class because of their inherited social advantages.

The extent to which individual talent will result in upward mobility is also limited by the general orientation of a society's activities. A tribal society that lives by hunting or is engaged in continual warfare will place a high value on physical strength and agility; a nation engaged in imperial expansion is likely to rate military qualities highly; and one that is chiefly concerned with industrial development may attach the greatest importance to business ability. The role of science and technology in present-day industrial societies makes certain kinds of intellectual ability (and the selection and training of such ability) particularly important, so that some writers have referred to the rise of a "meritocracy" or a "scientific and educational estate" or a "technocratic elite."

Mobility and changes in social structure

One of the most profound influences upon mobility comes from the general changes in social structure. A revolution that dispossesses an existing upper class or a national liberation movement that overthrows foreign rule creates opportunities for other individuals, groups, and whole classes to establish themselves in a dominant position. In the early years of the Socialist regimes in eastern Europe, for example, the universities were opened to young people from working class and peasant families, while restrictions were placed upon the entry of those from the middle class. More gradual changes, especially in the occupational structure, also affect mobility. The expansion of clerical, technical, and professional employment involves a continuing movement out of manual work and accounts for a large part of the upward mobility in Western industrial societies. In earlier periods economic changes had similar effects; the revival of trade and the growth of towns in European feudal societies made it possible for serfs to embark upon new occupations and to escape the constraints of the feudal system. A similar movement from country to town is occurring in the present-day developing countries and involves, in most cases, upward mobility in terms of rewards, opportunities, and social status. There are also movements from poorer to wealthier regions in a particular country. The postwar movement in Italy from the relatively poor agrarian south to the more prosperous industrial north and the movement of black Americans from the South to industrial areas provide examples.

After a revolution a new system of stratification may emerge, and the rate of social mobility may decline again. Similarly, if there is little economic development or the rate of growth slows down, opportunities for mobility will be diminished. In such conditions, the possibilities of upward mobility will depend largely upon the extent of downward mobility. One other factor, however, may influence this situation. If the upper and middle classes practice family limitation to such an extent that they do not reproduce themselves from generation to generation, the vacant places may be filled by individuals who rise from lower classes. Such differential fertility, however, has rarely, if ever, been a major influence, and most analysts believe that it has been declining in the developed industrial societies during the past few decades.

Mobility and migration

International movements of population have been important at various times in promoting upward mobility. The colonial expansion of the western European countries from the 16th century on provided opportunities for individuals to enrich themselves as traders or settlers, at the same time as they subjugated and enslaved other peoples in Asia, Africa, and America. Later, with the creation by European settlers of new societies, especially in North America, fresh opportunities for mobility were provided by large-scale immigration. The expansion of U.S. society across the continent during the 19th century brought millions of working class and peasant immigrants from Europe who were leaving conditions of poverty and subordination that seemed unalterable in their own societies.

Consequences of mobility. The determinants of mobility outlined above are difficult to isolate and measure, as are the social consequences of mobility. On one side, it may be said that the movement of large numbers of people up and down the social hierarchy tends to break down the exclusiveness of social classes and to create a more uniform national culture. Presumably, this would also lead to a diminution of class prejudices and class conflicts. It has often been claimed, in fact, that the allegedly lesser degree of class consciousness in U.S. society, as compared with European societies, is due to a higher rate of mobility in the former. Although this may be doubtful in the light of recent studies that show no great differences in overall mobility in the industrial societies, the widespread belief that opportunities for upward mobility are actually greater than in other societies may itself have had an important influence. From another aspect, however, it can be argued that the preoccupation with vertical mobility reinforces the class system; the individual who is concerned to rise or avoid falling in the social hierarchy accepts and indeed emphasizes the importance of class and status distinctions. It has been suggested that these distinctions might be diminished much more by the attainment of greater economic and social equality than by any amount of individual mobility.

Another consequence of mobility that is seen as beneficial from the point of view of society as a whole is the more effective use of ability. If individuals are confined to the social sphere in which they are born, many useful talents will remain undiscovered and unused. The expansion of education in modern industrial societies has been stimulated by the desire to provide opportunities for the development of all the abilities available in the population, though this is still very imperfectly achieved.

Possible undesirable consequences of mobility

On the other hand, vertical mobility may have some undesirable consequences. Such mobility, whether upward or downward, imposes strains upon the individual striving for success and adapting to new social milieus and may be disruptive of families and local communities. More generally, a high rate of mobility may be regarded as producing in society the condition that the French sociologist Émile Durkheim called "anomie" (normlessness and resultant disorientation and anxiety), in which there is insufficient regulation of behaviour and the individual suffers from the "malady of infinite aspiration." The existence of such strains may lead to a higher incidence of mental illness among highly mobile individuals, but it seems doubtful that mobility is a major factor in mental illness, given that such illness has been increasing rapidly in the industrial societies, while there has been no correspondingly marked increase in mobility. It seems likely that it is the pace of economic and cultural change in general and the conditions of urban-industrial life that impose the greatest strain upon individuals. (T.B.B.)

SOCIAL STRATIFICATION: CASTE

Caste systems are moral systems that differentiate and rank the whole population of a society in corporate units (castes) generally defined by descent, marriage, and occupation. Elaborately differentiated and ranked caste systems have developed especially in the regional societies of India and among adjacent Hindu and related populations in the territories of modern Pakistan, Bangladesh, Nepal, and Sri Lanka (Ceylon) over the past 2,000 years. Simpler caste systems have developed elsewhere. The nature, history, and variety of these systems, their significance for general understanding of human society and for particular understanding of South Asian society and religion are the subjects of this section.

So intricate are the connections of caste systems in South Asia with regional systems of ideas regarding such matters as genetics, physiology, kinship, ethics, and cosmogony that they have been equated by some scholars with the whole culture of Hinduism and declared to be unique phenomena inseparable from it. Caste systems are found, however, in combination with other religions and cultures in South Asia and elsewhere. Their features as a general type may be compared with other systems of social class or stratification.

Caste systems compared to other systems of social stratification

Caste systems resemble racial stratification in their biological concern with differences of birth and marriage; they resemble stratified plural societies in their presumption of profound differences in group behaviour; they depart from both in conceiving of themselves simultaneously as unitary societies that are culturally integrated. Unlike racially or culturally plural societies, where prior intractable difference among groups is taken as a moral reason for the divided constitution of the society, established caste systems may differentiate into further, new units or they may reassemble their units into one.

Caste systems resemble the much more widespread systems of social and economic classes in containing ranks that tend to be culturally marked, occupationally linked, hereditary, and endogamous (marrying within a group). They differ from such systems in having as their ranked units not only individual persons and families but also larger corporate groups. Corporate caste units differ from social classes in being necessarily defined—rather than incidentally distinguished—by occupation, descent, and marriage. Castes have no other necessary cultural markings, although classes, being assemblages of persons on the basis of their attributes, necessarily do. The ranking of corporate units in caste systems creates an illusion of individual immobility (a feature sometimes taken as typical of caste systems); yet rates of individual rise and fall are not known to differ between class systems and caste systems, as the types are here defined.

Caste systems are comparatively rare among large-scale moral systems in making the differentiation and interrelations of corporate groups into major concerns; but in this they are at one with a large and diverse category of small-scale societies. The sharings and exchanges by which caste systems reckon the boundaries, alliances, and ranks of their corporate units have much in common with widespread forms of marriage and parent–child relationships, gift giving and tribute, ritual, sports competitions, communication networks, and dominance in animal as well as human societies. But caste systems are unique in their ways of combining these features—in having biological and behavioral schemes for differentiating and ranking occupational and ethnic units according to the determined way they relate to each other in sharings and exchanges. They have proven uniquely capable, in each of the regional societies of South Asia, of relating hundreds of units in dozens of higher and lower ranks and of generating new societal forms through combinations and transformations of existing schemes.

Theories of caste. For about two centuries, the explanation and analysis of South Asian caste systems have been matters of interest and controversy among foreign observers, who have brought with them assumptions and sensitivities peculiar to their own respective societies and times.

No visitor of ancient or medieval times seems to have felt that rank, corporateness, heredity, and endogamy among the occupational and ethnic groups of India were worthy of special mention. The Portuguese seafarers who traded mainly on the west coast of India in the 16th and 17th centuries described groups they called *castas* (from which derive the English and French words *caste*), meaning "species" or "breeds" of animals or plants and "tribes," "races," "clans," or "lineages" among men. Only in the 18th century did the distinction that is now made between unranked tribes and ranked castes begin to be drawn clearly in the usage of the French and English.

Attempts by Western observers to explain castes

Consideration of the thoroughgoing inequality and exclusiveness of the castes, as well as the peculiar occupational division of labour prescribed for them, generally produced a sense of shock in Western missionaries and officials of the turbulent 19th and 20th centuries. Attempts were made to explain the castes as an imagined rational legislation by some unknown past authority, as a self-serving invention by the system's top-ranking priestly castes, or (after Charles Darwin) as the unplanned evolution of successively higher castes through the growth of superior technological specialization.

The insufficiency of these—or, indeed, of any—simple conjectures for explaining the origins of these apparently irrational systems strengthened a trend toward more complex historical speculation. Castes were seen as heirs of the Indo-European language family's common traditions of private kin-group ritual by Émile Senart, a late-19th-century Sanskrit scholar. In 1891 H.H. Risley, a British anthropologist, reported that proportionally wider noses correlate with lower caste rank and also with a geographic gradation from northern to southern India; this suggested a hypothesis of repeated invasion from Central Asia and from the north to the south by longer nosed racial elements. Other official British Indian research stressed separate surveys of the physical traits and customs of each of the thousands of castes and tribes. Compendious reports of such surveys reinforced a conception of Indian society as a geographical accident and a political composite. Many authors speculated on historical factors favouring the development of the Indian caste systems—such factors as diffusions of primitive customs, ancient movements of peoples, and the mutually distancing adaptations of allegedly alien peoples to each other—and these speculations were compiled in a list of 15 factors by J.H. Hutton, census commissioner and ethnologist, at the end of the British regime in 1946.

What was missing from these atomistic historical analyses was any conception of a caste system as an integrated whole. Such conceptions were provided in nationalistic political contexts in the late 19th and early 20th centuries. Max Weber, a German sociologist, emphasized the connections of caste with Hindu religion as part of his larger effort to reveal why the West had progressed more rapidly than Asia. Skeptical toward Hindu religious explanations, the French sociologist Célestin Bouglé in 1908 postulated, from descriptions of the relations of the castes, an underlying set of social structural "ideas"—hereditary specialization, rank, and mutual repulsion concerned with purity and impurity—governing the whole system. Hutton held that social differentiation in caste systems requires an assumption of a once widespread preliterate philosophy concerning the powers of bodily and other substances and the dangers of magical contagion. The British social anthropologist A.M. Hocart, combining observations in Sri Lanka (Ceylon) and Polynesia with Vedic (early Indian-language) texts, saw caste systems as organizations for royal rituals of worship. In 1952 M.N. Srinivas, an Indian anthropologist working in southern India, attempted to derive both the unity and the ranks of Indian caste systems from the imitation by diverse groups of the customs of the highest Hindu castes, a process that he called "sanskritization."

Western social values reflected in explanations

Western comparative social theorists have generally placed caste systems not in a unique religious ideological or structural category by themselves but at the end of a theoretical series of types beyond medieval and ancient Europe and furthest from the equalitarian, competitive tendencies of the modern West in matters of social differentiation and social stratification. The enclosure of individuals in culturally diverse ranked ethnic groups (castes) was seen as poles away from a culturally homogenizing and individualizing modern society. The multiplicity of the castes and their apparent social distance and segregation from each other through endogamy and restrictions on contact seemed to exceed in stringency any but the most extreme differentiations (*e.g.,* Black and white in the U.S.) among ethnic groups of a single society in the West. Castes were also commonly regarded in Western comparative perspective simultaneously as social-status groups and as economic classes. They were considered rigidly closed classes, permitting recruitment of individuals into caste membership only by birth, not by achievement. Caste systems were further mistakenly thought to allow vertical mobility only through movement by the whole corporate unit, not through individual or familial movement. They were seen as guaranteeing the virtual immobility of groups, as well, owing either to the preponderant power of the conservative high-caste elite or to the strength of cultural consensus and traditional sacred sanctions or to both.

Although comparative sociology thus attempted to look at South Asian caste systems as extreme instances of presumed universal features of human society and implied a view of them as social systems rather than as accidental collections of ideas and peoples, its typifications of these systems remained sociocentric, reflecting modern Western society's assumptions about essential elements and processes, while reversing some cherished Western values. It did not attempt to base its understanding closely upon the cognitive assumptions actually prevalent in South Asia. For understanding of this sort, one must examine certain Indian social concepts that are known first from Vedic texts from 1000 BC. These are maintained to the present by the predominant Hindu population and are widely shared among persons of all religions residing in South Asia.

Jāti

Indian social concepts. The organization of South Asian society is premised on the ancient and continuing cultural assumption that all living beings are differentiated into genera, or classes, each of which is thought to possess a defining coded substance. One of the commonest words for genus in most Indian languages, *jāti,* is derived from an Indo-European verbal root meaning "genesis," "origin," or "birth." It is applied to any species of living things, including gods and humans. Among humans, *jāti* can designate a distinct sex, a race, a caste, or a tribe; a family, a lineage, or a clan; an ethnic group, a regional population, the followers of an occupation or a religion, or a nation.

Every human genus (and therefore every caste) is thought to have as the shared or corporate property of its members a particular substance (*e.g., śarīra,* "body," *rakta,* "blood") embodying its code for conduct (*dharma*). Each caste's inborn code enjoins it to maintain its substance and morality, its particular occupation, and its correct exchanges with other castes. Indian thought does not separate "nature" and "morality" or "law," so that castes are, in Western terms, at once "natural" and "moral" units of society. These units make up a single order, one that is profoundly particularized.

The persons of each caste are believed to transmit the particles of substance that are peculiar to their caste from one generation to the next through a series of natural and moral acts extending from birth to marriage. The male and female sexes, like castes, are thought to be composed of nonequivalent, complementarily coded bodily substances; the reproductive essence of each, variously conceptualized but often identified with semen and uterine blood, is required to make the child complete from birth in his caste's particular nature. Acts that wrongly mix bodily substances, such as improper procreation, are thought to alter the embodied morality of the caste and its offspring. Each caste must therefore be concerned with regulating marriage.

Castes are related to each other externally not by hereditarily shared substance but by exchange of nonhereditary substances. Each caste is involved through its existence and its means of subsistence in transformations and exchanges of transformed natural substances. Of these substances, a caste's own bodily substances (*e.g.,* hair, blood, semen, and feces) and food (which is seen as becoming transformed into human bodily substance) are considered capable of creating, expressing, or altering relationships of rank and solidarity between it and other castes. The substances derived from each caste's pursuit of its occupation are also major media, along with services, of intercaste exchange. Each caste is thought to maintain or alter its moral standing (that is, its rank with respect to other castes), its hereditary substance, and its natural code according to the way that it receives or refuses to receive bodily substances, consumes or refuses to consume food, and gives or refuses to give services in exchanges with particular other castes.

The Vedic heritage. A conception of Indian society in terms of genus (*varṇa,* here synonymous with *jāti*) appeared first in the Rigveda (Ṛgveda), the oldest collection of Vedic priestly hymns, composed in Punjab about 1200–1000 BC. The Vedic people of that time saw themselves as members of a single "genus of respectable men" (*ārya-varṇa*), "the humans" (*manuṣa*). Their code enjoined them to offer sacrifice (*yajña*) to members of another genus of beings, the Vedic gods (*devas*); it also enjoined them to smite and enslave their enemies, "the nonhumans" (*dasyu-varṇa*), who were naturally and morally fit for service.

The conception of society as containing more than two *varṇa*s appeared only after 1000 BC, when descendants of these early Aryans began to settle in the upper Ganges region and built a more complex society containing towns and the beginnings of states. It appeared then in the increasingly elaborate body of Vedic sacred texts and ultimately reached its culmination in the classical moral code book (*Dharma-śāstra*) of the Manu school (*c.* 200 BC–AD 200).

Division of Puruṣa into specialized *varṇa*s

The whole later Vedic society continued to think of itself as sharing one original natural substance, conceived as the body of Puruṣa, the original code man. The division of Puruṣa by the gods into distinct specialized *varṇa*s mythically inaugurated the Vedic sacrifice. Henceforth, only those who had the substance of Puruṣa in their bodies were to follow the code of exchanges required by the sacrifice.

Sacrifice was seen as fundamental for upholding the natural and moral order of the cosmos. It required exchanges between the two specialized genera, gods and men. Gods, the beings who wield divine power (*brahman*), were seen as the ultimate source of bodily existence and well-being; men were the givers of the food that the gods needed to sustain themselves. Mutual exchange of food through the sacrifice thus created solidarity between gods and men.

Exchange in the sacrifice also created a ranked relationship between gods and men. Only pure food was fed to gods by men, while gods, after they had eaten, returned only their leavings for men to eat. These same asymmetrical events that created rank—that elevated the gods and lowered men—were conversely interpreted as expressions of rank previously established. Thus, men regarded their own food leavings as contaminated by saliva and therefore as impure and valueless to others but considered the leavings of the gods as transvalued—pure and incomparably valuable precisely because they were filled with the divine natural substance that brought human well-being.

Rank in this ancient South Asian moral order was thought of as being based less upon established possessions than upon generosity. Gods were of higher rank than men not so much because they possessed the attribute of divine power as because they were incomparably generous in bestowing that power on men. Men were of lower rank because they could not return this gift. They could give the gods the most valued foods, but they could not equal the value of the divine power contained in the gods' gift to men.

The same relationship existed among the specialized natural genera within the society of men—between the "once-born" and the genera of the "twice-born" (*dvi-jāti*), those who became divinized humans through a second, ritual birth. The once-born, called Śūdra (Servant), was not to hear the sacred Vedic words or to sacrifice directly or to receive the gods' leavings. He was instead to exchange with the twice-born in the same way that the twice-born, to whom the term *ārya* was exclusively applied, exchanged with the gods. Born of the lowest part of Puruṣa, the feet, the Śūdra followed a code enjoining him to serve the twice-born in exchange for maintenance. He could symbolize his ranked relationship, as was done in the sacrifice, by feeding pure food to his human superiors and eating their leavings in return. His service was a form of worship but an inferior one that properly brought only the divine gift of maintenance in return. Because the Śūdra was defined as being unable to return even this gift of maintenance, his moral rank was considered to be lower than the rank of the twice-born.

The twice-born was the divine human genus of Aryans, whose males came to share the divine sounds of the Veda through a second, divine birth and initiation into Vedic study and sacrifice. Their code also required married male householders to exchange gifts among themselves. The wife and children of a twice-born householder were forbidden, as was the Śūdra, from hearing Vedic sounds and from carrying on the sacrifice. But women and children were conceived as sharing the effects of those duties through having bodily substance in common with their household head.

The three types of twice-born persons

The code book enjoining sacrifice and exchange also defined three specialized genera of people within the twice-born genus. The substance of each of these had its distinctive ranked origin and inherent ranked powers, and the code for conduct of each was specialized. The Brāhmaṇa, born from the highest part of Puruṣa, his mouth, possessed godly power (*brahman*): he was to teach the Vedas, perform sacrifices for the Kṣatriya and Vaiśya, and accept gifts from them in exchange. The Kṣatriya, born from the arms of Puruṣa, possessed royal power (*kṣatra*): he was to fight enemies, give gifts and food to the Brāhmaṇa, and protect the Vaiśya; in exchange, he received a share in the leavings of the sacrifice from the Brāhmaṇa and wealth from the Vaiśya. The Vaiśya, born from the thighs of Puruṣa, possessed productive power (*viś*): he was to produce wealth for the Brāhmaṇa and the Kṣatriya through agriculture, commerce, and animal herding, and was to give a share of it as taxes to the Kṣatriya in exchange for protection.

Each superior genus is considered divine in relation to each inferior genus. The moral ranks of the genera were gauged in *varṇa* theory also by the values assigned to the different things they gave each other. Since the wealth given by the Vaiśya to the Kṣatriya was considered less valuable than the divine protection received in exchange, the Vaiśya ranked below the Kṣatriya. The Brāhmaṇa, acting as representative of the whole society, fed the gods through the sacrifice and was first to eat their leavings, to take their power. In comparison to the Kṣatriya and the Vaiśya, he was therefore a "god on earth." Because the gifts and food given by the Kṣatriya and the Vaiśya to the Brāhmaṇa were considered to be less in value than the remainder of the sacrificial leavings that he gave to them, he was thought to have a higher rank than they. Beneath all three was the Śūdra, with his offering of mere labour, benefitting by the activities of all.

The four *varṇa*s of Vedic theory may have had a waning existence in parts of northern India when the theory was first being formulated, between 1000 and 300 BC. But the authors of the classical moral code books in the centuries following (*c.* 200 BC–AD 200) were concerned with reconciling this theory with the multitude of specialized *jāti*s of uncertain *varṇa* that they actually saw about them. They explained the existence of non-Vedic *jāti*s by a theory of miscegenation: The *jāti*s were offspring of improper mixing and remixing among descendants of the four pure Vedic *varṇa*s. The numerous *jāti* units, their varied codes for conduct, and many ranks were all seen in the Vedic scheme itself as responses to the requirements of particular places, times, and genera of persons. The four *varṇa*s together with the sacrifice from which they were thought to spring were seen as the universal, eternal, and logically complete scheme for upholding nature and morality.

Hindu worship and the castes. The Vedic scheme continues to the present as a template, generating systematic conceptions of the local, everchanging castes (*jāti*s) and their exchanges in the closely connected scheme of Hindu image worship, called *pūjā* (literally, "respect"). Evidences of the proliferous systems of castes connected with image worship appeared gradually after AD 500.

Complex systems of castes required by Hindu image worship

The Hindu gods to be worshipped are related among themselves, like men, by shared and exchanged natural substance. They have not merely abstract qualities and representations like the Vedic gods, but also particular life-like images, biographies, bodily functions, and specialized relations to men. They are attached by particular codes of worship to particular occasions, communities, and genera of persons. The occasions for their worship are times of bodily events in their own or their worshippers' lives—births, weddings, deaths, and the routines of such things as the daily bath or meals—and in the natural cycles of the Sun, Moon, Earth, and other planets, whose heat and cold, dark and light, wet and dry, etc., are seen as bodily states affecting human beings.

The codes of Hindu worship require the existence of complex local communities of castes. Worship cannot proceed without a priest to bring the living substance of the god into the image, made by an image maker. The priest must utter formulas (*mantras*) that activate this divine life substance and other formulas that have the power to transform the ordinary materials of worship into sublime offerings. The priest must be a male of the highest, most godlike caste available—ideally a Brahmin skilled by heredity in the maintenance of ritual boundaries between substances, and empowered to transform them.

To sponsor the worship there must be a local worshipper of means, typically a ruler or man of wealth, who can by gifts entreat a priest to mediate with the god. There must be specialists of appropriate castes—*e.g.,* temple keeper, garland maker, cook, sweet maker, singer, musician, dancer—to feed, attend, and entertain the god. Before the worshipper can approach the god, he must prepare himself and his caste to be as godlike as possible and must remove as much as possible from his person and his caste any insulting, transmissible bodily substance. He does so through bathing and through engaging the services practiced by other castes, such as those who do the work of barber, washerman, midwife, funeral priest, leatherworker, scavenger, or sweeper. These castes are by their intimate receivings of bodily substance rendered subordinate to the worshipper, as a child's receiving of bodily substance renders him subordinate to his parents. Each caste contributes in its particular way to the worship, directly or indirectly, and each then receives in return and in order of rank a share, either in the transvalued leavings of the god's food (called *prasāda,* "favour") or in the leavings of those who do receive a share.

Hindu worship thus greatly ramifies the specialities assigned to the four *varṇa*s and invites specialization by caste. One effect of such worship is a ranking of all participating castes by a pattern of exchanges in which natural substances and services go up and divinely transvalued substances containing divine benefits go down. Another effect is the establishment of a solidarity of substance among all the castes.

Varieties of caste systems. Caste groups range in size from hundreds up to millions, with their estimated median population lying between 5,000 and 15,000. Such populations are likely to be spread over dozens or hundreds of villages and cities and over hundreds of miles of territory. Thus, few castes ever meet as wholes in one place. Nevertheless, the universality of concern for a caste's substance and code and for its rank (which can be altered for all by the conduct of a single member) guarantees at least an informal, diffuse organization of authority within each caste. Recruitment of allies or a shift to a larger context of competition may lead a caste to fuse with groups otherwise treated as separate categories.

In the eyes of others, a caste is most often categorized in an occupational role. Each caste in fact owns a set of related occupations among which its members may choose according to their particular economic situation. Thus, for example, each of the "royal," traditionally warrior, castes (called Rājput) includes landlords, cultivating tenants, and landless labourers, as well as princes and soldiers. Least specialized are some of the tribal *jāti*s in the hills and forests of central India, who traditionally do for themselves all work except metal-working and the literate professions. About one-third of all the South Asian population is in castes for whom some aspect of agriculture is the main traditional occupation. Another third, traditionally specialized by caste in nonagricultural work, in fact earns most of its living in agriculture.

Local communities and class

Settlements of fewer than 1,000 persons have contained most of the predominantly rural population of South Asia up to the 20th century. These settlements have commonly comprised members of from five to 20 castes—the number varying according to the density of the particular region and the wealth of the local population. Contact among persons of different castes is typically close and frequent in villages, being required by economic division of labour and by the forms of ceremony. Persons are readily identified by one another as members of known castes and the import of their actions accurately assessed. On the average, villagers agree with each other by majorities of 90 percent or more concerning the ranks of eight out of any 12 local castes.

Apart from their belonging to castes, the households of most villages are organized in economic and political systems that concentrate power in the hands of a landed or commercial minority. Membership in the local ruling class rarely corresponds closely with the membership of any local caste group: individuals, households, and families may rise or fall in their economic resources and influence, apart from the ranks of their castes. Local caste groups often differ greatly in their members' aggregate or average wealth and influence. Castes that have no other economic resources must serve and defer in ritual ways to those with power in order to subsist. Rank is traded in order to eat. Thus, the actual distribution of politico-economic power within a community can effect a rearrangement of the order of ranks among the castes.

The rankings of castes in villages are alterable but are not quickly altered, locally supported as they usually are by politico-economic interests, restrained by complex multilateral exchanges, and remembered by minds accustomed to a hereditary order. The mild but ubiquitous inclination of castes to improve their ranks can be estimated from the finding that the average villager believes the correct rank of his own caste to be about half a rank higher than what nonmembers of his caste are willing to concede. Efforts by rural caste groups to depart from the locally accepted line of precedence are numerous, even if moderate and slow to take effect. That castes' local ranks do change is known from case histories and from the differences in the same castes' ranks that occur from place to place.

The caste systems of rural people, as they are believed to have existed at the beginning of British rule in 1858 and as they continue to some degree into the present, can be divided into five major regions of South Asia.

Caste systems in southern India and Sri Lanka. The southern, Dravidian-speaking portion of India is characterized by a proliferation of castes, consistent with regional concern for female bodily substance. Except in the matrilineal castes (in which membership is gained through the mother) of Kerala, both parents are thought of as contributing caste code to their offspring. Marriages outside Kerala are uniformly made within the caste. Endogamous castes here tend to become very small and close, to divide into numerous smaller circles by the preferred repetition of reciprocal, protective marriages among cousins and other known kin. Distinctions between castes are more often marked by visible attributes. Only whole, uncooked foods, such as betel leaf and nut, are much exchanged beyond the caste and household.

The relatively small spread of each caste leaves it available to be ranked consistently by its members' exchanges with other castes in just a few nearby localities; there is little danger of conflict over precedence through comparisons with possibly inconsistent rankings of the same castes in any remote villages or kingdoms. At the same time, local communities are the largest and most complex in South Asia, averaging 10 to 15 local caste groups in the southeast and 15 to 20 in the southwest.

The stability of rice agriculture and government in the lowlands of southern India have also favoured a formal administration of land tenure from the state down to the cultivator (*raiyatvāri* tenure), operating either through hereditary local officials whose castes (*e.g.,* the Nāyar) have been or become regarded as the high castes or through land-endowed temples administered by Brahmins. Dynasties have had long durations, and conquests by other than neighbouring kingdoms have been few. One result has been a frequently close identification of land-tenure positions with certain castes and a closer correlation than elsewhere between castes' ranks and their powers. Brāhmaṇa caste groups have in some areas taken the positions of rulers. Another result of these conditions has been the formal social organization of the whole village community on caste lines for the several purposes of legal administration, occupation, and ritual.

Rankings of local castes in 20 or more grades are encountered in Kerala. An extreme etiquette of formal speech, clothing, gesture, and distancing among the castes is required there. On the other hand, repetition of the sets of priestly and pollution-removing (*e.g.,* barber, washerman) castes for patrons at three or four ranks and marriage of women above their caste (intercaste hypergamy) integrate the many castes into single, ranked schemes. The central role of Hindu temples in focussing the services and ranks of castes continues from Kerala up the western coast of Mysore and also down to Sri Lanka (Ceylon), especially in its Kandyan highland region. In much of Sri Lanka, however, the small village settlements have few castes, favouring their Buddhist caste system.

The caste systems of villages in Tamilnadu and Andhra Pradesh are less centrally structured and somewhat less elaborately ranked than in Kerala. Local communities contain fewer castes, castes whose wealth and power are less closely correlated with their ranks and whose residences are divided into hamlets with separate caste-linked temples. In some districts castes are organized in competing sets—*e.g.,* castes of the "right hand" versus those of the "left hand"—reflecting rivalries between ruling castes, between indigenous and immigrant castes, or between agriculturalists on the one hand and traders and artisans on the other.

Castes in the northern plains

Caste systems in the northern plains of India. The region of the later Aryan settlement in eastern Punjab and in the upper Ganges Valley (in Uttar Pradesh) and of the plains in the north Indian area where Indo-Aryan languages continue to be spoken today (Bihār, West Bengal, and Gujarāt) are generally densely populated with moderately sized villages having caste systems somewhat less elaborate than those in southeastern India. Each settlement averages about 500 persons belonging to ten castes, nine of which are Hindu and one is Muslim.

Centralization and formal organization are slight. Villages in the northern parts generally have no temple at all; those of Gujarāt generally have many private temples but none that can bring all the castes of a locality together in a ranked order of worship; those of Bihār and West Bengal that have centralized worship, organized by a local landlord, are few. The intercaste organization in the villages of the northern plains, both for worship and for control of land and conflict, tends to be based upon a plurality of landed households, mostly of the higher castes. Each household conducts its separate worship, and each separately employs specialist members of five to 15 castes. Such specialists serve by heredity on fixed rates of annual payment either in land or crop shares.

The main occasions for proving the ranks of the castes, beyond the regularly rendered services and payments, are household-sponsored feasts. Intercaste feasts in this region use a class of perfected (*pakkā*) cooked foods that are based on clarified butter, need not be served hot, and are therefore thought to be not so open to combination with bodily substance. Such feasts are offered to representatives of every caste in the village, and the resulting acceptances and rejections of food make public show of each caste's relative standing.

Local rankings of the castes in most of the northern plains are less isolated and more subject to contravention by rankings occurring elsewhere, primarily as a result of the very wide range of marriage. Except in Bengal, clan terms of kinship are used locally among persons of all castes, and locally born women must marry outside their localities. Caste by caste, linear rankings of local castes become impossible. Nevertheless, high local agreement occurs on blocks of castes; about three-quarters of the possible ranks are distinguishable.

Some castes of the northernmost plains, with some higher castes of the central region and all castes of the western region, hold that their natural code is transmitted by the father alone. Many of these marry hypergamously, linking higher with lower clans and castes. Castes here tend to have larger spreads and the largest regional populations.

The small kingdoms and pockets of Hindu and Buddhist settlement on the Himalayan slopes of northern South Asia differ greatly from the plains and among themselves in their environments. Many villages here are composed of only one, two, or three castes or tribes and thus have a very simple division of labour. These villages may be quite isolated settlements, practicing close reciprocal marriages.

The more densely populated Himalayan areas, such as the valley of Nepal, were strongly Hinduized by Rājput and Brāhmaṇa immigrants, increasingly so after the Muslim conquests of the plains. Nepal contains dozens of Buddhist and Hindu castes that are elaborately graded and linked, both by position in the state and sometimes by intercaste hypergamous marriages. Both parents are believed to transmit caste-coded substance, and offspring of mixed marriages are assigned to third castes having ranks between those of their two parents' castes.

Caste systems in central India

Caste systems in central India. The central Hindu region extending from Rājasthān in the north through Madhya Pradesh to Mahārāshtra on the south and Orissa on the east is characterized by relatively sparse rural population that tends to be concentrated in villages of moderately large size—500 to 700. The average number of castes present in these villages is less than in the northern plains, for in each state and locality two or three out of a few very large clusters of castes preponderate.

Within the central zone, the regions of Rājasthān, Kutch, Kāthiāwār, and much of Madhya Pradesh contain large desert or hilly areas that can be exploited fully only by herders and semimigratory farmers. Where permanent villages exist, they organize hereditary intercaste services by household. But wide and flexible networks of organization are advantageous within the castes of these mobile farmers. They are essential for the controlling castes of Rājput and other warriors and among the numerous small Hindu and Jaina (a religion founded in India in the 6th century BC) trading castes of the region. Such networks are structured for these higher castes not by close caste councils such as those that prevail in southern India but rather by the rulers' deep and widespread corporate lineages and clans (recorded for them by genealogist castes), by belief in exclusively male transmission of caste code, and by broad rules of marriage like those found in the northern plains. Within such rules, alliances may be sought and broken. Equipped with such effective structures for mutual aid in an otherwise famine-prone environment, cultivating (Rājput), herding (Gūjar, Ahīr), and merchant (Mārwāri) castes of the dry lands have spread into the neighbouring plains, often as predatory forces in politics and commerce. At the same time, the nondominant castes of Rājasthān, although widely dispersed, follow an opposite strategy of reciprocating, caste-endogamous marriage (inside the caste) that tends to consolidate rather than expand their scarce resources.

South of Rājasthān and north of Mahārāshtra is the fertile plateau of Mālwa, a crossroads of trade, migrations, and routes of conquest in all directions. Villages here, like villages in parts of southeastern India, are often divided between two sets of castes: on the one hand, those who eat meat, worship the Hindu goddesses, and follow the Rājput and local Brāhmaṇa codes and, on the other hand, those who are vegetarian, worship the Hindu god Vishnu (Viṣṇu), follow the Vaiśya code, and are often immigrant castes.

Farther south in the central zone is Mahārāshtra. The Mahārāshtrian village makes public feasting an important focus of intercaste relationships, although restricting feasting to the feeding of kinsmen and traditional servants; it also separates its lowest castes (*e.g.,* Mahār) from the others in a distinct residential hamlet, regards its specialist castes as serving the village as a whole, and organizes all its castes around one or more temples. What distinguishes and moderates intercaste relations in the Mahārāshtrian village is the strongly communitarian nature of village institutions: the village runs its affairs by decision of a village council that includes representatives of all castes and typically makes its main temple into a community centre. It shares some of these features with villages in adjacent sections of western Andhra Pradesh (Telingana or Telengana) and northwestern Mysore.

On the wide eastern and southern margins of the central zone and through its heart in the hills of the west and on both sides of the Narmada River Valley are forested areas inhabited largely by hundreds of tribal peoples—endogamous ethnic groups that have distinct moral systems, that trade their products in peasant markets, but that engage little if at all in the exchanges related to worship that would make them part of any neighbouring Hindu moral system of ranked groups. Non-Hindu tribes are also numerous in Assam and in the North East Frontier Agency (*e.g.,* Nāga). Many Hindu castes derive from such tribes.

Caste systems in Bangladesh and Pakistan

Caste systems in Bangladesh. The densely populated floodplain of the Ganges and Brahmaputra rivers differs radically from northern India in the ecology of its human settlement. A conception of the stable village of locally exchanging castes does not exist. Instead, the scattered households, loosely linked in shifting affiliation with the stronger leaders among them, seek to fill their needs for caste-specialized products through resort to cash exchanges at rotating markets. What specialized services and goods are not traded at markets are brought door-to-door by waterborne artisans and merchants. Even agricultural labourers are often itinerant and indistinguishable from mobile fishermen; only domestic service is generally rendered on long-term local contracts.

The rural population of this socially fluid region is also for the most part relatively undifferentiated by caste and even by kinship groupings of any considerable depth. Localities for miles may be inhabited by a single genus of Muslim cultivators: specialized Hindu castes concentrate themselves apart or cluster at the rare permanent markets and centres of land administration. Thus, neither the human materials nor the institutions of domestic or community worship are present to encourage any elaborate rankings of castes. Indeed, the region is not known to have been Hindu in any large degree or for any appreciable time. Its older past was Buddhist in regime; it came at last under the rule of a Hindu dynasty only a century before its conversion to Islām began in the 13th century. Both Buddhist and Islāmic treatments of caste are congenial to the peculiarities of the region.

Caste systems in Pakistan. Like Bangladesh, the provinces of Pakistan—the desert of Baluchistan, the hills of the North-West Frontier Province and Kashmir, and the plains of Sind and West Punjab—passed from a Buddhist to an Islāmic regime without elaborate development of caste ranking in the Hindu style. Here, again, the ecology of much of the region does not permit the co-residence of many ethnic groups. Neither the mobile herdsmen of the desert and hills nor the cultivators in tiny hamlets on the Indus Plain can assemble many castes to construct complex local systems. Exchanges among the castes of these areas take place more often in towns and for cash at market rates.

Where villages are larger and richer, as in West Punjab, they have attracted members of numerous castelike lineages, or clans, and have become more than equal in ethnic complexity to villages of Hindu eastern Punjab. But here conditions of instability—like those noted for Rājasthān and Mālwa in the central region of India—play a part, for both West Punjab and the North-West Frontier Province have been shattered socially by repeated tribal incursions. Parts of the same tribes and lineages have taken widely varying ranks, according to their political fortunes in different localities. Although marriage is spatially closer than in south India, tribal and lineal names are spread over great distances and tend to confound ranks based on local exchanges. Fewer than one-third of the possible ranks are therefore found to be distinguished among castes in villages and towns of this region.

Caste systems outside South Asia

South Asian caste systems overseas. Caste systems have been carried by South Asians to other continents but are found there only in drastically reduced versions.

Traders and settlers from India established elite cultures of Vedic, then Buddhist religion in many parts of Southeast Asia during the early centuries of the Christian Era, but only on the small Hindu island of Bali are traces of a caste society known to have survived beyond the advent of Islām. In Bali, the four Vedic *varṇa*s are employed as a ranked classification of patrilineal titles (those handed down through the father's line of descent) held by generally endogamous (sometimes hypergamously related) lineages. Higher titleholders are more likely to marry nonlocally. They often receive elaborate verbal deference from lowers. They may refuse cooked food from lowers to assert superiority without believing that food transfers alone can alter rank. Apart from some specialization in worship, Balinese caste titles in practice designate claims to rank, rather than major occupational groupings.

The millions of South Asians who went as labourers to Africa, Malaya, Fiji, Trinidad, Suriname, and Guyana in the 19th and 20th centuries proved unable to carry complex caste systems with them. The regional heterogeneity of those who migrated permanently made their caste identities and their codes of exchange mutually unintelligible. Original caste names have been changed, and identities of code and substance are everywhere merged by marriage into composite regional or linguistic and religious castes. Such a composite Hindu caste is commonly further divided by a single distinction into a priestly and nonpriestly caste. Only those overseas traders (in Burma, Malaya, or East Africa) who return periodically to their home regions are able to adhere to their original codes of marriage and caste identity. Nowhere abroad have modern South Asians set up local communities having a caste division of labour in work or worship, again excepting the Brāhmaṇa role among Hindus.

Caste systems in cities. Caste systems are also modified by the conditions of city life. Like overseas migrants, urban residents in South Asian cities of all regions and since ancient times have had to cope with the difficulty of identifying and ordering great numbers of regionally heterogeneous castes. They have attracted and supported specialized groups little known in rural areas. They have also had to cope with largeness of scale, with recognizing the castes of thousands of mobile individuals. Their larger populations within each caste have furthermore tended to segregate themselves in self-sufficient enclaves, each with its own internal division of labour and its separate cult or sect. Finally, their intercaste exchanges have always tended to be mediated more by shifting power and prices in the marketplace than by corporate group codes.

Specialized mercantile, administrative, and military camp cities of the northern and western regions have seen these tendencies carried furthest, also suffering frequent depopulation and relocation with changes in their specific functions. Certain more permanent royal and religious cities, especially in the south, have seen these tendencies counteracted most fully by centralized exchanges at palace or temple among formal representatives of the various castes. Royal regulation of service, payment, residence, and consumption according to caste rank and unilateral arbitration of disputes over rank also reduced the urban potentialities of anonymity and disorder.

Caste and the religions related to Hinduism. Belonging to a caste is thought by South Asians of the indigenous religions to raise difficulties for the perfection of a person's "soul," his unique natural code (*ātman*). Yet a person cannot exist embodied in the natural order without belonging by birth to one of that order's natural genera and without being subject to its inborn generic code.

Religious attempts to find liberation from caste

Vedic liberation and caste. In the Vedic view, the soul is contained in two kinds of body: a gross body that is abandoned at death and a subtle body that clings to the soul. The gross body belongs to a natural genus, while the subtle body does not; yet, being still a unit in the natural order, the subtle body must go on to be reborn sooner or later within a gross body belonging to some natural genus. The moral actions (*karman*) of a person in following the code of his genus affect the quality of his natural substance; and this determines where in the ranked order of natural genera his soul will be reborn—among plants, animals, demons, gods, or in the same or some other caste of humans.

Further Vedic views stress not the unknowable ultimate rebirth of the soul but its long continuity in its original birth as a heaven-dwelling spiritual ancestor, fed by offerings from corporeal descendants, or as a dissatisfied terrestrial ghost, troubling living persons. These and still other Vedic views continue in the beliefs of some Hindu castes, along with popular theories of this-worldly punishment and reward for moral actions.

Liberation (*mokṣa*) from the endless flux (*saṃsāra*) of birth and rebirth—often called transmigration—is the highest goal of the Vedic religion. Liberation means absorption of the individual code or soul into Brahman, the divine, perfect substance that is the ultimate source of the natural and moral order. This goal is attainable only by a male of twice-born genus who has followed the sacrificial code through the three life stages of student, householder, and forest dweller. Only he is to enter the fourth, or ascetic (*sannyāsi*), stage by abandoning the sacrifice and performing his own funeral. As an ascetic, he is morally dead and thus need have no more concern for his caste nor it for him. He is to own nothing and give nothing to others but is to beg from others of any caste a minimum of food. Thus abasing himself, the ascetic reverses his former high rank. Through knowledge and meditation and aided by such austerities, he strives to obtain control over the relationship of his soul and body in order to effect their complete separation. At his second, natural death, if he has achieved control, his soul is thought to attain perfect liberation from the caste-bound natural order.

Hindu sects and caste. Later, in Hinduism, liberation is thought to be attained through the individual's participation in one of the numerous Śaiva, Vaiṣṇava, or Śākta "sects." Each sect is both a divine and a human moral order thought to have a code above and beyond the natural and moral codes that are inborn in the ranked human genera. Each sectarian order consists of three kinds of persons—gods, preceptors (gurus), and devotees. Persons of any caste or sex are thought eligible to participate. They are related within the sect not by their previous particular natures but by manipulations of divinely coded substance.

In some sects, liberation is attained by assembling these persons more closely. By means of direct personal devotion (*bhakti*) to a chosen supreme deity, devotees are believed to attain a permanent relationship of superior love (*parā-bhakti*) equivalent or superior to liberation. By accepting or eating substances, especially food leavings transvalued by preceptor and deity, the devotees bind themselves to the god and to each other as members of his divine family, overriding the ranks of their castes as they are ordinarily maintained through image worship.

In other sects, those called Tantric, liberation is attained by reversal. By means of secret words and prescribed collective ritual acts directly contrary to ordinary moral codes—*e.g.*, eating meat, drinking wine, having promiscuous sexual intercourse—a person is thought to destroy the bonds that normally link him to the lower human order. He is thereby thought to attain a superior liberated state (*parā-mukti*) while still in human form.

Liṅgāyats and Sikhs. The Liṅgāyat community of Mysore arose within Hinduism in opposition to Jaina and Brāhmaṇa domination during the 12th century; the Sikh community of the Punjab arose as a synthesis of Islām and Hinduism during the 16th century. Both religious communities make radical revisions by manipulating together the mutually immanent, nondual constituents of the Hindu systems that are seen either as encoded substances or embodied codes. Liṅgāyats and Sikhs both believe that they become a divine human genus when they are given initiation (*dīkṣā*) into the making of natural sacred sounds by their teachers (gurus). The Liṅgāyats believe that they become human incarnations of their supreme god, Śiva. Being divine, they reason that their bodily substances at birth, death, or menstruation no longer pollute, as do those of ordinary Hindus. The Sikhs believe that through initiation they incorporate their teachers, who are believed to be perfect and divine, even though human. Both Sikhs and Liṅgāyats believe that they attain liberation while still belonging to their many original castes. These castes are thought to continue as a simply human, therefore lower, order. Ranks of castes in these orders, as among Muslims, are expressed in upward transfer of daughters in marriage.

The higher divine order within both of these communities consists of the categories of teachers, priests, and disciples. These categories are also conceived to be both moral and natural. Any person of any caste or sex who comes to possess the sacred natural sounds of the Liṅgāyats or Sikhs may become a Liṅgāyat or Sikh disciple; and any Liṅgāyat or Sikh may in theory become a priest or teacher. In fact, it tends to happen that members of the higher castes are more likely to become teachers or priests, and members of the lowest castes are liable to be regarded as inferior disciples and therefore excluded in some ways. What distinguishes these communities from their Hindu neighbours, despite such discrimination, is their members' mutual sharing and exchange of transvalued substances. Thus, Liṅgāyat disciples and gurus mutually regard each other as gods, drinking the dirt of each other's feet, while Sikhs of all castes drink nectar from a common bowl and take food reciprocally.

Caste among non-Hindus in South Asia. Other major religious communities in South Asia share with Vedic and Hindu peoples some basic assumptions regarding the particularity of human genera, the unity of substance and code, and most of the meanings of exchanges. Unlike Hinduism, however, both indigenous Buddhism and the foreign religions are at pains to distinguish human from divine substance and the inferior order of castes from another order of superior morality.

Buddhism and caste. The Buddha and his followers advocated a universal moral code transcending all particular codes. By this code, each person should honour the Buddha and his followers with gifts and should be personally nonviolent, moderate, generous, and self-denying in emulation of the Buddha. Living by this code, any person eventually escapes both from the social order and from the world of nature. He attains a superior nonnatural state called Nirvāṇa.

Buddhist conception of a universal moral code

Buddhism relegates the natural Vedic *varṇa*s to a lower moral order and distinguishes from it a higher moral order conceived as containing three new categories—householders, king, and monks. Householders of all castes, rather than only Vaiśyas and Śūdras, are seen as the providers of food, wealth, and service. The king and his officials are to be chosen from among the most qualified persons rather than from among Kṣatriyas only. The highest teachers of morality, the Buddhist monks, are to be chosen by personal qualifications, rather than being hereditary Brāhmaṇas.

Ranks among these social categories are conceived in Buddhist thought as emerging from exchanges like the Vedic exchanges. The Buddhist exchanges differ, however, in being conceived as contractual or voluntary in origin, rather than inherent in any genus. Householders owe taxes to the king as a reward for his godlike gifts of protection and charity. Both king and householders make gifts to the propertyless monks; monks should accept these regardless of the nature of the gift or the donor's sex or caste but should return no food substance, not even leavings, to their benefactors. Instead, monks are to return only the supreme gift—instruction in the words of the Buddha regarding knowledge of nature and regarding the code leading to liberation of the soul from the natural order.

The Buddhist conception of society prevailed among peoples on the trade-oriented peripheries of South Asia (the Himalayan borderland, the present Bengal-Bihār, Pakistan, and Sri Lanka [Ceylon]) and in scattered urban areas until the 8th to 13th century, when it was supplanted in most areas by Hinduism or Islām. Only in Sri Lanka did a Buddhist society survive along with castes until the British period. There, many gods and ranked systems of 20 or more natural castes are found, much as in Hindu South Asia. Above is the distinctive higher moral order of Buddhism, consisting of the Buddha and several orders of monks, the state, and householders, all concerned with the goal of personal spiritual liberation. In the modern Buddhist democracy of Sri Lanka, natural castes tend to lose their moral significance as units of exchange and rank, while the state's legislation is seen as "returning" the society more and more to the remade social order conceived by the Buddha.

Jainism and caste. Jainism arose at the same time and in the same region (Bihār) as Buddhism, and, like Buddhism, it conceives of certain householders and ascetics—not castes—as the social units of a transcendent moral order. It differs from Buddhism in that it does not present an alternative conception capable of embracing the whole of society. The Jaina code for conduct is more extreme than that of Buddhism, requiring adherence to a rule of thoroughgoing nonviolence. In theory, persons of any *varna* or *jāti* may become Jainas; but virtually all occupations, with exceptions such as banking and commerce, are considered to be violent. Hence, most adherents have been members of castes of merchants. Hundreds of small Jaina castes are now found in western India (Gujarāt, Mahārāshtra, Rājasthān).

Islām and caste. Islām was introduced into South Asia by merchants, conquering armies, and missionary Ṣūfīs (Muslim mystics). It became established by the 13th and 14th centuries and became the predominant religion in Pakistan and Bangladesh, where the Buddhist pattern of social organization had predominated. Like the Buddhists, Muslims of South Asia conceive of society as containing a lower and a higher moral order. The lower moral order consists of ranked genera—caste brotherhoods (*berādarīs*) thought to be related by the shared natural substance of blood.

Muslim conception of a lower and a higher moral order

The moral ranks of these rural blood brotherhoods in Pakistan are often based on a reputation for past violence and on the local, hypergamous exchange of women. A brotherhood should take care of its own women; hence, the brotherhood that gives women to another is thought to be morally deficient and to rank lower. In the northern plains of India, where Muslim castes live among Hindu majorities, their moral ranks are more likely to be based on food exchanges. In Muslim-majority communities, however, the public donation of cooked food shows a brotherhood's piety and raises its rank, although the receiving of such donations is not considered to lower its rank.

The moral order of blood brotherhoods, or castes, is felt by Muslims to be transcended by a number of purely social, contractual brotherhoods whose members are supposed to behave toward each other as if they were related by blood even though they are not. The largest purely social brotherhood in the higher moral order is a religious one and consists of those persons who worship Allāh and submit to the Islāmic code for conduct enshrined in the Qur'ān. The single Islāmic deity is said to be the source of the natural and moral order but to transcend it himself. Men depend on him for their bodily existence and salvation, but he does not depend on them, not even for food. Muslims offer up their total submission, but no material goods. However, exchanges of food are made among Muslims on ritual occasions to express the equality and solidarity of the members of a social brotherhood, and also to express the differences of rank in the higher social order. By reciting the sounds of the Qur'ān, eating substances animated by the words of saints, and venerating the remains of dead saints and of the Prophet (Muḥammad) himself, the worshipper becomes more like the saint or the Prophet. Servants of a landholder or followers of a village headman—also brotherhoods in the Muslim moral order—express their subordination by taking ordinary cooked food from him.

The relative ranks of these purely religious and social roles in the higher moral order usually coincide with the relative caste ranks of their occupants in the lower moral order. Persons who occupy the higher roles of saint, pious Muslim, and landholder are usually members of the higher castes, and persons who occupy the lower roles of casual Muslim and servant are usually members of the lower castes. Yet Muslims are quite clear that, when there is a discrepancy between the two, the role occupied by a person in the higher moral order takes precedence. As a result, there is a persistent strain in Muslim communities to bring the caste membership of a person into accord with the role he occupies in the higher Islāmic moral order.

Judaism and Christianity and caste. Other religions accommodate what are seen in South Asia as natural castes. Thus, in Bombay there are Jews of white, black, and recent immigrant descent; in Cochin, besides white and black, there are Jews of slave-mixed descent. Christian converts of many centuries marry and eat within their respective original castes or adopted sects, especially in south India. Adaptations in the higher moral spheres of these religions are much like those made within Buddhism and Islām.

Legal and political roles of caste. *During the pre-British period.* By Hindu conception, castes are categories essential to defining morality (*dharma,* sometimes translated "law"), but not to the calculations of political or economic advantage (*artha*) that all men make as individuals. Morality is conceived to be as highly particularized as nature. Its stratified principles are partly compiled in the moral code books, but its full contents include the varied and innumerable unwritten moral codes (*ācāra*) that are thought to inhere in particular corporate units of each specified kind—lineages (*kula*), castes, or localized communities (*deśa*). If individuals or groups of different castes compete for the same advantages or ends, their competition is not necessarily a matter of moral concern. However, one point of contact between self-interested behaviour and morality is elaborately considered by Hindu legalists: the office of the king. It is the moral duty of the ruler (properly a Kṣatriya) to use force (*daṇḍa*) so as to establish the moral order, especially in order to maintain the rank and separation of the castes, so that their internal self-government and their proper exchanges may continue.

Content of Hindu moral law

All rules, rewards, and punishments among the castes are systematically unequal. Lower castes are customarily prohibited not only from participating in higher worship but also from such acts as touching the higher castes' water, wells, utensils, clothing, persons, or food and from trespassing on their furniture, buildings, roadways, light, sight, and air. Such acts of trespass, along with many more crimes, are scaled from more to less consequential, as the castes involved are ranged from high to low. Done at the command of a person of higher caste to a person of lower caste, such acts are thought to be privileged or acceptable, but, done on the lower's initiative, they cause insult or "pollution" to the higher and deserve graded retaliation or punishment.

During the British period. British administration took an interest in the Vedic moral code books as statements of abstract principle but followed their own views of law as essentially above nature and as unconnected with particular categories of men. As part of a policy of nonintervention in religious and social matters, they declined to play explicitly either the Hindu ruler's role of enforcing caste distinctions or the reformer's role of abolishing them. Local officials were generally as unsympathetic to reform as were local communities, so that, in effect, local caste codes of exchange tended to be supported. At the same time, other governmental measures and widened communications unintentionally encouraged an intensification of internal caste organization. Military units were recruited on a caste basis. The caste membership of individuals was registered in censuses and published in composite totals, and competitive claims by about one-fifth of the castes to affiliation with Vedic *varṇa*s were stimulated by court applications of Hindu personal law.

Limited reforms of the political relationships among castes were undertaken after 1919 in connection with the widening of political franchise. These reforms included reserved legislative seats and benefits in education and employment favouring the lowest ("Scheduled"; *i.e.,* officially listed) and very low ("Backward") castes and tribes. Such reforms underlined the separateness of caste groups and rewarded political allegiance to them and to the imperial government awarding the new advantages. Certain larger castes or merged clusters of similar castes also became politicized, forming associations to press for advantages in the widened arena of governmental competition. In all regions, effective organization for gaining power required less narrowly based political alliances, factions, and parties that could incorporate members of many castes.

Effects of the Indian independence movement

Such political changes during the later British era, along with the British exclusion of religion from governmental concern, stimulated leaders of the independence movement to try to reshape the Hindu social system, seen as a whole made up of particular natural genera, into a structure suited to the winning and the running of a nation of united citizens. B.G. Tilak in Mahārāshtra in the 1890s mobilized urban ceremonies of communal worship in which representatives of all Hindu castes and neighbourhoods offered prayers for political salvation to the Vedic and regional god Gaṇapati. Mahatma Gandhi, Tilak's pupil and the architect of India's freedom, in the 1930s urged a return from the multiplicity of *jāti*s and secular individualism to a neo-Vedic system of four *varṇa*s and stages of personal life (*varṇāśrama-dharma*). This was to be a system emphasizing the mutual responsibility of each to all, one in which men would affiliate themselves with *varṇa*s (interpreted as organically related socioeconomic classes) according to the way they reshaped their bodily natures by following codes of vegetarian and nonalcoholic diet, fasting, sexual abstinence, and public service. Gandhi's hope was to abolish the category of the lowest, sub-Śūdra castes—whose members he labelled Harijans (Offspring of God)—by reversing their former position in the code of exchanges. Harijans were made to serve cooked food to workers of the higher castes, while these in turn swept the garbage from the Harijans' streets. Harijans entered previously forbidden temples of southern India under Gandhi's leadership. Ultimately, as the reformers hoped, the many previous grades in worship tended to become reduced for all persons to the minimal two grades of priest and worshipper, if not to a single grade. The wide popular appeal of these and similar programs of revival and reform must be ascribed to their speaking initially in the caste idiom of substance and code. Their successes led to the strengthening and activation of still larger regional and religious castes, such as Hindus, Muslims, and Dravidian nationalists, which sometimes clashed with others in subsequent competition for power.

A still more innovative and prominent change in caste thinking, devised first in 19th-century Bengal, declares the land substance of India to be the nourishing mother of all her people, Hindu and Muslim, high caste and low. Mother India's code for conduct enjoins unity on all residents in her service and defense. This universalizing scheme spread officially through the nationalist movement to the new nations of India and Bangladesh.

Modernizing tendencies in caste thinking

Since 1947. The independent governments of South Asia have moved vigorously to modify caste conduct through legislative action. Most consistent with earlier Buddhist and Islāmic roles was Pakistan's declaration to be an Islāmic state having transcendent law and therefore having no need to recognize the natural castes.

A secular legislative Indian approach to conduct is seen in the Hindu code laws of the 1950s. These replace the particularistic Vedic moral code books by a uniform code taking no account of *varṇa.* They also abolish caste endogamy as a requisite for valid marriage. The legislative approach is seen fully in the Indian constitution of 1950, authored chiefly by B.R. Ambedkar, leader of a party of the lowest castes. The constitution rests on the notion that government must regulate individuals directly and not through autonomous corporate castes. It abolishes caste "untouchability" (now to be legally defined by courts) and forbids any other restriction on public facilities arising out of caste membership. Subsequent legislation against caste restrictions has proven generally influential at the urban temple or the crossroad coffeeshop but not at the village well. Ambedkar advised his followers to renounce Hinduism in favour of Buddhism, which he saw as providing a casteless social system.

Many modern South Asians consider caste-reform actions by government as necessary and desirable but regard the distribution of economic resources among castes as the most crucial problem. South Asian Marxists go further in their economic program, yet less far in their social program, holding that castes are secondary and need not be a subject of policy; Marxists expect that castes will disappear by general intermarriage once the only real forces—the antagonisms of the economic classes—are resolved by revolution. The overall official Indian position on caste systems by the 1970s was a compromise: natural ethnic groups are relegated to a lower order that is devoid of legality, and provision of comprehensive moral, economic, and political codes is a function reserved to the processes of the transcendent national state.

Persistence and evolution of caste

The ideas of South Asian citizens as to what caste systems are and should be vary widely among the positions sketched. Practice also differs. Thus, norms of caste endogamy appear generally to have stiffened, while marriages outside of certain small urban groups of high caste nevertheless may exceed 20 percent. Traditional caste occupations are followed by an overall minority that grows less each year, but these occupations are often replaced by similar substitutes. The forms of Hindu worship are subject to constant innovation, but worship itself—ranked exchange between the genera of men and gods—is not abandoned. Where prevailing opinion is conservative, as in most rural localities, caste codes of exchange may continue to be observed publicly although individuals violate them privately; where opinion follows the official position, as it often does in cities, ranked exchanges among the castes may be confused or transcended publicly but maintained by conservative families in the domestic sphere. Most individuals move among social compartments that differently organize the enduring linkage between substance and code.

Caste systems in societies outside South Asia. Caste systems need not be thought of as unique to South Asia or to its emigrants. Rudimentary caste systems have appeared elsewhere to the extent that similar situations, structures, and cultural concepts were present.

Situations of encroachment between alien ethnic groups have generally been productive of simple, short-lived caste systems like that of Aryans and Dasyus in northern India in early Vedic times. Many a society has opposed itself as a superior whole to a despised immigrant population of traders, artisans, or Gypsies but usually without developing systematic conceptions or mutual consensus regarding ranks. Many a society has incorporated but ranked separately, then in time assimilated, a caste of foreign conquerors (*e.g.,* Peru), slaves (*e.g.,* Brazil), or noncitizens (*e.g.,* imperial Rome). Occasionally, a society such as Mexico, with its duality of European and Indian cultures, has preserved contrasting ethnic styles of life long after rank, descent, and marriage have ceased to insulate the onetime castes from each other socially.

A relatively small number of ethnically heterogeneous, expanding societies—certain medieval African tribal kingdoms and some recent European colonial states such as the United States, Rhodesia (now Zimbabwe), and the Republic of South Africa—have used political means to perpetuate the ranks, occupations, and biological relationships of two to four ethnic categories. The legality of the black and white caste systems in the United States did not endure two centuries, and the moral future of other recently legislated colour-caste systems is in question, but there is no doubt that ritualized composite tribal kingdoms such as Mali and Rwanda in Africa did maintain simple caste systems with moral authority for some five centuries.

Castes have rarely developed out of a single society's classes. In East Asia, several homogeneous societies sharing Buddhist ideas regarding the bodily pollution of death (Tibet, Korea, Japan) have generated internally their own local fractions of morally abhorred specialists, originally leatherworkers. These occupationally outcaste persons constitute small, scattered populations, separated in marriage and descent from the local majorities, although not thought of as ethnic or corporate units. In the lower Mississippi Valley of North America, the Natchez and Chitimacha tribes developed internal caste systems tied to hereditary offices in village war and temple cults while absorbing alien ethnic groups as the lowest of four ranks. The relative success of the Natchez in developing an internal caste system may be attributed to their facing, like the later Aryans, sharply defined alien ethnic units with an internal moral system of sharing and exchange.

(Mc.M./R.B.I.)

MINORITY GROUPS

The most common general description of a minority group used is of an aggregate of people who are distinct in religion, language, or nationality from other members of the society in which they live and who think of themselves, and are thought of by others, as being separate and distinct. Separation, too, often implies that the members of such a minority are excluded from taking a full share in the life of the society because they differ in certain ways from the dominant group—a situation that tends to develop attitudes of discrimination and prejudice toward the minority in question, attitudes that may also be assumed by other groups or minorities in the society. The minority itself is likely to respond with strong attitudes of group loyalty and to develop forms of behaviour that, by design or not, help to segregate its members still further from the rest of society.

"Minority" and "minority group"

The sociological employment of the word minority, then, largely agrees with common usage in denoting a distinct, separate group of people who are different in certain easily recognized aspects from the majority. But the term minority group implies rather more; for to a sociologist a social group of any kind is an aggregate of people with defined aims and rules of membership and with its own obligatory rules of behaviour and a subculture that publicly mark it off from the rest of society. It is the use of particular forms of cultural behaviour, used as criteria or emblems, that marks off a minority group from other types of group.

The word minority denotes by implication a part of a larger whole, but a minority group in the sociological sense is not always a numerical minority of the population. In parts of the southern states of the U.S., blacks form a clear majority of the population, but they are nonetheless a minority group in relation to the numerically smaller dominant group of whites. A similar situation existed in East African towns under the former colonial rule of the British. Under British rule immigrants from India had settled in East Africa as traders and skilled artisans. Most of them lived and worked in the towns, where they formed the overwhelming majority of the population, although, in East Africa as a whole, the Indians numbered less than 1 percent of the total African population. Despite appearances in the towns, the Indians were clearly a "minority group." In South Africa, to take another example, the Bantu population, although many times more numerous than the dominant white group, are nevertheless a "minority group."

A minority group is generally but not always politically less influential than other sections of the population, but its political position is only one factor in distinguishing a minority from other types of grouping that may also occupy subordinate positions. What marks off a minority group in the eyes of members of a society are a number of other distinctive features: race (as, for example, in South Africa or the U.S.), culture, language, religion, or economic function. To ordinary citizens it is characteristics of this kind that matter when they think of minority groups. A few examples of different kinds of minority groups should make these various points clearer.

After the Norman conquest of England in AD 1066, the native Anglo-Saxon inhabitants still constituted the majority of the population, but their culture and the social forms sustaining it were subordinated to Norman-French political, social, and cultural institutions. Like Africans, many Asians, and American Indians under the former European colonial empires, they lived as a linguistic and cultural minority group.

Religious minorities

The demarcation of minorities by religion is possibly more familiar from historical sources than from contemporary situations, where, especially in the Western world, the adherents of separate religious persuasions are more commonly regarded as categories of the population than as coherent groups of people with distinctive religious, cultural, economic, and political aims, which they pursue in their own interests. But even today such situations are not uncommon. In the northern parts of western Malaysia, for example, Thai-speaking Buddhists form an important religious and cultural minority group, whereas a few miles across the border in southern Thailand it is the Muslim Malays who are a religious and cultural minority. In India and Pakistan the confrontation of Hinduism by Islām led to the emergence of minority groups whose differences, though expressed primarily in a religious idiom, had important political and economic aspects. When the English consolidated their conquest of Ireland in the 16th and 17th centuries, a small number of English-speaking Protestants gradually spread out through a much larger Irish-speaking population of Roman Catholics and subordinated them religiously, culturally, economically, and politically. But only in some contexts were the Irish regarded as a purely religious minority group: in most situations they were seen, and felt themselves, to be primarily political and economic subordinates.

Economic minorities

Any social group formed for particular aims is likely to develop distinctive cultural features, but a group whose ends are primarily economic seems to do so less frequently, unless its organization is also intimately associated with other principles, such as race or kinship. Shopkeepers in any society are people with a number of shared interests. They are usually united in common associations, such as chambers of commerce, but it is unusual for them to constitute a minority group in the strict sense. In certain circumstances, however, they may do so. During the 19th and 20th centuries large numbers of Chinese people emigrated to Southeast Asia. Their motives in leaving China were almost exclusively economic: it was hard for them to earn a living at home. In Southeast Asia most of them eventually became traders, forming closely knit economic groups. The indigenous inhabitants of Southeast Asia, however, as much as the Chinese themselves and their colonial rulers, always saw these groups as Chinese first and only secondly as economic. In other words, the immigrants lived as Chinese ethnic groups with economic functions rather than as economic groups, most of whose members happened to be foreigners.

During the same period there was also a large emigration from the subcontinent of India to places as far apart as the Caribbean, Fiji, East and South Africa, Mauritius, and Great Britain. Like the Chinese, these Indian emigrants left their homes for economic reasons and settled abroad as trading or labouring minorities, but, as in Southeast Asia, the people among whom they settled regarded them primarily as cultural and ethnic groups.

The form of a minority group and its position in a particular society depend on many variable factors. Its members, as has been seen, may be bound to one another by common religious ideals or by similar political, cultural, or economic interests, or the tie may be one of race. In East Africa the Indian immigrant settlers distinguished themselves from the Africans and the Europeans first by the differences of their Indian languages and their Indian culture, but they also used the criterion of race to separate themselves, especially from the Africans. In Great Britain the dominant English group uses race as the main criterion to segregate ethnic minorities such as immigrants from the Caribbean, India, Pakistan, or East Africa. But however a minority group may be demarcated, for it to maintain itself as a group its organization normally must include political, economic, and cultural aspects. The emphasis placed on these various factors determines the form that the minority group takes and the place that it is assigned in society. Further, the emphasis itself varies with factors such as the group's size, its ideology, and, to some extent the outcome of these two, its wealth and political power.

The nature of ethnic groups and other types of minority groups. A social group, as noted earlier, may be founded on almost any kind of common interest, but, as also noted, strict reference to minority groups should be made only when the interests that bind the members also serve to separate them from the rest of society in ways that are felt to be significant by everybody. In most Western countries Freemasons are an organized minority of the male population, but to most people their characteristics do not mark them off as a minority group in quite the same way as do the distinctions of race for blacks in the U.S., for example.

"Race" and ethnic groups

Physically men vary widely, and in some societies selected bodily characteristics, lumped together under the broad heading of "race," are used as marks of social differentiation and for the formation of minority groups. If race is thought to be important, then it is likely to be used in forming groups that today are more usually called ethnic groups than minorities. (Some writers, especially in the U.S., have applied the term ethnic group also to immigrant groups who are distinguished by cultural differences in language and national origin but who have no distinguishing physical characteristics. They can, if they wish, sometimes shed their particular cultural characteristics and disappear into the wider population in ways not open to the members of ethnic or racial groups who carry the marks of difference on their persons.)

Not every society whose population is racially diverse elects to make use of race for social purposes. In much of the Greco-Roman world, in many parts of past and modern Islām, and in some areas of Latin America a man's racial attributes may have as little significance as does the possession in northern Europe of dark or fair hair. The variations in the use of physical characteristics for social purposes can be placed illuminatingly on a scale. At one end are those societies in which racial attributes, if they are noticed at all, merely place people in ethnic categories, in much the same way as the males and females of a population are often separated for social purposes. Among the ruling classes of Mexico, for instance, the physical differences between people of predominantly Caucasian and Amerindian descent are very apparent, but in many situations of social life they are unimportant compared with the possession of a common pattern of social positions in society and a shared culture based on the Spanish language. Such a society is at the opposite end of the scale contrasted to one like South Africa, which makes extensive use of racial criteria in the allocation of group membership.

An important characteristic of a social category is that it can easily become a group if its members are led to hold values in common and to follow rules binding on all members. In the same way a social group, whose members lose their common purpose and sense of corporate identity, can become no more than a social category or even disappear altogether. Indian immigrants in South Africa, for example, were heterogeneous in origin and social composition, and in the earlier stages of their settlement they were loosely united only by some common, but vague, qualities of "Indianness." Nevertheless, they did constitute a distinct and recognizable element in the population, thus making them a social grouping more cohesive than a mere category but less cohesive than a fully organized group. Later changes in South African law compelled the Indians to assume more and more common, obligatory rules of behaviour that were different from those of other sections of the population, thus consolidating them into an increasingly corporate and separate ethnic group.

Characteristics of cultural diversity. It is clear that any society may include (and that most have and do include) a large number of subgroups, or constituent groups. The manner in which the constituent groups are arranged and related to one another determines the social form or structure of the society. Minority and ethnic groups do not in principle differ from other social subgroups; however, their aims, the rules by which they recruit members, and their internal organization may well be different and impose special forms of behaviour on their members. Just how a minority group is related to the other sections of a society depends in part on the structure of the society itself and in part on the structure of the minority group.

An integral aspect of the structure of any social group is the set of ideas and beliefs held by its members, which largely determines their aims and, to a greater or lesser degree, guides their actions. The ideology of a minority group is an important element in the place that it is allowed to occupy in a society, but as important, and often more significant, is the ideology of the dominant group in the society. Its members and those of the other constituent groups exercise all kinds of political, economic, and other pressures to produce a situation that as far as possible satisfies them. But it is rare for any group, whether dominant or not, to have the power to impose a situation that is wholly of its own choosing: the process is almost always a two-way, reciprocal one.

There are, roughly speaking, two ways in which the members of a society may try to solve the problems posed by the presence of a minority group. They may either attempt to eliminate it altogether or decide to tolerate it. Both methods, separately or combined, with good grace or bad grace, have been used many times in the past. They are still being used today.

Elimination of minorities: assimilation, suppression, ejection. When the physical features, aims, or behaviour of the members of a minority group are unacceptable to the rest of society, there may be attempts to get rid of the minority altogether—either by forcing or encouraging its members to adopt the dominant group's culture and so be assimilated or by completely ejecting the minority from the society.

Assimilation

Assimilation (sometimes called acculturation) has been defined as those phenomena that result when groups of individuals having different cultures come into continuous first-hand contact, with subsequent changes in the original culture patterns of either or both groups. The process is a familiar one in the modern world. During the 19th and 20th centuries very large groups of immigrants, many of them speaking foreign languages, were encouraged to settle in the United States, but the possibility that they might develop into permanent, compact minority groups was felt to be a threat by the largely Protestant, English-speaking dominant group of American society. Prolonged and successful efforts were therefore made by means of school education and ideological indoctrination to assimilate the immigrants and their children as rapidly as possible. A similar policy of directed acculturation has also been successfully followed in Thailand, where the governing elite felt that the large and economically powerful minority of immigrant Chinese traders presented a political and cultural danger to the Thai people. The government, therefore, by means of legal inducements and restrictions on business and family life, combined with compulsory Thai education, have, without physical force and in a relatively short time, almost completely eliminated the Chinese minority as an ethnic group.

Ejection and suppression

The contrasting method of ejecting an unacceptable minority (and coercing those who chance to remain) is perhaps only a little less common than peaceful assimilation. A combination of ejection and suppression has been used many times in many parts of the world. In the 17th century Protestants in France were banished or driven into concealment if they refused to change their religion. In the same century the small minority of Christians in Japan was proscribed and eradicated. In modern Indonesia the Chinese trading minority, as in Thailand, was seen as a threat to national interests; the traders were therefore violently driven out of the country or forced rapidly to adopt most aspects of Indonesian culture. In the newly independent countries of East Africa the dominant African group, considering the long-settled Indian traders to be a threat to their interests, physically ejected as many as they could and severely limited all spheres of activity open to those who remained. Perhaps the most dramatic example of extermination or attempted extermination was the fate of the Jews under Nazi rule.

Assimilation necessarily takes time. It is usually a reciprocal process, and it is not always the culture of the dominant group that survives. The Greek invaders who followed Alexander III's armies and set up independent kingdoms in northwestern India disappeared almost without trace. For 200 years after the Norman conquest of England, the French language and culture were supreme, but English, a modification of Anglo-Saxon, was the language finally spoken in the country. The dominant culture evolved into a combination of French, Norman, and Saxon institutions; the same eclectic process can be seen at work in any postcolonial country in the world today.

Toleration of minorities: pluralism. A minority may be tolerated either because its aims and organization are not felt to be sufficiently objectionable to be eradicated or because the dominant group in the society, even if it dislikes the minority, may think that it is unable for political or ideological reasons or that it is morally wrong to assimilate, eject, or suppress it.

The rulers of several countries in western Europe in the 12th and 13th centuries found it hard to tolerate Jews, but so many of the rulers relied on Jewish trade and finance that they could not afford to destroy them. The rulers were consequently forced to allow Jews to live partly on their own terms, difficult though it was in that particular feudal society to accommodate non-Christian aliens or persons without feudal status. In England this was done by segregating them in ghettos; in Spain and Portugal, where the feudal system was less rigidly organized, their freedom was greater.

During the 20 years after 1950, immigrants from the Caribbean, Pakistan, and India settled in Britain. The British did not welcome their arrival, even though their services were essential for running the economic and social system, nor did the British like certain aspects of the immigrants' culture, which combined with their physical appearance made the new minority grops socially very visible. But the ideology of the British people made it equally difficult for them seriously to consider eliminating these new groups either by ejection or by assimilation, and so they were reluctantly compelled to accept forms of cultural, social, and ethnic heterogeneity that had not before existed in their society.

The toleration of ethnic and cultural minority groups can produce a wide range of different types of social organization, most of which involve some form of separation. A minority may be completely segregated (as in medieval Jewish ghettos) even when separatism may not otherwise be a major structural principle in the society. Alternatively, separatism may run right through the system, as it does, for example, in Switzerland. In that country separate linguistic and cultural groups are effectively segregated by living in different districts, although all are united by a central administrative system, which, unlike that in most multiracial and multicultural societies, is not controlled by one dominant group. Other less extreme forms of separatism are produced by discrimination of varying degrees of intensity and by excluding a minority group from goods valued by members of the society. A society divided into minorities resolves the problems of pluralism by balancing the factors that make either for the complete elimination of the minorities on the one hand, or for complete separatism, short of fragmentation of the whole society, on the other hand.

The pluralist model

Types of societies with cultural divisions. Pluralistic situations are common in all parts of the world and have given rise to much discussion. The term pluralism is used by political scientists for the view that political, cultural, and social systems are most usefully considered as being made up of separate but interdependent groups whose qualities together produce a characteristic and valuable moral order or unity. After World War II, pluralism was used and developed to describe and interpret societies divided into cultural and ethnic minorities and held together by the political power of one dominant group. J.S. Furnivall, who first began to limit the term in this way, wrote in 1948:

> In Burma, as in Java, probably the first thing that strikes the visitor is the medley of peoples—European, Chinese, Indian and native. It is in the strictest sense a medley, for they mix but do not combine. Each group holds by its own religion, its own culture and language, its own ideas and ways. As individuals they meet, but only in the market-place, in buying and selling. There is a plural society, with different sections of the community living side by side, but separately, within the same political unit. Even in the economic sphere there is a division of labour along racial lines.

The stratification model

Whether or not such "plural societies" are usefully considered as a special type is a matter that has provoked much argument. For one thing, critics would contend that all complex societies are more or less culturally heterogeneous and that to list alternative values and institutions within individual societies does not often help to distinguish one type of complex society from another. For reasons of this kind most social scientists have preferred to use the traditional models of stratification in analyzing social diversity. A society in which technology and the economy are able to support a large population is usually divided into broad divisions or strata—classes, castes, or estates—that form a hierarchy of wealth, power, and prestige. Such strata are marked off from one another by special rules of behaviour or forms of subculture, which are to a greater or lesser extent obligatory on all the members of each stratum, and which give them a graded access to goods (such as wealth, power, and prestige) that are valued by everybody in the society. Even though these broad divisions may not always correspond with the sections of a plural society or with ethnic or minority groupings, they do correspond often enough to merit further consideration of stratified and unstratified forms of society in reference to minorities and ethnic groups.

Relatively stratified societies. In any system of stratification the forms of behaviour and the rights of graded access to valued goods, which mark off one stratum from another, must be both objectively observable and subjectively accepted by most of the members of the society, all of whom must also occupy only one place in the system at one time. These criteria are sufficient to distinguish a system of ranked social groups, as, for example, that of the upper, middle, and lower classes in England during the 19th century, from an ordering of ranked statuses or positions, such as was also found in England in the same period in the rules of precedence that placed the queen's subjects in descending order below her, from the degrees of peerage to the degrees of commoner status. These criteria alone are, however, insufficient for distinguishing different types of stratification, any of which may be found in a society where ethnic or other forms of minority group are component elements of the system.

A system of social stratification may be closed or open. If an individual's place, once assigned, can in theory or practice never or rarely be altered, the system is closed. If, in contrast, social mobility from one stratum to another is permitted, then the system is open, and the relation of one stratum to another is necessarily competitive. In a closed system not only is place assigned but shares in scarce and valued resources are also differentially assigned to the separate strata. As places and resources are graded and fixed, it follows that if a closed system is to work the relations between the strata must be cooperative to a considerable degree.

Caste systems

A traditional Hindu caste system was a small self-contained hierarchy of endogamous groups in which an individual was given a permanent position at birth. Relations between the groups were governed by religious rules, the breach of which was believed to produce a state of dangerous impurity that carried ritual, legal, and other penalties.

The English word caste refers to two separate systems of social relations. In the one, the traditional fourfold division of Hindu society into four *varṇa* or castes (priests, soldiers and administrators, businessmen, and labourers) is sometimes described in Sanskrit literature in terms of what appear to be ethnic differences, but at least in historical times the *varṇa* have not constituted a system of stratification. Sociologically speaking they are categories of value, which refer to all parts of the Hindu world and which people use, among other things, for placing themselves and the subcaste group to which they belong in the hierarchy of districts outside their own local caste system.

The word caste also refers to the several thousand small, self-contained, local hierarchies of endogamous subcaste groups, or *jāti,* of which Hindu society was made up and of which a Hindu state might include a great many. Each *jāti* was named and ranked, though the ranking might not be the same even in adjacent local hierarchies. Traditionally each subcaste had a particular service assigned to it, so that the sons of goldsmiths or priests, for example, though not bound to follow these occupations, could prevent others from practicing them. In India neither castes nor subcastes were usually ethnic or cultural minority groups in the usual sense of the word, but in Iran and other parts of the Near East, cultural minorities are sometimes components in small-scale caste systems. It is possible, too, that in some areas of the southern states of the United States society could be most usefully described and interpreted by using a similar model of a local caste system.

Estate systems

If one is content to say, as are many, that in both South Africa and India social stratification is based on caste, then the concept of closed stratification need be examined no further. But writers on India do not usually agree that "colour-bar" societies are necessarily caste systems, and certainly in South Africa, although individuals are assigned positions in ranked strata at birth, the society is large scale and is not made up of many small, local hierarchies. Most important, as in most industrial societies, the ties between individuals are not multiple: the fact that a man employs another does not usually mean that he is also his landlord, his creditor, his lawyer, a member of the same church congregation, and attached to the same political party. Thus in many of its aspects the South African system of closed stratified ethnic groups is nearer to the model of a feudal estate system than to that of a caste system.

In Europe feudal estates were defined and ranked by law, and penalties for a breach of the rules of rank were legal, not ritual as in a caste system. Admission to the estates was regulated by birth and a few other carefully stated criteria, so that a man's social standing as a noble, a cleric, or a commoner was legally defined and carried with it rights and duties shared by all members of his estate and not by others. Estates were also political in nature in that representatives for assemblies such as Parliament in England and France were recruited according to their rank. An estate was a closed group whose members could not in theory move out of it into another.

The population included in an estate system might, unlike that in a local caste system, be so large that the people in it did not have the multiple ties and close knowledge of one another, characteristic of a caste system and perhaps necessary for the working of the intricate and close-knit network of ties and regulations that held it together.

In writing of ethnic minorities in contemporary societies or of slaves in older societies, authors often speak of "caste"; and though they are generally aware of the fact that they are using an analogy, nevertheless misleading implications are frequently carried by the metaphor and are not always discounted. In analyzing the place of highly segregated minority groups in a large-scale society, a model based on a caste system is seldom as useful as one based on an estate system.

Open stratification

In some situations and places, as in parts of the Caribbean, ethnic or culturally distinct groups are often treated as, and in fact may be, social classes. Sociologists generally consider a social class to be an aggregate of people in roughly the same social position, which is different from that of other social classes, and which, unlike position in a caste or an estate system, allows movement from one stratum to another. Because mobility is permitted it is not always easy to draw a boundary between one class and the next, nor is it always easy to decide how far the members of a class are bound to one another by ties that make the whole aggregate an institutionalized group. Furthermore, in an open system of stratification the scarce resources valued by the society are not overtly and explicitly allocated to each ranked stratum, and relations between classes are therefore necessarily primarily competitive, not cooperative as they often are in a closed system, where, within limits, an individual's share of resources is prescribed by his rank.

When some of the qualifications for belonging to a social class are also those for belonging to a minority group the difficulties of analysis may become very great indeed. In the later stages of the Spanish Empire in Mexico the qualities of being Indian—qualities such as physical appearance, the use of an Indian language, and a peasant or servile status—were also the marks of low class position. At a later period, with the acquisition of wealth, Spanish culture, and political influence, individuals from the lower classes could and did move in relatively large numbers from one social class to another. In the process the boundaries both between the social classes and between the ethnic groups often became blurred.

When social mobility in a system of stratification becomes very general and rapid, the system may sometimes change quite quickly into one of ranked individuals and not of ranked groups. In the opinion of some students of society, this condition has already been reached in the United States and may be a contributory factor in the difficulty that is sometimes met in absorbing certain minority groups. For instance, other members of American society are more likely to regard Negroes as stereotyped representatives of a minority group with prescribed social roles to play than they are to regard them as individuals making their own social roles outside fixed social groups. To the degree that Negroes are seen in this way they are encouraged to conform with the general view and assume stereotyped roles. Because they lack the greater freedom that other citizens have in choosing roles and in moving from one position to another, they often become even more effectively segregated.

The traditional models of stratified societies, which have been discussed here, are of course merely "ideal types" that have been constructed for studying social processes, which in different combinations produce different types of society. Any real-life society may in fact never quite match any of the ideal types, whose functions are simply to isolate the social processes that, acting together, produce particular situations. The models are useful in helping to think about such situations and in allowing a minority group to be placed in the structure of the society of which it forms a part. Much of the thought about minority groups in the past has been confused by focussing attention on inessential attributes. It is not the blackness of the Bantu in South Africa that is socially significant: it is their position in the formal structure of the society.

Relatively unstratified societies. In all societies there is a division of social labour and some diversification, but not all are stratified. Most of the discussion of pluralism has been of racially and culturally diverse populations within the jurisdiction of a central government—in such countries as Switzerland, Malaysia, and Canada. But there are other types of plural situations in which separate ethnic or cultural groups live within a common and accepted system of relationships that does not amount to either an effective central administration or a system of ranked groups. Systems of independent but allied groups of this kind were not uncommon, especially in tribal areas, before the expansion of European colonial regimes. In the eastern woodlands of North America, for instance, a group of Indian tribes known as the Iroquois had confederated for purposes of making decisions in common, even though central administrative control was lacking.

The Swiss model

In Switzerland, as has been noted, cultural and linguistic groupings are sharply differentiated and maintained. Within itself each group may be separately and differently stratified, but access to power in the wider society does not seem to be limited by local or cultural origins, nor does it seem to be limited by any ranking of the separate cultural groupings.

In western Malaysia the population is evenly divided between Malays and Chinese. There is also a small minority of Indians and even smaller ones of aboriginals, Thai-speaking Buddhists, and a few other small groups. The aboriginal population is almost wholly separated from the rest, and the majority of the Malays live in rural districts, whereas the Chinese are mostly urban dwellers. The ethnic groups are not segregated by law. The political constitution provides that the men who direct the machinery of government and hold legislative powers shall be impartially selected by voting. In practice electoral procedures and the staffing of the army and the civil service have put most political power in the hands of Malays. The greatest economic strength is in the hands of the Chinese. This distribution makes a permanent balance of power difficult. The fortunes of the Chinese could be threatened by political action, and they might use their wealth to upset the political balance and so threaten the Malays. Under the colonial rule of the British, the Malays, the Chinese, and the Indians were unorganized ethnic categories; but since independence the categories have for political purposes become more and more organized as corporate ethnic groups. Neither the Malay nor Chinese group is completely dominant, and in the strict sense neither is a minority group.

The situation in Malaysia, where no ethnic group is completely able to dominate the central government, is exceptional. The situation in Kenya, Tanzania, or Uganda is more usual. Since independence the Africans have been the dominant group in all those countries, which also contain Arab, European, and Indian minorities. Like the Chinese in Malaysia, the Indians in East Africa are economically strong, but their numbers have always been far too few for them ever to think of making a bid to take over the central administration and run the country in their own interest. Indians, for various reasons, are now regarded by the Africans as an objectionable minority group, and for ideological reasons they are being driven out of Kenya and Uganda.

The effects of the rise of the nation-state

Toleration of minority groups depends in part, as noted earlier, on such factors as size, wealth, access to power, and ideology. In modern nation-states wide differences of ideology are not as a rule very willingly permitted, possibly because the idea of a nation implies that if the nation and the dominant group controlling it are to survive, its citizens ought to hold strong, common, national values. The difficulties in deciding how far divergent minorities and ideas can be tolerated is not new, nor are they confined to new nation-states with ethnic and cultural minority groups held together only by the political power of a dominant group. In the 16th century France was divided between Roman Catholic and Protestant Christians, and each group, for its own safety, felt that it had to capture the central government. The victorious Roman Catholics, for their own sake and the safety of the French nation, as they saw it, did not cease their war until the Protestant minority was virtually eliminated.

The idea of a nation-state and the idea of tolerating a minority are to some extent in opposition to one another. The very long-standing solution to the problem reached by a country like Switzerland is unusual. The balance reached in such countries as Malaysia or Canada sometimes gives the impression of a temporary solution, which may at any time give way to a unitary state dominated by one group or to the secession of the minority to form a new nation. The dilemma is important in many parts of the world, especially in such areas as the Caribbean, the United States, and southern and southeastern Asia, where cultural distinctions have often fused with political hostility. It is this widespread fusion of ethnic and cultural hostilities as one of the main elements of political programs that has undoubtedly drawn many to the concept of a plural society. The concept does not give a satisfactory account of the structural characteristics of pluralistic societies, but what its advocates may be attempting to describe and analyze is a political situation in which ethnic and cultural loyalties are felt to be more important than the interests of class or party or any other group in society. This transformation of ethnic and cultural hostilities into political programs is closely related to the problems of minority groups and their place in society.

In the past, pluralism has usually been discussed as if it were a feature only of societies with a central administration. It is, however, not uncommon to find ethnic and cultural groups, whether adjacent or intermingled, that live within a common system of relations constituting an alliance of some kind but not an effective central administration.

"Tribal pluralism"

In northern Thailand, Burma, Borneo, New Guinea, and other parts of the world small independent groups of ethnically and culturally distinct peoples were dispersed and mingled over wide areas without being subject in any effective sense to a central government. Among the Kachin of highland Burma, peoples of different languages and cultures, sometimes resident in the same villages, were united in a common network of marital, political, and trading alliances. On the Baram and Rajang rivers of Sarawak in Borneo, villages having different cultures and speaking different languages were interspersed along the banks. Each village was politically independent with its own characteristic social structure. In most villages the inhabitants were rigidly ranked: those of the lower ranks married and lived, for the most part, within their native villages; but those of the upper ranks, both men and women, married and often moved into other villages and other cultures for political and economic advantage. In most of these societies ranking was a fundamental principle of organization, but the different villages and cultures were not ranked against one another. Each was different from and none was considered better than the other. The system of alliances that held together a considerable population, culturally and linguistically divided, had certain similarities to the system by which modern nation-states maintain their independence and diversity.

The functions of minorities and ethnic groups. Minority groups have performed almost every type of function in almost all known types of social system. Minorities have played economic roles as specialist bankers, traders, or craftsmen; more often they have been manual labourers, whether slave, serf, or free. As revolutionary groups they have acted as ideological leaders in political life; sometimes their function has been to hold the political balance; more often they have been the subject of exploitation. Some, because they have special skills, may be valued for their artistic contributions to social life and yet not be altogether socially acceptable to the rest of society. Earlier it was noted that the role played by a minority group within a society varies with such factors as its size, its skills, and its ideology and also with the roles of the dominant and other groups in the society. The amounts of wealth that the members of the group can accumulate and the political influence that they can exercise vary with the ways in which these factors arrange themselves. Factors such as these obviously work differently in different types of society, and even in the same type the balance is never likely to be identical in any two societies.

Their economic roles. A division of labour is a feature of all societies, even the simplest, but unless the population is large enough to free specialists from the task of finding sufficient food and shelter for themselves and their dependents, the society is unlikely to be able to support wholly specialized groups. Without a sufficiently large population, too, a system of social stratification cannot develop; for a graded distribution of resources among the social strata is possible only if there is a sufficiency of basic goods for everybody.

In a system of closed stratification, in which movement from one stratum to another is forbidden, and in which resources are differentially allocated to the strata, functional specialization is likely to be a feature of the system. Specialized groups, handling resources that are the property, so to speak, of their members and of their stratum, are likely to resist change and defend their specialization and the resources that are needed for it. In an open system of stratification, on the other hand, where social mobility is the reward for talent of various kinds, the temptation to form perpetual, specialized minority groups is probably not so great.

In caste and estate systems

In a Hindu caste system the allocation of economic, political, and ritual roles is probably more formally and rigidly arranged than in almost any other kind of society; but it is not usual to consider the specialized subcastes, which make up the system, as minority groups. On the other hand, the outcaste groups, who were said by the Hindu to be altogether outside the system and whose menial occupations and servile structural position separated them, often physically, from the rest of society, are probably best understood when regarded as minority groups. Similarly, in a feudal system the roles of the estates were strictly prescribed: the nobles fought for all, the clergy prayed for all, and the commoners worked for all. Insofar as the nobles dominated society, both the clergy and the commoners could be considered minority groups, though it is not usual to call them so, for medieval European society had in it a number of smaller groups that were unambiguously minorities in the usual sense. The structure of a feudal system allows little place in it for trade, which indeed was not an approved activity in medieval Europe, though of course it was necessary and was carried on by specialized minority groups such as the Jews or by anomalous minorities such as the merchants in semi-independent towns.

In the colonial and postcolonial societies of the modern world, the population is more often than not ethnically and culturally diverse and is often also rigidly stratified. In these societies specialized minority groups are common. When the British established a colonial regime in East Africa at the end of the last century, the African population lacked most of the technical and trading skills that were needed to develop the area and produce a revenue for the administration. Immigrants from the west coast of India were therefore encouraged to move into the interior regions and set themselves up as a trading minority. They differed from the Africans and the Europeans in physique, language, religion, and culture. They were absorbed into the hierarchy of colonial society as a specialized minority group, with a virtual monopoly of retail trade and a position of wealth, influence, and prestige greater than that of the Africans but lower than that of the Europeans.

After independence the hierarchies in the East African countries were rearranged. The position of a non-African minority group with a major share of the country's retail trade was ideologically unacceptable to the new African rulers. The countries they ruled were ethnically and culturally highly diversified, and their main concern was to encourage the growth of a unified nation, whose citizens shared common, patriotic values, even, if need be, at the expense of equality of opportunity. The rulers therefore adopted policies designed to eliminate non-African minorities such as the Indians. With local variations the same situation can be found in all parts of Southeast Asia, where rulers of dominant groups have attempted either to absorb or to drive away Chinese and Indian minority groups that have specialized in trade or crafts.

Minority groups have occupied servile or labouring positions in every type of society. The San (Bushmen) of Botswana in southern Africa are a minority group loosely attached as herdsmen and servants to Tswana society. Pagan tribes, immigrants, and conquered groups have on many occasions been brought into Hindu society as servile minorities, usually as outcasts outside the caste system, but sometimes (though more rarely) they have been incorporated as subcastes on the lower rungs of the hierarchy. In England after the Norman conquest the servile roles of the feudal system were mostly occupied by the previously free Saxons, the Welsh, and other native minority groups. Ethnic and cultural minority groups of African slaves and Amerindians filled the lower strata of the rigidly organized class systems of the 18th century in the Caribbean area. In the less rigid class system of modern Britain immigrant minorities from the Caribbean, Pakistan, and India for the most part are labourers.

The labour of ethnic groups

Because the distinguishing marks of an ethnic group are physically carried by its members, it is harder for them to be assimilated. And if racial attributes are important in the ideology of the dominant group of the society, it may be easier for a cultural minority to improve its status than it is for an ethnic group. The Irish who settled in the United States as labourers and various eastern Europeans who followed them have found it possible (at least in the second or third generation) to be socially mobile by changing their cultural identities in a way that has not yet been really open to blacks. It is not entirely accidental that ethnic groups tend to be more segregated and restricted and therefore regularly found among the poorer classes of a society. As long as the ethnic groups reside in the lower ranks of the social system, the privileged classes—unthreatened—can talk easily and abstractly of the desirability of equality of opportunity and the undesirability of segregation and restriction of opportunity. If an ethnic minority does acquire wealth or influence, however, either through its own efforts within society or through circumstances outside it, then the dominant group is usually compelled to make the necessary changes in its practices as well as ideology to accommodate if not assimilate the minority. But the process is never easy and is seldom willingly undertaken.

Their political roles. The political roles open to a minority depend on many of the same factors as limit its economic opportunities. Its numbers may be too small, as were those of the Indian immigrants in East Africa, for it ever to contemplate making a bid for control of the central political machinery, whereas in Malaysia the Chinese were sufficiently numerous to threaten a takeover of the government and thus could protect themselves. In fact, they accomplished the takeover in Singapore. Many minorities, however—as, for instance, merchants whose main interest in a political system is the preservation of law and order needed for peaceful trading—do not possess an active political ideology. It is only if circumstances force them and the structure of the society permits it that they think of protecting themselves by direct action. A minority group by definition is subordinate, and its usual position is near the bottom of the power structure, especially if its values diverge markedly from those of the main body politic.

Minorities acquiring power and leadership

The opportunities for acquiring power depend not only on a minority's own characteristics but also on the nature of the larger society. In a Hindu caste system the hierarchy is one of ritual prestige. The priestly groups are endowed with the greatest sanctity and honour. But ideally secular power ought not to be in their hands. It is the prerogative of the administrative or warrior castes. In practice, political power under Hinduism has often been held by dominant priestly or merchant castes, and even by labouring castes, without wrecking the system. In theory, too, a society of feudal estates ought to be ruled only by nobles; but if, as happened at the Norman invasion of England, the conquerors were of mixed origins, they were soon attributed noble status and the system survived.

In an open class system, if a minority group achieves power in a modern society and displaces a dominant group, the distribution of roles among most of the ruled classes is unlikely to be fundamentally altered, whatever the fate of individuals. This situation can be documented from many parts of the Communist and the postcolonial worlds. In Ghana under the British colonial regime, the African elite was a minority group that was drawn from several indigenous linguistic and cultural groups and whose members were often foreign educated and English speaking. When this group took over the colonial administration from the British as the dominant group, the society in its broad essentials altered little, except that the Ghanian elite ceased to be a minority group.

A favourably placed minority may occupy a strategically important position in a political system without being the dominant group. After their defeat in the Boer War at the beginning of this century, the Afrikaners were sociologically speaking a minority group though numerically a majority of the white population of South Africa. Their ideology differed from that of the dominant English group, though both had much in common concerning ideas of race, stratification, and the political superiority of whites. These ideas were directly opposed to the idea of equality of opportunity, which was paradoxically held concurrently by the English. The ideology of the Afrikaners was less ambiguous, and in a real sense they were the ideological leaders of the whites in South Africa. When they eventually gained control of the central administrative system and, as the dominant group, began to put their ideas on race and minorities into practice, there was little real resistance.

In South Africa a linguistic and cultural minority was able, as in other parts of the postcolonial world, to become dominant without a military conquest or revolution. In the Soviet Union, in Mainland China, and in other Communist countries, the Communist Party has usually gained power by force. In all countries the party began as a minority group based on a political philosophy in which differences of race, language, and culture were not considered to be relevant criteria of membership. In the period before it gained control of the central administrative machinery, the party was undoubtedly, especially in many of the undeveloped tropical areas, a minority group that acted as an ideological leader and an instrument of social change.

A minority group may be in an important political position in a society and yet not be able to take over the administration and rule it in its own interests. In parts of the European colonial empires, for example, trading immigrants were disenfranchised minorities; but because their economic power was great in the countries in which they had settled, they could frequently persuade, and sometimes force, colonial rulers to modify policies unfavourable to trade, though possibly advantageous to the rulers or other groups in the society. In Hawaii in the 1890s the white men who dominated the economy even persuaded the United States to annex the islands after a new U.S. tariff had threatened their sugar exports.

In the latter part of the 19th century the Irish formed a minority group in Great Britain, and their parliamentary representatives in London were sufficiently numerous to hold the balance for a time in the political struggles of the rest of the United Kingdom. A not dissimilar position was held for a few years by the Cape Coloureds of the Cape Province of South Africa before their vote was removed.

The risks of arbitration

The role of arbiter, of choosing whether to give or refuse support, can be a dangerous one for a minority group. Even that of a "go-between" who mediates between the rulers and the ruled members of society may bring hatred and disaster in its train, as the Anglo-Indian community discovered when British rule ended in India; and as Chinese and Indian groups who had specialized in working with colonial civil services also discovered when former colonial territories gained independence. In situations where a minority holds a balance of power, it may feel obliged for the protection of its own interests, as were the Afrikaners, to become dominant in the society. On the other hand, it may be safer for the minority, if it can, to secede altogether and become independent before it is stripped, as was the Cape Coloured population of South Africa, of all political power. The Irish solved the dilemma by secession from the United Kingdom, and the Chinese of Singapore by secession from Malaysia.

In the discussion of the economic functions of minority groups it was noted that most often they were servile or exploited and that both the form and function of the group largely depended on the form and ideology of the society of which they were a part. It was also seen that the degree of subordination or exploitation and the type of relation that a minority can have with a dominant group is variable. Thus, for instance, Brāhmaṇ priests in a caste system are the ritual servants of the other subcaste groups and may be exploited politically and economically by a lower ranked but dominant group, and yet the ritual relationship is not a servile one. Similarly in feudal Europe the nobles and the clergy exploited and segregated minorities such as the Jewish trading community, but the relationship differed in its degree of subordination and exploitation from the relationship of the nobles to serfs and slaves. Nor was the relationship of the Indian trading community in East Africa with the British rulers a wholly subordinate one, although the British exploited and used the Indians' commercial skills and political weakness. In British Guiana or Fiji the situation was different. There Indian coolies were imported as indentured plantation labourers when the suppression of the slave trade put an end to that source of cheap African labour, and as a result both the freed African and the indentured Indians, though not technically slaves, occupied an almost servile status. (H.S.M.)

BIBLIOGRAPHY. The principal general studies of social differentiation are GEORG SIMMEL, *Über soziale Differenzierung: Soziologische und psychologische Untersuchungen* (1890); and C.C. NORTH, *Social Differentiation* (1926). Differentiation in terms of social roles is discussed in R. LINTON, *The Study of Man* (1936); H.H. GERTH and C. WRIGHT MILLS, *Character and Social Structure* (1953); and M.P. BANTON, *Roles* (1965).

Age differentiation is discussed in KARL MANNHEIM, "The Problem of Generations," in *Essays on the Sociology of Knowledge,* 2nd ed. (1952); S.N. EISENSTADT, *From Generation to Generation* (1956); and MARGARET MEAD, *Culture and Commitment: A Study of the Generation Gap* (1970). The division of labour is treated at length in EMILE DURKHEIM, *De la division du travail social* (1893; Eng. trans., *The Division of Labour in Society,* 1933), which also deals with the differentiation of societies into types. Other classifications of the types of society include FERDINAND TONNIES, *Gemeinschaft und Gesellschaft* (1887; Eng. trans., *Community and Association,* 1955), KARL MARX and FRIEDRICH ENGELS, *Die deutsche Ideologie* (1846; Eng. trans. of pt. 1 and 3, *The German Ideology,* 1938), *Manifest der kommunistischen Partei* (1848; Eng. trans., *The Communist Manifesto,* 1933); RAYMOND ARON, *The Industrial Society* (1967). The literature on stratification is extensive, but the following works provide a view of the major aspects: S. OSSOWSKI, *Class Structure in the Social Consciousness* (Eng. trans., 1963); MAX WEBER, "Class, Status, Party," in *Wirtschaft und Gesellschaft* (1922), Eng. trans. in H.H. GERTH and C. WRIGHT MILLS (eds.), *From Max Weber* (1946); J.A. SCHUMPETER, "Social Classes in an Ethnically Homogeneous Environment," *Imperialism and Social Classes* (1951); KINGSLEY DAVIS and WILBERT MOORE, "Some Principles of Stratification," in R. BENDIX and S.M. LIPSET (eds.), *Class, Status and Power,* 2nd rev. ed. (1966). On elite groups, see C. WRIGHT MILLS, *The Power Elite* (1956); and T.B. BOTTOMORE, *Elites and Society* (1964). There have been many studies of social mobility; two national inquiries which also deal with wider issues are D.V. GLASS (ed.), *Social Mobility in Britain* (1954); and PETER BLAU and OTIS D. DUNCAN, *The American Occupational Structure* (1967). See also ABNER COHEN, *The Politics of Elite Culture* (1981); and ARTHUR MARWICK, *Class: Image and Reality in Britain, France and the USA since 1930* (1980).

Women and sex differentiation: F. BEACH (ed.), *Sex and Behaviour* (1965); and E.E. MACCOBY (ed.), *The Development of Sex Differences* (1967), symposia by well-known experts discussing sex differences in temperament and abilities and their possible origin and explanation; SIGMUND FREUD, "The Psychology of Women," in *Neue Folge der Vorlesungen zur Einführung in die Psychoanalyse* (1933; Eng. trans., *New Introductory Lectures on Psychoanalysis,* ch. 33, 1942); and "Some Psychological Consequences of the Anatomical Distinction Between the Sexes," *Int. J. Psychoanal.,* 8:133–142 (1927), two classical formulations by the founder of psychoanalysis on his interpretation of feminine psychology; VIOLA KLEIN, *The Feminine Character: History of an Ideology* (1946, 1971), a comparative analysis of various theories about the psychology of women; MARGARET MEAD, *Sex and Temperament in Three Primitive Societies* (1935) and *Male and Female* (1949), anthropological studies of temperamental sex differences, based on field work in primitive societies; L.M. TERMAN and C.C. MILES, *Sex and Personality* (1936, reprinted 1968); and L.M. TERMAN and L.E. TYLER, "Psychological Sex Differences," in L. CARMICHAEL (ed.), *Manual of Child Psychology* (1954; 3rd ed., 1970), psychometric studies of personality differences between men and women; ELIZABETH BOTT, *Family and Social Network* (1957), a small-scale, intensive investigation of the connection between the distribution of roles between husbands and wives and their wider network of family and other social relationships; W.J. GOODE, *World Revolution and Family Patterns* (1963), a study by the well-known American sociologist on current changes in the structure and form of the family and their connection with technical and industrial developments; A.C. KINSEY *et al., Sexual Behavior in the Human Female* (1953); A. MYRDAL and VIOLA KLEIN, *Women's Two Roles: Home and Work,* 2nd rev. ed. (1968), a comparative study of the changing role of women in four highly industrialized countries (United States, Britain, France, and Sweden), its causes and consequences; F.I. NYE and L.W. HOFFMAN (eds.), *The Employed Mother in America* (1963), an anthology of detailed and statistically substantiated studies of the manifold aspects of maternal employment in the U.S.: the reason why so many married women go out to work, what effect this employment has on different types of children and in different situations, on the relationships between husbands and wives and on the women themselves; RAPHAEL PATAI (ed.), *Women in the Modern World* (1967), a symposium of essays by 24 social scientists, mostly women, from different parts of the world, describing recent changes in the status of women in their respective countries, including Asian, American, African, and European nations. See also LISA LEGHORN and KATHERINE PARKER, *Woman's Worth: Sexual Economics and the World of Women* (1981).

Youth: E.A.M. DOUVAN and J. ADELSON, *The Adolescent Experience* (1966), is an authoritative description of normal as opposed to deviant youth, and is particularly perceptive on differential sex roles. S.N. EISENSTADT, *From Generation to Generation* (1956), remains perhaps the most complete sociological analysis of the status, function, and role of youth in modern societies and is valuable for understanding peer groups. J.S. COLEMAN, *The Adolescent Society* (1961); PAUL GOODMAN, *Growing Up Absurd* (1960); J.B. MAYS, *The Young Pretenders* (1965); E.A. SMITH, *American Youth Culture* (1962), examine a variety of aspects of teenage culture in several different societies. D. GOTTLIEB, J. REEVES, and W.D. TENHOUTEN, *The Emergence of Youth Societies* (1966), provides a bibliography for cross-cultural analysis that attempts to include the whole world; and MARGARET MEAD, *Coming of Age in Samoa* (1943, reprinted 1961), and *Growing Up in New Guinea* (1942, reprinted 1962), are interesting studies of more primitive cultures that have a bearing on modern industrial urban life. Psychological works that complement and expand the sociological approach to adolescence are among the most valuable texts on the subject. P. BLOS, *On Adolescence* (1962), is a well-known essay; and C.M. FLEMING, *Adolescence: Its Social Psychology,* 2nd rev. ed. (1963), synthesizes the psychological, social, and physiological aspects. ERIK ERIKSON, *Childhood and Society,* 2nd rev. ed. (1963), and *Identity, Youth and Crisis* (1968), are classic statements of the psychoanalytical approach to adolescent personality structure in relation to cultural influences. M. ROSENBERG, *Society and the Adolescent Self-Image* (1965), gives the results of large-scale study of American children and young people. R.J. HAVIGHURST, *Human Development and Education* (1953), examines at length the important idea of the various developmental tasks facing the adolescent in general, while R.F. PECK *et al., The Psychology of Character Development* (1960), offers research evidence to show that, in spite of widespread views to the contrary, the influence of parents on their children's moral values is more often than not still paramount. PATRICIA M. SPACKS, *Adolescent Idea: Myths of Youth and the Adult Imagination* (1981), explores the conceptions adults maintain about adolescents.

The aged: Extensive empirical and theoretical work has been collected in the following three handbooks, comprising essays on varied aspects of gerontology: J.E. BIRREN (ed.), *Handbook of Aging and the Individual: Psychological and Biological Aspects* (1959); C. TIBBITTS (ed.), *Handbook of Social Gerontology: Societal Aspects of Aging* (1960); and E.W. BURGESS (ed.), *Aging in Western Societies* (1960). More recent information on several of these topics may be found in a special issue of the *International Social Science Journal,* entitled "Old Age," vol. 15, no. 3 (1963). Findings from some 3,000 social science studies concerning people in their middle and later years have been condensed and organized for ready reference in M.W. RILEY and A. FONER *et al., Aging and Society,* vol. 1, *An Inventory of Research Findings* (1968). The findings from this inventory are interpreted for use in those professions concerned with older people in M.W. RILEY, J.W. RILEY, JR., and M.E. JOHNSON (eds.), *Aging and Society,* vol. 2, *Aging and the Professions* (1969). Systematic research findings for selected times and places appear in the analysis of ethnographic reports in L.W. SIMMONS, *The Role of the Aged in Primitive Society* (1945); H.D. SHELDON, *Older Population of the United States* (1958); and in the parallel surveys reported for Great Britain, Denmark, and the U.S. in E. SHANAS *et al.*, *Old People in Three Industrial Societies* (1968).

Race differentiation: T.W. ADORNO *et al.*, *The Authoritarian Personality* (1950), an influential study of the psychodynamics of prejudice from a psychoanalytical perspective; G.W. ALLPORT, *The Nature of Prejudice* (1954), a standard text by a social psychologist; B.N. COLBY and P.L. VAN DEN BERGHE, "Ethnic Relations in Southeastern Mexico," *American Anthropologist*, 63:772–792 (Aug. 1961), a description of a system of ethnic relations between Maya Indian groups and Spanish-speaking ladinos; O.C. COX, *Caste, Class and Race* (1948), a detailed critique of the field of race and ethnic relations by a Marxian sociologist; A.W. LIND (ed.), *Race Relations in World Perspective* (1956), a collection of essays dealing with racial situations in various parts of the world; A. MARCHANT, *From Barter to Slavery* (1942), the early history of racial and cultural contacts in Brazil; L. MARQUARD, *The Peoples and Policies of South Africa* (1952), the most concise introduction to racial problems in South Africa; GUNNAR MYRDAL, *An American Dilemma* (1944), a monumental study of black America in the 1940s; R.E. PARK, *Race and Culture* (1950), a collection of sociological essays on race and culture; T.F. PETTIGREW, *A Profile of the Negro American* (1964), a useful summary of studies on black Americans; D. PIERSON, *Negroes in Brazil: A Study of Race Contact at Bahia* (1942), a pioneer study of Afro-Brazilians in the North East; T. SHIBUTANI and K.M. KWAN, *Ethnic Stratification: A Comparative Approach* (1965), a text on comparative race and ethnic relations; G.E. SIMPSON and J.M. YINGER, *Racial and Cultural Minorities*, 3rd ed. (1965), a standard text stressing North America; K.M. STAMPP, *The Peculiar Institution: Slavery in the Ante-Bellum South* (1964), a good account of antebellum slavery in the U.S.; F. TANNENBAUM, *Slave and Citizen: The Negro in the Americas* (1947), a comparative study of slavery in the Americas; P.L. VAN DEN BERGHE, *Race and Racism* (1967), a text on comparative race relations in Mexico, Brazil, the U.S., and South Africa, and *South Africa: A Study in Conflict* (1965), a comprehensive sociological analysis of South Africa; C. WAGLEY, *Amazon Town: A Study of Man in the Tropics* (1964), a community study of a multiracial town in the Amazon basin, and (ed.), *Race and Class in Rural Brazil* (1952), a series of studies of race relations in various regions of Brazil; and R.M. WILLIAMS *et. al.*, *Strangers Next Door: Ethnic Relations in American Communities* (1964), a compilation of research findings on race and ethnic relations in the U.S.; STEPHEN STEINBERG, *The Ethnic Myth: Race, Ethnicity, and Class in America* (1981), depreciates cultural values and emphasizes class to explain differential race mobility.

Social class: Two short introductions to the study of social class are KURT B. MAYER and WALTER BUCKLEY, *Class and Society*, 3rd ed. rev. (1969); and T.B. BOTTOMORE, *Classes in Modern Society*, 2nd ed. (1966). The historical development of ideas and theories of class has not been very fully discussed, but some treatment is given in STANISLAW OSSOWSKI, *Struktura klasowa w społecznej świadomości* (1957; Eng. trans., *Class Structure in the Social Consciousness*, 1963), which also contains a good account of the Marxist theory. Another exposition and critical examination of Marx's theory is provided by RALF DAHRENDORF, *Soziale Klassen und Klassenkonflikt in der industriellen Gesellschaft* (1957; Eng. trans., *Class and Class Conflict in Industrial Society*, 1959). The ideas of the early American sociologists are discussed in CHARLES H. PAGE, *Class and American Sociology*, new ed. (1969), which contains a long introduction surveying much of the research on social class and mobility over the past 30 years. MAX WEBER'S writings on social class and status are to be found mainly in his work *Wirtschaft und Gesellschaft*, 4th ed., 2 vol. (1922; Eng. trans., *Economy and Society*, ed. by GUENTHER ROTH and CLAUS WITTICH, 3 vol., 1968). The theory of elites is expounded by VILFREDO PARETO in *Trattato di sociologia generale*, 2nd ed., 3 vol. (1923; Eng. trans., *The Mind and Society*, 4 vol., 1935); more recent studies include C. WRIGHT MILLS, *The Power Elite* (1956); and T.B. BOTTOMORE, *Elites and Society* (1964). Among the general studies of class are REINHARD BENDIX and S.M. LIPSET (eds.), *Class, Status and Power*, 2nd ed. (1966); and GERHARD E. LENSKI, *Power and Privilege* (1966). Particular social classes are discussed in C. WRIGHT MILLS, *White Collar* (1951); G.A. BRIEFS, *The Proletariat* (1937); J.H. GOLDTHORPE *et al.*, *The Affluent Worker*, 3 vol. (1968–69); and ERIC R. WOLF, *Peasants* (1966).

Social mobility: A pioneer survey of social mobility is P.A. SOROKIN, *Social Mobility* (1927, reprinted with an additional chapter on cultural mobility, 1959). Two major national studies are D.V. GLASS (ed.), *Social Mobility in Britain* (1954); and PETER M. BLAU and OTIS D. DUNCAN, *The American Occupational Structure* (1967). Comparisons between societies are discussed in S.M. LIPSET and REINHARD BENDIX, *Social Mobility in Industrial Society* (1959); and S.M. MILLER, "Comparative Social Mobility," *Current Sociology*, 9:1–89 (1960). Some aspects of the social context in which mobility occurs are examined, with reference to the United States, in ANSELM L. STRAUSS, *The Contexts of Social Mobility* (1971).

Caste: L. DUMONT, *Homo Hierarchicus* (1966; Eng. trans. 1970), description, history, and theories of the Hindu caste system in contrast to Western philosophies, with comprehensive bibliography up to 1962; J. SILVERBERG (ed.), *Social Mobility in the Caste System in India* (1968), corrects Western stereotypes. Important classical analyses are those of C. BOUGLE, *Essais sur le régime des castes* (1908; Eng. trans., *Essays on the Caste System*, 1971); A.M. HOCART, *Les Castes* (1938; Eng. trans., *Caste: A Comparative Study*, 1950); J.H. HUTTON, *Caste in India*, 4th ed. (1963); and MAX WEBER, *Hinduismus und Buddhismus* in vol. 2 of *Gesammelte Aufsätze zur Religionssoziologie* (1921; Eng. trans., *The Religion of India: The Sociology of Hinduism and Buddhism*, 1958). D.M. SCHNEIDER, *American Kinship* (1968), provides the theoretical approach used here; it was first applied to an Indian caste by S. Barnett. P.V. KANE, *History of Dharmáśāstra*, 5 vol. (1930–62), an encyclopaedic digest of Vedic codes; G.S. GHURYE, *Caste and Race in India*, 5th ed. rev. (1969), historical interpretation of social developments; the *Census of India* (1881–1971), statistics and descriptions; D.G. MANDELBAUM, *Society in India*, 2 vol. (1970), summary of recent field studies with bibliography; M. MARRIOTT, *Caste Ranking and Community Structure in Five Regions of India and Pakistan* (1960), ecological comparisons and bibliography of older descriptive works.

Minority groups: H. KOHN, *The Idea of Nationalism* (1944); L. WIRTH, "The Problem of Minority Groups" in R. LINTON (ed.), *The Science of Man in the World Crisis* (1945); and A.M. and C.B. ROSE (eds.), *Minority Problems* (1965), give a good background to the problems implicit in the ideas of nationalism and minorities, especially with reference to the West. L. KUPER and M.G. SMITH (eds.), *Pluralism in Africa* (1969), and H.S. MORRIS, *Indians in Uganda* (1968), examine the problem of ethnic and minority groups in non-Western societies and consider the subject of social and cultural pluralism in terms first raised by J.S. FURNIVALL in *Colonial Policy and Practice: A Comparative Study of Burma and Netherlands India* (1948). G.A. DE VOS and H. WAGATSUMA (eds.), *Japan's Invisible Race* (1966), discuss the same problems in a psychological rather than a sociological idiom. P. MASON, *Patterns of Dominance* (1970), treats a selected number of societies in different parts of the world and at different historical periods and evaluates various solutions for the problems of ethnic and cultural minorities. Of special interest for minority problems in the Western world are M. FREEDMAN (ed.), *A Minority in Britain: Social Studies of the Anglo-Jewish Community* (1955); E.J.B. ROSE *et al.*, *Colour and Citizenship* (1969); and C. WAGLEY and M. HARRIS, *Minorities in the New World* (1958). B. BENEDICT, *Indians in a Plural Society* (1961); F. and L.O. DOTSON, *The Indian Minority Zambia, Rhodesia and Malawi* (1968); S. PATTERSON, *Colour and Culture in South Africa* (1953); and D.E. WILLMOTT, *The Chinese of Semarang* (1960), give useful and illuminating descriptive accounts of ethnic groups and minorities outside Europe and America. The essays in F. BARTH (ed.), *Ethnic Groups and Boundaries: The Social Organization of Culture Difference* (1969), re-examine the traditional problems of maintaining ethnic or minority identities. STANLEY LIEBERSON, *A Piece of the Pie: Blacks and White Immigrants Since 1880* (1980), examines the different economic progress made by blacks and other minority groups; THOMAS SOWELL, *Markets and Minorities* (1981), studies the effect of discrimination on the economic progress of various minority groups.

The Social Sciences

The social sciences, which deal with human behaviour in its social and cultural aspects, include the following disciplines: cultural (or social) anthropology, sociology, social psychology, political science, and economics. Also frequently included are social and economic geography (see GEOGRAPHY) and those areas of education that deal with the social contexts of learning and the relation of the school to the social order (see TEACHING). History is regarded by many as a social science, and certain areas of historical study are almost indistinguishable from work done in the social sciences. Most historians, however, still consider history as one of the humanities. It is generally best, in any case, to consider history as marginal to the humanities and social sciences, since its insights and techniques pervade both (see HISTORY).

The study of comparative law may also be regarded as a part of the social sciences, although it is ordinarily pursued in schools of law rather than in departments or schools containing most of the other social sciences.

During the past quarter of a century, the term behavioral sciences has become more and more commonly used for the disciplines cited as social sciences. Those who favour this term do so in part because these disciplines are thus brought closer to some of the sciences, such as physical anthropology, linguistics, and physiological psychology, which also deal with human behaviour. Whether the term behavioral sciences will in time supplant "social sciences" or whether it will, as neologisms so often have before, disappear without trace in a few years is impossible to say. For the purposes of this article, the two terms may be considered synonymous.

The concern in this article is much less with the conclusions and results of the individual social sciences than with their disciplinary identity in modern society. The concern is with the social sciences as vital elements in the aftermath of the two great revolutions, the political and industrial, which opened the 19th century, with the pattern the social sciences assumed in that century, and their extraordinary development in the 20th century.

The article is divided into the following sections:

History of the social sciences

Although, strictly speaking, the social sciences do not precede the 19th century—that is, as distinct and recognized disciplines of thought—one must go back farther in time for the origins of some of their fundamental ideas and objectives. In the largest sense, the origins go all the way back to the ancient Greeks and their rationalist inquiries into the nature of man, state, and morality. The heritage of both Greece and Rome is a powerful one in the history of social thought as it is in so many other areas of Western society. Very probably, apart from the initial Greek determination to study all things in the spirit of dispassionate and rational inquiry, there would be no social sciences today. True, there have been long periods of time, as during the Western Middle Ages, when the Greek rationalist temper was lacking. But the recovery of this temper, through texts of the great classical philosophers, is the very essence of the Renaissance and the Age of Reason in modern European history. With the Age of Reason, in the 17th and 18th centuries, one may begin.

HERITAGE OF THE MIDDLE AGES AND THE RENAISSANCE

The same impulses that led men in that age to explore the earth, the stellar regions, and the nature of matter led them also to explore the institutions around them: state, economy, religion, morality; above all, the nature of man himself. It was the fragmentation of medieval philosophy and theory, and, with this, the shattering of the medieval world view that had lain deep in thought until about the 16th century, that was the immediate basis of the rise of the several strands of specialized thought that were to become in time the social sciences.

Adverse effects of medieval theology

Medieval theology, especially as it appears in St. Thomas Aquinas' *Summa theologiae,* contained and fashioned syntheses from ideas about man and society—ideas indeed that may be seen to be political, social, economic, anthropological, and geographical in their substance. But it was partly this close relation between medieval theology and ideas of the social sciences that accounts for the longer time it took these ideas—by comparison with the ideas of the physical sciences—to achieve what one would today call scientific character. From the time of the great Roger Bacon in the 13th century, there were at least some rudiments of physical science that were largely independent of medieval theology and philosophy. Historians of physical science have no difficulty in tracing the continuation of this experimental tradition, primitive and irregular though it was by later standards, throughout the Middle Ages. Side by side with the kinds of experiment made notable by Roger Bacon were impressive changes in technology through the medieval period and then, in striking degree, in the Renaissance. Efforts to improve agricultural productivity; the rising utilization of gunpowder, with consequent development of guns and the problems that they presented in ballistics; growing trade, leading to increased use of ships and improvements in the arts of navigation, including use of telescopes; and the whole range of such mechanical arts in the Middle Ages and Renaissance as architecture, engineering, optics, and the construction of watches and clocks—all of this put a high premium on a pragmatic and operational understanding of at least the simpler principles of mechanics, physics, astronomy, and, in time, chemistry.

In short, by the time of Copernicus and Galileo in the 16th century, a fairly broad substratum of physical science existed, largely empirical but not without theoretical implications on which the edifice of modern physical science could be built. It is notable that the empirical foundations of physiology were being established in the studies of the human body being conducted in medieval schools of medicine and, as the career of Leonardo da Vinci so resplendently illustrates, among artists of the Renaissance, whose interest in accuracy and detail of painting and sculpture led to their careful studies of human anatomy.

Very different was the beginning of the social sciences. In the first place, the church, throughout the Middle Ages and even into the Renaissance and Reformation, was much more attentive to what scholars wrote and thought about man's mind and his behaviour in society than it was toward what was being studied and written in the physical sciences. From the church's point of view, while it might be important to see to it that thought on the physical world corresponded as far as possible to what Scripture said—witnessed, for example, in the famous questioning of Galileo—it was far more important that such correspondence exist in matters affecting the nature of man, his mind, spirit, and soul. Nearly all the subjects and questions that would form the bases of the social sciences in later centuries were tightly woven into the fabric of medieval scholasticism, and it was not easy for even the boldest minds to break this fabric.

Adverse effects of reverence for classics and Cartesianism

Then, when the hold of scholasticism did begin to wane, two fresh influences, equally powerful, came on the scene to prevent anything comparable to the pragmatic and empirical foundations of the physical sciences from forming in the study of man and society. The first was the immense appeal of the Greek classics during the Renaissance, especially those of the philosophers Plato and Aristotle. A great deal of social thought during the Renaissance was little more than gloss or commentary on the Greek classics. One sees this throughout the 15th and 16th centuries.

Second, in the 17th century appeared the powerful influence of the philosopher René Descartes. Cartesianism, as his philosophy was called, declared that the proper approach to understanding of the world, including man and society, was through a few simple, fundamental ideas of reality and, then, rigorous, almost geometrical deduction of more complex ideas and eventually of large, encompassing theories, from these simple ideas, all of which, Descartes insisted, were the stock of common sense—the mind that is common to all human beings at birth. It would be hard to exaggerate the impact of Cartesianism on social and political and moral thought during the century and a half following publication of his *Discourse on Method* and his *Meditations.* Through the Age of Reason and down through the Enlightenment in the later 18th century, the spell of Cartesianism was cast on nearly all those who were concerned with the problems of the nature of man and society.

Both of these great influences, reverence for the classics and fascination with the geometrical-deductive procedures advocated by Descartes must be seen from today's vantage point as among the major influences retarding the development of a science of society comparable to the science of the physical world. It is not as though data were not available in the 17th and 18th centuries. The emergence of the national state carried with it evergrowing bureaucracies concerned with gathering information, chiefly for taxation, census, and trade purposes, which might have been employed in much the same way that physical scientists employed their data. The voluminous and widely published accounts of the great voyages that had begun in the 15th century, the records of soldiers, explorers, and missionaries who perforce had been brought into often long and close contact with primitive and other non-Western peoples, provided still another great reservoir of data, all of which might have been utilized in scientific ways as such data were to be utilized a century or two later in the social sciences. Such, however, was the continuing spell cast by the texts of the classics and by the strictly rationalistic, overwhelmingly deductive procedures of the Cartesians that, down until the beginning of the 19th century, these and other empirical materials were used, if at all, solely for illustrative purposes in the writings of the social philosophers.

HERITAGE OF THE ENLIGHTENMENT

There is also the fact that, especially in the 18th century, reform and even revolution were often in the air. The purpose of a great many social philosophers was by no means restricted to philosophic, much less scientific, understanding of man and society. The dead hand of the Middle Ages seemed to many vigorous minds in western Europe the principal force to be combatted, through critical reason, enlightenment, and, where necessary, major reform or revolution. One may properly account a great deal of this new spirit to the rise of humanitarianism in modern Europe and in other parts of the world and to the spread of literacy, the rise in the standard of living, and the recognition that poverty and oppression need not be the fate of the masses. The fact remains, however, that social reform and social science have different organizing principles, and the very fact that for a long time, down indeed through a good part of the 19th century, social reform and social science were regarded as pretty much the same thing could not have helped but retard the development of the latter.

Nevertheless, it would be wrong to discount the significant contributions to the social sciences that were made during the 17th and 18th centuries. The first and greatest of these was the spreading ideal of a science of society, an ideal fully as widespread by the 18th century as the ideal of a physical science. Second was the rising awareness of the multiplicity and variety of human experience in the world. Ethnocentrism and parochialism, as states of mind, were more and more difficult for educated people to maintain given the immense amount of information about—or, more important, interest in—non-Western peoples, the results of trade and exploration. Third was the spreading sense of the social or cultural character of human behaviour in society—that is, its purely historical or conventional, rather than biological, basis. A science of society, in short, was no mere appendage of biology but was instead a distinct discipline, or set of disciplines, with its own distinctive subject matter.

Ideas of structure and developmental change

To these may be added two other very important contributions of the 17th and 18th centuries, each of great theoretical importance. The first was the idea of structure. First seen in the writings of such philosophers as Hobbes, Locke, and Rousseau with reference to the political structure of the state, it had spread by the mid-18th century to highlight the economic writings of the Physiocrats and Adam Smith. The idea of structure can also be seen in certain works relating to man's psychology and, at opposite reach, to the whole of civil society. The ideas of structure that were borrowed from both the physical and biological sciences were fundamental to the conceptions of political, economic, and social structure that took shape in the 17th and 18th centuries. And these conceptions of structure have in many instances, subject only to minor changes, come down to 20th-century social science.

The second major theoretical idea was that of developmental change. Its ultimate roots in Western thought, like those indeed of the whole idea of structure, go back to the Greeks, if not earlier. But it is in the 18th century, above all others, that the philosophy of developmentalism took shape, forming a preview, so to speak, of the social evolutionism of the next century. What was said by such writers as Condorcet, Rousseau, and Adam Smith was that the present is an outgrowth of the past, the result of a long line of development in time, and, furthermore, a line of development that has been caused, not by God or fortuitous factors, but by conditions and causes immanent in human society. Despite a fairly widespread belief that the idea of social development is a product of prior discovery of biological evolution, the facts are the reverse. Well before any clear idea of genetic speciation existed in European biology, there was a very clear idea of what might be called social speciation—that is, the emergence of one institution from another in time and of the whole differentiation of function and structure that goes with this emergence.

As has been suggested, these and other seminal ideas were contained for the most part in writings, the primary function of which was attack on the existing order of government and society in western Europe. Another way of putting the matter is to say that they were clear and acknowledged parts of political and social idealism—using that word in its largest sense. Hobbes, Locke, Rousseau, Montesquieu, Adam Smith, and other major philosophers had as vivid and energizing sense of the ideal—ideal state, ideal economy, ideal civil society—as any earlier utopian writer. These men were, without exception, committed to visions of the good or ideal society. Their interest in the "natural"—that is, natural morality, religion, economy, or education, in contrast to the merely conventional and historically derived—sprang as much from the desire to hold a glass up to a surrounding society that they disliked as from any dispassionate urge simply to find out what man and society are made of. The fact remains, however, that the ideas that were to prove decisive in the 19th century, so far as the social sciences were concerned, arose during the two centuries preceding.

THE 19TH CENTURY

Effects of the democratic and industrial revolutions

The fundamental ideas, themes, and problems of the social sciences in the 19th century are best understood as responses to the problem of order that was created in men's minds by the weakening of the old order, or European society, under the twin blows of the French Revolution and the Industrial Revolution. The breakup of the old order—an order that had rested on kinship, land, social class, religion, local community, and monarchy—set free, as it were, the complex elements of status, authority, and wealth that had been for so long consolidated. In the same way that the history of 19th-century politics, industry, and trade is basically about the practical efforts of human beings to reconsolidate these elements, so the history of 19th-century social thought is about theoretical efforts to reconsolidate them—that is, to give them new contexts of meaning.

In terms of the immediacy and sheer massiveness of impact on human thought and values, it would be difficult to find revolutions of comparable magnitude in human history. The political, social, and cultural changes that began in France and England at the very end of the 18th century spread almost immediately through Europe and the Americas in the 19th century and then on to Asia, Africa, and Oceania in the 20th. The effects of the two revolutions, the one overwhelmingly democratic in thrust, the other industrial-capitalist, have been to undermine, shake, or topple institutions that had endured for centuries, even millennia, and with them systems of authority, status, belief, and community.

It is easy today to deprecate the suddenness, the cataclysmic nature, the overall revolutionary effect of these two changes and to seek to subordinate results to longer, deeper tendencies of more gradual change in western Europe. But as many recent historians have pointed out, there was to be seen, and seen by a great many sensitive minds of that day, a dramatic and convulsive quality to the changes that cannot properly be subsumed to the slower processes of continuous evolutionary change. What is crucial, in any event, from the point of view of the history of the social thought of the period, is how the changes were actually envisaged at the time. By a large number of social philosophers and social scientists, in all spheres, those changes were regarded as nothing less than of earthquake intensity.

The coining or redefining of words is an excellent indication of men's perceptions of change in a given historical period. A large number of words taken for granted today came into being in the period marked by the final decade or two of the 18th century and the first quarter of the 19th. Among these are: industry, industrialist, democracy, class, middle class, ideology, intellectual, rationalism, humanitarian, atomistic, masses, commercialism, proletariat, collectivism, equalitarian, liberal, conservative, scientist, utilitarian, bureaucracy, capitalism, and crisis. Some of these words were invented; others reflect new and very different meanings given to old ones. All alike bear witness to the transformed character of the European social landscape as this landscape loomed up to the leading minds of the age. And all these words bear witness too to the emergence of new social philosophies and, most pertinent to the subject of this article, the social sciences as they are known today.

Major themes resulting from democratic and industrial change. It is illuminating to mention a few of the major themes in social thought in the 19th century that were almost the direct results of the democratic and industrial revolutions. It should be borne in mind that these themes are to be seen in the philosophical and literary writing of the age as well as in social thought.

First, there was the great increase in population. Between

Effects of population growth

1750 and 1850 the population of Europe went from 140,000,000 to 266,000,000; in the world from 728,000,000 to well over 1,000,000,000. It was an English clergyman-economist, Thomas Malthus, who, in his famous *Essay on Population,* first marked the enormous significance to human welfare of this increase. With the diminution of historic checks on population growth, chiefly those of high mortality rates—a diminution that was, as Malthus realized, one of the rewards of technical progress—there were no easily foreseeable limits to growth of population. And such growth, he stressed, could only upset the traditional balance between population, which Malthus described as growing at geometrical rate, and food supply, which he declared could grow only at arithmetical rate. Not all social scientists in the century took the pessimistic view of the matter that Malthus did but few if any were indifferent to the impact of explosive increase in population on economy, government, and society.

Second, there was the condition of labour. It may be possible to see this condition in the early 19th century as in fact better than the condition of the rural masses at earlier times. But the important point is that to a large number of writers in the 19th century it seemed worse and was defined as worse. The wrenching of large numbers of people from the older and protective contexts of village, guild, parish, and family, and their massing in the new centres of industry, forming slums, living in common squalor and wretchedness, their wages generally behind cost of living, their families growing larger, their standard of living becoming lower, as it seemed—all of this is a frequent theme in the social thought of the century. Economics indeed became known as the "dismal science," because economists, from David Ricardo to Karl Marx, could see little likelihood of the condition of labour improving under capitalism.

Third, there was the transformation of property. Not only was more and more property to be seen as industrial—manifest in the factories, business houses, and workshops of the period—but also the very nature of property was changing. Whereas for most of the history of mankind property had been "hard," visible only in concrete possessions—land and money—now the more intangible kinds of property such as shares of stock, negotiable equities of all kinds, and bonds were assuming ever greater influence in the economy. This led, as was early realized, to the dominance of financial interests, to speculation, and to a general widening of the gulf between the propertied and the masses. The change in the character of property made easier the concentration of property, the accumulation of immense wealth in the hands of a relative few, and, not least, the possibility of economic domination of politics and culture. It should not be thought that only socialists saw property in this light. From Edmund Burke through Auguste Comte, Frédéric Le Play, and John Stuart Mill down to Karl Marx, Max Weber, and Émile Durkheim, one finds conservatives and liberals looking at the impact of this change in analogous ways.

Effects of urbanization and technological change

Fourth, there was urbanization—the sudden increase in the number of towns and cities in western Europe and the increase in number of persons living in the historic towns and cities. Whereas in earlier centuries, the city had been regarded almost uniformly as a setting of civilization, culture, and freedom of mind, now one found more and more writers aware of the other side of cities: the atomization of human relationships, broken families, the sense of the mass, of anonymity, alienation, and disrupted values. Sociology particularly among the social sciences turned its attention to the problems of urbanization. The contrast between the more organic type of community found in rural areas and the more mechanical and individualistic society of the cities is a basic contrast in sociology, one that was given much attention by such pioneers in Europe as the French sociologists Frédéric Le Play and Émile Durkheim; the German sociologists Ferdinand Tönnies, Georg Simmel, and Max Weber; the Belgian statistician Adolphe Quetelet; and, in America, by the sociologists Charles H. Cooley and Robert E. Park.

Fifth, there was technology. With the spread of mechanization, first in the factories, then in agriculture, social thinkers could see possibilities of a rupture of the historic relation between man and nature, between man and man, even between man and God. To thinkers as politically different as Thomas Carlyle and Karl Marx, technology seemed to lead to dehumanization of the worker and to exercise of a new kind of tyranny over human life. Marx, though, far from despising technology, thought the advent of socialism would counteract all this. Alexis de Tocqueville declared that technology, and especially technical specialization of work, was more degrading to man's mind and spirit than even political tyranny. It was thus in the 19th century that the opposition to technology on moral, psychological, and aesthetic grounds first made its appearance in Western thought.

Sixth, there was the factory system. The importance of this to 19th-century thought has been intimated above. Suffice it to add that along with urbanization and spreading mechanization, the system of work whereby masses of workers left home and family to work long hours in the factories became a major theme of social thought as well as of social reform.

Effects of the rise of the masses

Seventh, and finally, mention is to be made of the development of political masses—that is, the slow but inexorable widening of franchise and electorate through which ever larger numbers of persons became aware of themselves as voters and participants in the political process. This too is a major theme in social thought, to be seen most luminously perhaps in Tocqueville's *Democracy in America,* a classic written in the 1830s that took not merely America but democracy everywhere as its subject. Tocqueville saw the rise of the political masses, more especially the immense power that could be wielded by the masses, as the single greatest threat to individual freedom and cultural diversity in the ages ahead.

These, then, are the principal themes in the 19th-century writing that may be seen as direct results of the two great revolutions. As themes, they are to be found not only in the social sciences but, as noted above, in a great deal of the philosophical and literary writing of the century. In their respective ways, the philosophers Hegel, Coleridge, and Emerson were as struck by the consequences of the revolutions as were any social scientists. So too were such novelists as Balzac and Dickens.

New ideologies. One other point must be emphasized about these themes. They became, almost immediately in the 19th century, the bases of new ideologies. How men reacted to the currents of democracy and industrialism stamped them conservative, liberal, or radical. On the whole, with rarest exceptions, liberals welcomed the two revolutions, seeing in their forces opportunity for freedom and welfare never before known to mankind. The liberal view of society was overwhelmingly democratic, capitalist, industrial, and, of course, individualistic. The case is somewhat different with conservatism and radicalism in the century. Conservatives, beginning with Edmund Burke, continuing through Hegel and Matthew Arnold down to such minds as John Ruskin later in the century, disliked both democracy and industrialism, preferring the kind of tradition, authority, and civility that had been, in their minds, displaced by the two revolutions. Theirs was a retrospective view, but it was a nonetheless influential one, affecting a number of the central social scientists of the century, among them Auguste Comte and Tocqueville and later Max Weber and Émile Durkheim. The radicals accepted democracy but only in terms of its extension to all areas of society and its eventual annihilation of any form of authority that did not spring directly from the people as a whole. And although the radicals, for the most part, accepted the phenomenon of industrialism, especially technology, they were uniformly antagonistic to capitalism.

These ideological consequences of the two revolutions proved extremely important to the social sciences, for it would be difficult to identify a social scientist in the century—as it would a philosopher or a humanist—who was not, in some degree at least, caught up in ideological currents. In referring to such minds as Saint-Simon, Comte, Le Play among sociologists, to Ricardo, the Frenchman Jean-Baptiste Say, and Marx among economists, to Jeremy

Bentham and John Austin among political scientists, even to anthropologists like the Englishman Edward B. Tylor and the American Lewis Henry Morgan, one has before one men who were engaged not merely in the study of society but also in often strongly partisan ideology. Some were liberals, some conservatives, others radicals. All drew from the currents of ideology that had been generated by the two great revolutions.

New intellectual and philosophical tendencies. It is important also to identify three other powerful tendencies of thought that influenced all of the social sciences. The first is a positivism that was not merely an appeal to science but almost reverence for science; the second, humanitarianism; the third, the philosophy of evolution.

Effects of Positivism

The Positivist appeal of science was to be seen everywhere. The rise of the ideal of science in the Age of Reason was noted above. The 19th century saw the virtual institutionalization of this ideal—possibly even canonization. The great aim was that of dealing with moral values, institutions, and all social phenomena through the same fundamental methods that could be seen so luminously in such areas as physics and biology. Prior to the 19th century, no very clear distinction had been made between philosophy and science, and the term philosophy was even preferred by those working directly with physical materials, seeking laws and principles in the fashion of a Newton or Harvey—that is, by persons whom one would now call scientists.

In the 19th century, in contrast, the distinction between philosophy and science became an overwhelming one. Virtually every area of man's thought and behaviour was thought by a rising number of persons to be amenable to scientific investigation in precisely the same degree that physical data were. More than anyone else, it was Comte who heralded the idea of the scientific treatment of social behaviour. His *Cours de philosophie positive,* published in six volumes between 1830 and 1842, sought to demonstrate irrefutably not merely the possibility but the inevitability of a science of man, one for which Comte coined the word "sociology" and that would do for man the social being exactly what biology had already done for man the biological animal. But Comte was far from alone. There were many in the century to join in his celebration of science for the study of society.

Humanitarianism, though a very distinguishable current of thought in the century, was closely related to the idea of a science of society. For the ultimate purpose of social science was thought by almost everyone to be the welfare of society, the improvement of its moral and social condition. Humanitarianism, strictly defined, is the institutionalization of compassion; it is the extension of welfare and succour from the limited areas in which these had historically been found, chiefly family and village, to society at large. One of the most notable and also distinctive aspects of the 19th century was the constantly rising number of persons, almost wholly from the middle class, who worked directly for the betterment of society. In the many projects and proposals for relief of the destitute, improvement of slums, amelioration of the plight of the insane, the indigent, and imprisoned, and other afflicted minorities could be seen the spirit of humanitarianism at work. All kinds of associations were formed, including temperance associations, groups and societies for the abolition of slavery and of poverty and for the improvement of literacy, among other objectives. Nothing like the 19th-century spirit of humanitarianism had ever been seen before in western Europe—not even in France during the Enlightenment, where interest in mankind's salvation tended to be more intellectual than humanitarian in the strict sense. Humanitarianism and social science were reciprocally related in their purposes. All that helped the cause of the one could be seen as helpful to the other.

Effects of evolutionary theory

The third of the intellectual influences is that of evolution. It affected every one of the social sciences, each of which was as much concerned with the development of things as with their structures. An interest in development was to be found in the 18th century, as noted earlier. But this interest was small and specialized compared with 19th-century theories of social evolution. The impact of Charles Darwin's *Origin of Species,* published in 1859, was of course great and further enhanced the appeal of the evolutionary view of things. But it is very important to recognize that ideas of social evolution had their own origins and contexts. The evolutionary works of such social scientists as Comte, Herbert Spencer, and Marx had been completed, or well begun, before publication of Darwin's work. The important point, in any event, is that the idea or the philosophy of evolution was in the air throughout the century, as profoundly contributory to the establishment of sociology as a systematic discipline in the 1830s as to such fields as geology, astronomy, and biology. Evolution was as permeative an idea as the Trinity had been in medieval Europe.

Development of the separate disciplines. Among the disciplines that formed the social sciences, two contrary, for a time equally powerful, tendencies at first dominated them. The first was the drive toward unification, toward a single, master social science, whatever it might be called. The second tendency was toward specialization of the individual social sciences. If, clearly, it is the second that has triumphed, with the results to be seen in the disparate, sometimes jealous, highly specialized disciplines seen today, the first was not without great importance and must also be examined.

Unification versus specialization

What emerges from the critical rationalism of the 18th century is not, in the first instance, a conception of need for a plurality of social sciences, but rather for a single science of society that would take its place in the hierarchy of the sciences that included the fields of astronomy, physics, chemistry, and biology. When, in the 1820s, Comte wrote calling for a new science, one with man the social animal as the subject, he assuredly had but a single, encompassing science of society in mind—not a congeries of disciplines, each concerned with some single aspect of man's behaviour in society. The same was true of Bentham, Marx, and Spencer. All these minds, and there were many others to join them, saw the study of society as a unified enterprise. They would have scoffed, and on occasion did, at any notion of a separate economics, political science, sociology, and so on. Society is an indivisible thing, they would have argued; so, too, must be the study of society.

It was, however, the opposite tendency of specialization or differentiation that won out. No matter how the century began, or what were the dreams of a Comte, Spencer, or Marx, when the 19th century ended, not one but several distinct, competitive social sciences were to be found. Aiding this process was the development of the colleges and universities. With hindsight it might be said that the cause of universities in the future would have been strengthened, as would the cause of the social sciences, had there come into existence, successfully, a single curriculum, undifferentiated by field, for the study of society. What in fact happened, however, was the opposite. The growing desire for an elective system, for a substantial number of academic specializations, and for differentiation of academic degrees, contributed strongly to the differentiation of the social sciences. This was first and most strongly to be seen in Germany, where, from about 1815 on, all scholarship and science were based in the universities and where competition for status among the several disciplines was keen. But by the end of the century the same phenomenon of specialization was to be found in the United States (where admiration for the German system was very great in academic circles) and, in somewhat less degree, in France and England. Admittedly, the differentiation of the social sciences in the 19th century was but one aspect of a larger process that was to be seen as vividly in the physical sciences and the humanities. No major field escaped the lure of specialization of investigation, and clearly, a great deal of the sheer bulk of learning that passed from the 19th to the 20th century was the direct consequence of this specialization.

Classical economists and socialists

Economics. It was economics that first attained the status of a single and separate science, in ideal at least, among the social sciences. That autonomy and self-regulation that the Physiocrats and Adam Smith had found, or thought they had found, in the processes of wealth, in the operation of prices, rents, interest, and wages during the 18th century became the basis of a separate and distinctive economics—or, as it was often called, "political economy"—in the 19th. Hence the emphasis upon what came to be widely called laissez-faire. If, as it was argued, the processes of wealth operate naturally in terms of their own built-in mechanisms, then not only should these be studied separately but they should, in any wise polity, be left alone by government and society. This was, in general, the overriding emphasis of such thinkers as David Ricardo, John Stuart Mill, and Nassau William Senior in England, of Frédéric Bastiat and Jean-Baptiste Say in France, and, somewhat later, the Austrian school of Carl Menger. This emphasis is today called "classical" in economics, and it is even now, though with substantial modifications, a strong position in the field.

There were almost from the beginning, however, economists who diverged sharply from this laissez-faire, classical view. In Germany especially there were the so-called historical economists. They proceeded less from the discipline of historiography than from the presuppositions of social evolution, referred to above. Such men as Wilhelm Roscher and Karl Knies in Germany tended to dismiss the assumptions of timelessness and universality regarding economic behaviour that were almost axiomatic among the followers of Adam Smith, and they strongly insisted upon the developmental character of capitalism, evolving in a long series of stages from other types of economy.

Also prominent throughout the century were those who came to be called the Socialists. They too repudiated any notion of timelessness and universality in capitalism and its elements of private property, competition, and profit. Not only was this system but a passing stage of economic developments; it could be—and, as Marx was to emphasize, would be—shortly supplanted by a more humane and also realistic economic system based upon cooperation, the people's ownership of the means of production, and planning that would eradicate the vices of competition and conflict.

Political science. Rivalling economics as a discipline during the century was political science. The line of systematic interest in the state that had begun in modern Europe with Machiavelli, Hobbes, Locke, and Rousseau, among others, widened and lengthened in the 19th century, the consequence of the two revolutions. If the Industrial Revolution seemed to supply all the problems frustrating the existence of a stable and humane society, the political-democratic revolution could be seen as containing many of the answers to these problems. It was the democratic revolution, especially in France, that created the vision of a political government responsible for all aspects of human society and, most important, possessed the power to wield this responsibility. This power, known as sovereignty, could be seen as holding the same relation to political science in the 19th century that capital held to economics. To a very large number of political scientists, the aim of the discipline was essentially that of analyzing the varied properties of sovereignty. There was a strong tendency on the part of such political scientists as Bentham, Austin, and Mill in England and Francis Lieber and Woodrow Wilson in the United States to see the state and its claimed sovereignty over human lives in much the same terms in which classical economists saw capitalism.

Concern with sovereignty

Among political scientists there was the same historical-evolutionary dissent from this view, however, that existed in economics. Such writers as Sir Henry Maine in England, Numa Fustel de Coulanges in France, and Otto von Gierke in Germany declared that state and sovereignty were not timeless and universal nor the results of some "social contract" envisaged by such philosophers as Locke and Rousseau but, rather, structures formed slowly through developmental or historical processes. Hence the strong interest, especially in the late 19th century, in the origins of political institutions in kinship, village, and caste, and in the successive stages of development that have characterized these institutions. In political science, as in economics, in short, the classical analytical approach was strongly rivalled by the evolutionary. Both approaches go back to the 18th century in their fundamental elements, but what is seen in the 19th century is the greater systematization and the much wider range of data employed.

Cultural anthropology. In the 19th century, anthropology also attained clear identity as a discipline. Strictly defined as "the science of man," it could be seen as superseding other specialized disciplines such as economics and political science. In practice and from the beginning, however, anthropology concerned itself overwhelmingly with primitive man. On the one hand was physical anthropology, concerned chiefly with the evolution of man as a biological species, with the successive forms and protoforms of the species, and with genetic systems such as stocks and races in the world. On the other hand was social and cultural anthropology: here the interest was in the full range of man's institutions but confined to those found in fact among existing preliterate or "primitive" peoples in Africa, Oceania, Asia, and the Americas. Above all other concepts, "culture" was the central element of this great area of anthropology, or ethnology, as it was often called to distinguish it from physical anthropology. Culture, as a concept, called attention to the nonbiological, nonracial, noninstinctual basis of the greater part of what one calls civilization: its values, techniques, ideas in all spheres. Culture, as defined in Tylor's landmark work of 1871, *Primitive Culture,* is the part of man's behaviour that is learned. From cultural anthropology more than from any other single social science has come the emphasis on the cultural foundations of man's behaviour and thought in society.

Anthropological focus on primitive man and evolutionism

Scarcely less than political science or economics, cultural anthropology shared in the themes of the two revolutions and their impact on the world. If the data that cultural anthropologists actually worked with were generally in the remote areas of the world, it was the effects of the two revolutions that, in a sense, kept opening up these parts of the world to more and more systematic inquiry. And, as was true of the other social sciences, the cultural anthropologists were immersed in problems of economics, polity, social class, and community, albeit among preliterate rather than "modern" peoples.

Overwhelmingly, without major exception indeed, the science of cultural anthropology was evolutionary in thrust in the 19th century. Edward B. Tylor and Sir John Lubbock in England, Lewis Henry Morgan in the United States, Adolf Bastian and Theodor Waitz in Germany, and all others in the main line of the study of primitive culture saw existing native societies in the world as prototypes of their own "primitive ancestors," fossilized remains, so to speak, of stages of development that western Europe had once gone through. Despite the vast array of data compiled on non-Western cultures, the same basic European-centred objectives are to be found among cultural anthropologists as among other social scientists in the century. Almost universally, then, the modern West was regarded as the latest point in a line of progress that was single and unilinear and on which all other peoples in the world could be fitted as illustrations, as it were, of Western man's own past.

Sociology. Sociology came into being in precisely these terms, and during much of the century it was not easy to distinguish between a great deal of so-called sociology and social or cultural anthropology. Even if almost no sociologists in the century made empirical studies of primitive peoples, as did the anthropologists, their interest in the origin, development, and probable future of mankind was not less great than what could be found in the writings of the anthropologists. It was Auguste Comte who coined the word sociology, and he used it to refer to what he imagined would be a single, all-encompassing, science of society that would take its place at the top of the hierarchy of sciences—a hierarchy that Comte saw as including astronomy (the oldest of the sciences historically) at the bottom and with physics, chemistry, and biology rising in that order to sociology, the latest and grandest of the sciences. There was no thought in Comte's mind—nor was there in the mind of Herbert Spencer, whose general view of sociology was very much like Comte's—of there being other, competing social sciences. Sociology would be to

The "grand" view of sociology

the whole of the social world what each of the other great sciences was to its appropriate sphere of reality.

Both Comte and Spencer believed that civilization as a whole was the proper subject of sociology. Their works were concerned, for the most part, with describing the origins and development of civilization and also of each of its major institutions. Both declared sociology's main divisions to be "statics" and "dynamics," the former concerned with processes of order in society, the latter with processes of evolutionary change in society. Both men also saw all existing societies in the world as reflective of the successive stages through which Western society had advanced in time over a period of tens of thousands of years.

The "problems" view of sociology

Not all sociologists in the 19th century conceived their discipline in this light, however. Side by side with the "grand" view represented by Comte and Spencer were those in the century who were primarily interested in the social problems that they saw around them—consequences, as they interpreted them, of the two revolutions, the industrial and democratic. Thus in France just after midcentury, Frédéric Le Play published a monumental study of the social aspects of the working classes in Europe, *Les Ouvriers européens,* which compared families and communities in all parts of Europe and even other parts of the world. Alexis de Tocqueville, especially in the second volume of his *Democracy in America* (1835), provided an account of the customs, social structures, and institutions in America, dealing with these—and also with the social and psychological problems of Americans in that day—as aspects of the impact of the democratic and industrial revolutions upon traditional society.

At the very end of the 19th century, in both France and Germany, there appeared some of the works in sociology that were to prove most lasting in their effects upon 20th-century sociology. Ferdinand Tönnies, in his *Gemeinschaft und Gesellschaft* (1887; translated as *Community and Society*), sought to explain all major social problems in the West as the consequence of the West's historical transition from the communal, status-based, concentric society of the Middle Ages to the more individualistic, impersonal, and large-scale society of the democratic-industrial period. In general terms, allowing for individual variations of theme, these were the views of Max Weber, Georg Simmel, and Émile Durkheim (all of whom also wrote in the late 19th and early 20th century). These were the men who, starting from the problems of Western society that could be traced to the effects of the two revolutions, did the most to establish the discipline of sociology as it is found for the most part in the 20th century.

Social psychology. Social psychology as a distinct discipline also originated in the 19th century, although its outlines were perhaps somewhat less clear than was true of the other social sciences. The close relation of the human mind to the social order, its dependence upon education and other forms of socialization, was well known in the 18th century. In the 19th century, however, an ever more systematic discipline came into being to uncover the social and cultural roots of human psychology and also the several types of "collective mind" that analysis of different cultures and societies in the world might reveal. In Germany, Moritz Lazarus and Wilhelm Wundt sought to fuse the study of psychological phenomena with analyses of whole cultures. Folk psychology, as it was called, did not, however, last very long in scientific esteem.

Early group studies

Much more esteemed, and closer to 20th-century conceptions of social psychology, were the works of such men as Gabriel Tarde, Gustave Le Bon, Lucien Lévy-Bruhl, and Émile Durkheim in France and Georg Simmel in Germany (all of whom also wrote in the early 20th century). Here, in concrete, often highly empirical studies of small groups, associations, crowds, and other aggregates (rather than in the main line of psychology during the century, which tended to be sheer philosophy at one extreme and a variant of physiology at the other) are to be found the real beginnings of social psychology. Although the point of departure in each of the studies was the nature of association, they dealt, in one degree or other, with the internal processes of psychosocial interaction, the operation of attitudes and judgments, and the social basis of personality and thought—in short, with those phenomena that would, in the 20th century, be the substance of social psychology as a formal discipline.

Social statistics and social geography. Two final manifestations of the social sciences in the 19th century are social statistics and social (or human) geography. At that time, neither achieved the notability and acceptance in colleges and universities that such fields as political science and economics did. Both, however, were as clearly visible by the latter part of the century as any of the other social sciences. And both were to exert a great deal of influence on the other social sciences by the beginning of the 20th century: social statistics on sociology and social psychology pre-eminently; social geography on political science, economics, history, and certain areas of anthropology, especially those areas dealing with the dispersion of races and the diffusion of cultural elements. In social statistics the key figure of the century was a Belgian, Adolphe Quetelet, who was the first, on any systematic basis, to call attention to the kinds of structured behaviour that could be observed and identified only through statistical means. It was Quetelet who brought into prominence the momentous concept of "the average man" and his behaviour. The two major figures in social or human geography in the century were Friedrich Ratzel in Germany and Paul Vidal de la Blache in France. Both broke completely with the crude environmentalism of earlier centuries, which had sought to show how topography and climate actually determine human behaviour, and they substituted the more subtle and sophisticated insights into the relationships of land, sea, and climate on the one hand and, on the other, the varied types of culture and human association that are to be found on earth.

In summary, by the end of the 19th century all the major social sciences had achieved a distinctiveness, an importance widely recognized, and were, especially in the cases of economics and political science, fully accepted as disciplines in the universities. Most important, they were generally accepted as sciences in their own right rather than as minions of philosophy.

THE 20TH CENTURY

What is seen in the 20th century is not only an intensification and spread of earlier tendencies in the social sciences but also the development of many new tendencies that, in the aggregate, make the 19th century seem by comparison one of quiet unity and simplicity in the social sciences.

In the 20th century, the processes first generated by the democratic and industrial revolutions have gone on virtually unchecked in Western society, penetrating more and more spheres of once traditional morality and culture, leaving their impress on more and more nations, regions, and localities. Equally important, perhaps in the long run far more so, is the spread of these revolutionary processes to the non-Western areas of the world. The impact of industrialism, technology, secularism, and individualism upon peoples long accustomed to the ancient unities of tribe, local community, agriculture, and religion was first to be seen in the context of colonialism, an outgrowth of nationalism and capitalism in the West. The relations of the West to non-Western parts of the world, the whole phenomenon of the "new nations," are vital aspects of the social sciences.

So too are certain other consequences, or lineal episodes, of the two revolutions. The 20th century is the century of nationalism, mass democracy, and large-scale industrialism beyond reach of any 19th-century imagination so far as magnitude is concerned. It is the century of mass warfare, of two world wars with toll in lives and property greater perhaps than the sum total of all preceding wars in history. It is the century too of totalitarianism: Communist, Fascist, and Nazi; and of techniques of terrorism that, if not novel, are to be seen on a scale and with an intensity of scientific application that could scarcely have been predicted by those who considered science and technology as unqualifiedly humane in possibility. It is a century of affluence in the West, without precedent for the masses of people, to be seen in a constantly rising standard of living and a constantly rising level of expectations.

The last is important. A great deal of the turbulence in the 20th century—political, economic, and social—is the result of desires and aspirations that have been constantly escalating and that have been passing from the white people in the West to ethnic and racial minorities among them and, then, to whole continents elsewhere. Of all manifestations of revolution, the revolution of rising expectations is perhaps the most powerful in its consequences. For, once this revolution gets under way, each fresh victory in the struggle for rights, freedom, and security tends to magnify the importance of what has not been won.

Once it was thought that by solving the fundamental problems of production and large-scale organization, man could ameliorate other problems, those of a social, moral, and psychological nature. What in fact occurred, on the testimony of a great deal of the most notable thought and writing, was a heightening of such problems. It would appear that as man satisfies, relatively at least, the lower order needs of food and shelter, his higher order needs for purpose and meaning in life become ever more imperious. Thus such philosophers of history as Arnold Toynbee, Pitirim Sorokin, and Oswald Spengler have dealt with problems of purpose and meaning in history with a degree of learning and intensity of spirit not seen since perhaps St. Augustine wrote his monumental *The City of God* in the early 5th century when signs of the disintegration of Roman civilization were becoming overwhelming in their message to so many of that day. In the 20th century, though the idea of progress has certainly not disappeared, it has been rivalled by ideas of cyclical change and of degeneration of society. It is hard to miss the currency of ideas in modern times—status, community, purpose, moral integration, on the one hand, and alienation, anomie, disintegration, breakdown on the other—that reveal only too clearly the divided nature of man's spirit, the unease of his mind.

Effects of alienation and social isolation

There is to be seen too, especially during later decades of the century, a questioning of the role of reason in human affairs—a questioning that stands in stark contrast with the ascendancy of rationalism in the two or three centuries preceding. Doctrines and philosophies stressing the inadequacy of reason, the subjective character of human commitment, and the primacy of faith have rivalled—some would say conquered—doctrines and philosophies descended from the Age of Reason. Existentialism, with its emphasis on the basic loneliness of the individual, on the impossibility of finding truth through intellectual decision, and on the irredeemably personal, subjective character of man's life, has proved to be a very influential philosophy in the writings of the 20th century. Freedom, far from being the essence of hope and joy, is the source of man's dread of the universe and of his anxiety for himself. Søren Kierkegaard's 19th-century intimations of anguished isolation as the perennial lot of the individual have had rich expression in the philosophy and literature of the 20th century.

It might be thought that such intimations and presentiments as these have little to do with the social sciences. This is true in the direct sense perhaps but not true when one examines the matter in terms of contexts and ambiences. The "lost individual" has been of as much concern to the social sciences as to philosophy and literature. Ideas of alienation, anomie, identity crisis, and estrangement from norms are rife among the social sciences, particularly, of course, those most directly concerned with the nature of the social bond, such as sociology, social psychology, and political science. In countless ways, interest in the loss of community, in the search for community, and in the individual's relation to society and morality have had expression in the work of the social sciences. Between the larger interests of a culture and the social sciences there is never a wide gulf—only different ways of defining and approaching these interests.

Marxist influences. The influence of Marxism in the 20th century must not be missed. Currently the works of Lenin have outstripped the Bible in distribution in the world. For hundreds of millions of persons today the ideas of Marx, as communicated by Lenin, have profound moral, even religious significance. But even in those parts of the world, the West foremost, where Communism has exerted little direct political impact, Marxism remains a potent source of ideas. Not a few of the central concepts of social stratification and the location and diffusion of power in the social sciences come straight from Marx's insights. Far more is this the case in the Communist countries—the Soviet Union, other eastern European countries, China, and even Asian countries in which no Communist domination exists. In all these countries, Marx's name is virtually sacrosanct. There is not the same degree of differentiation of social sciences in these countries that is found in the West. As an example, sociology hardly exists as a recognized discipline in these countries, and by the standards of the West, the other social sciences have little more than a rather rudimentary existence. Economics alone tends to be favoured, and this is, of course, largely Marxian economics—the economics of Marx's *Das Kapital.*

But even though Marxism has had relatively little direct impact on the social sciences as disciplines in the West, it has had enormous influence on states of mind that are closely associated with the social sciences. Especially was this true during the 1930s, the decade of the Great Depression. Today signs are not lacking of a strong revival of interest in Marx that could very well, through sheer numbers of its adherents, affect the nature of the social sciences in the years ahead. Socialism remains for a great many persons an evocative symbol and creed. Marx remains a formidable name among intellectuals and is still, without any question, the principal intellectual source of radical movements in politics. Such a position cannot help but influence the contexts of even the most abstract of the social sciences.

Triumph of social planning and Keynesianism

What Marx's ideas have suggested above all else in a positive way is the possibility of a society directed, not by blind forces of competition and struggle among economic elements, but instead by directed planning. This hope, this image, has proved a dominant one in the 20th century even where the influence of Marx and of Socialism has been at best small and indirect. It is this profound interest in central planning and governance that has given almost historic significance to the ideas of the English economist J.M. Keynes. What is called Keynesianism has as its intellectual base a very complex modification of the classical doctrines of economics—one set forth in Keynes's famous *The General Theory of Employment, Interest and Money,* published in 1935–36. Of greater influence today, however, than the strictly theoretical content of this general theory is the political impact that Keynesian ideas have had on Western democracies. For out of these ideas came the clear policy of governments dealing directly with the business cycle, of pumping money and credit into an economic system when the cycle threatens to turn downward, and of then lessening this infusion when the cycle moves upward. Above all other names in the West, that of Keynes has become identified with such policy in the democracies and with the general movement of central governments toward ever more active and constant regulation of processes once thought best left to what the classical economists thought of as natural laws. True, the root ideas of the classical economists are found in modified form even today in the works of such economists as the American Milton Friedman. But it would not be unfair to say that Keynes's name has become associated with democratic economic planning and direction in much the way that Marx's name is associated with Communist economic policies.

Freudian influences. In the general area of personality, mind, and character, the writings of Sigmund Freud have had influence on 20th-century culture and thought scarcely less than Marx's. His basic theories of the role of the unconscious mind, of the lasting effects of infantile sexuality, and of the Oedipus complex have gone beyond the discipline of psychoanalysis and even the larger area of psychiatry to areas of several of the social sciences. Anthropologists have applied Freudian concepts to their studies of primitive cultures, seeking to assess comparatively the universality of states of the unconscious that Freud and his followers held to lie in the whole human race. Some political scientists have used Freudian ideas to

Freudian emphases on the unconscious and the unrational illuminate the nature of authority generally, and political power specifically, seeing in totalitarianism, for example, the thrust of a craving for the security that total power can give. Sociology and social psychology have been influenced by Freudian ideas in their studies of social interaction and motivation. From Freud came the fruitful perspective that sees social behaviour and attitudes as generated not merely by the external situation but also by internal emotional needs springing from childhood—needs for recognition, authority, self-expression. Whatever may be the place directly occupied by Freud's ideas in the social sciences today, his influence upon 20th-century thought and culture generally, not excluding the social sciences, has been hardly less than Marx's.

Diffusion of the social sciences

Specialization and cross-disciplinary approaches. A major point to make about the social sciences of the 20th century is the vast increase in the number of social scientists involved, in the number of academic and other centres of teaching and research in the social sciences, and in their degree of both comprehensiveness and specialization. The explosion of the sciences generally in the 20th century—an explosion responsible for the fact that a majority of all scientists who have ever lived in human history are now alive—has had, as one of its signal elements, the explosion of the social sciences. Not only has there been development and proliferation but there has also been a spectacular diffusion of the social sciences. Beginning in a few places in western Europe and the United States in the 19th century, the social sciences, as bodies of ongoing research and centres of teaching, are today to be found almost everywhere in the world. In considerable part this has followed the spread of universities from the West to other parts of the world and, within universities, the very definite shift away from the hegemony once held by humanities alone to the near-hegemony held today by the sciences, physical and social.

Specialization has been as notable a tendency in the social sciences as in the biological and physical sciences. This is reflected not only in varieties of research but also in course offerings in academic departments. Whereas not very many years ago, a couple of dozen advanced courses in a social science reflected the specialization and diversity of the discipline even in major universities with graduate schools, today a hundred such courses are found to be not enough.

Side by side with this strong trend toward specialization, however, is another, countering trend: that of cross-fertilization and interdisciplinary cooperation. At the beginning of the century, down in fact until World War II, the several disciplines existed each in a kind of splendid isolation from the others. That historians and sociologists, for example, might ever work together in curricula and research projects would have been scarcely conceivable prior to about 1945. Each social science tended to follow the course that emerged in the 19th century: to be confined to a single, distinguishable, if artificial, area of social reality. Today, evidences are all around of cross-disciplinary work and of fusion within a single social science of elements drawn from other social sciences. Thus there are such vital areas of work as political sociology, economic anthropology, psychology of voting, and industrial sociology. Single concepts such as "structure," "function," "alienation," and "motivation" can be seen employed variously to useful effect in several social sciences. The techniques of one social science can be seen consciously incorporated into another or into several social sciences. If history has provided much in the way of perspective to sociology or anthropology, each of these two has provided perspective, and also whole techniques, such as statistics and survey, to history. In short, specialization is by no means without some degree at least of countertendencies such as fusion and synthesis.

Professionalization

Another outstanding characteristic of each of the social sciences in the 20th century is its professionalization. Without exception, the social sciences have become bodies of not merely research and teaching but also practice, in the sense that this word has in medicine or engineering. Down until about World War II, it was a rare sociologist or political scientist or anthropologist who was not a holder of academic position. There were economists and psychologists to be found in banks, industries, government, even in private consultantship, but the numbers were relatively tiny. Overwhelmingly the social sciences had visibility alone as academic disciplines, concerned essentially with teaching and with more or less basic, individual research. All this has changed profoundly, and on a vast scale, during the past three decades. Today there are as many economists and psychologists outside academic departments as within, if not more. The number of sociologists, political scientists, and demographers to be found in government, industry, and private practice rises constantly. Equally important is the changed conception or image of the social sciences. Today, to a degree unknown before World War II, the social sciences are conceived as policy-making disciplines, concerned with matters of national welfare in their professional capacities in just as sure a sense as any of the physical sciences. Inevitably, tensions have arisen within the social sciences as the result of processes of professionalization. Those persons who are primarily academic can all too easily feel that those who are primarily professional have different and competing identifications of themselves and their disciplines.

Nature of the research. The emphasis upon research in the social sciences has become almost transcending within recent decades. This situation is not at all different from that which prevails in the physical sciences and the professions in this age. Prior to about 1945, the functions of teaching and research had approximately equal value in many universities and colleges. The idea of a social (or physical) scientist appointed to an academic institution for research alone, or with research preponderant, was scarcely known. Research bureaus and institutes in the social sciences were very few and did not rival traditional academic departments and colleges as prestige-bearing entities. All of that was changed decisively beginning with the period just after World War II. From governments and foundations, large sums of money passed into the universities—usually not to the universities as such, but rather to individuals or small groups of individuals, each eminent for research. Research became the uppermost value in the social sciences (as in the physical) and hence, of course, in the universities themselves.

Use of mathematics and computers

Probably the greatest single change in the social sciences during the past generation has been the widespread introduction of mathematical and other quantitative methods. Without question, economics is the discipline in which the most spectacular changes of this kind have taken place. So great is the dominance of mathematical techniques here—resulting in the eruption of what is called econometrics to a commanding position in the discipline—that, to the outsider, economics today almost appears to be a branch of mathematics. But in sociology, political science, social psychology, and anthropology, the impact of quantitative methods, above all, of statistics, has also been notable. No longer does statistics stand alone, a separate discipline, as it did in effect during the 19th century. This area today is inseparable from each of the social sciences, though, in the field of mathematics, statistics still remains eminently distinguishable, the focus of highly specialized research and theory.

Within the past decade or two, the use of computers and of all the complex techniques associated with computers has become a staple of social-science research and teaching. Through the data storage and data retrieval of electronic computers, working with amounts and diversity of data that would call for the combined efforts of hundreds, even thousands of technicians, the social sciences have been able to deal with both the extensive and intensive aspects of human behaviour in ways that would once have been inconceivable. The so-called computer revolution in modern thought has been, in short, as vivid a phase of the social as the physical sciences, not to mention other areas of modern life. The problem as it is stated by mature social scientists is to use computers in ways in which they are best fitted but without falling into the fallacy that they can alone guide, direct, and supply vital perspective in the study of man.

Closely related to mathematical, computer, and other

Emphasis on empirical evidence and data gathering

quantitative aspects of the social sciences is the vast increase in the empiricism of modern social science. Never in history has so much in the way of data been collected, examined, classified, and brought to the uses of social theory and social policy alike. What has been called the triumph of the fact is nowhere more visible than in the social sciences. Without question, this massive empiricism has been valuable, indispensable indeed, to those seeking explanations of social structures and processes. Empiricism, however, like quantitative method, is not enough in itself. Unless related to hypothesis, theory, or conclusion, it is sterile, and most of the leading social scientists of today reflect this view in their works. Too many, however, deal with the gathering and classifying of data as though these were themselves sufficient.

It is the quest for data, for detailed, factual knowledge of human beliefs, opinions, and attitudes, as well as patterns and styles of life—familial, occupational, political, religious, and so on—that has made the use of surveys and polls another of the major tendencies in the social sciences of this century. The poll data one sees in his newspaper are hardly more than the exposed portion of an iceberg. Literally thousands of polls, questionnaires, and surveys are going on at any given moment today in the social sciences. The survey or polling method ranks with the quantitative indeed in popularity in the social sciences, both being, obviously, indispensable tools of the empiricism just mentioned.

Specific versus grand theory

Theoretical modes. It is not the case, however, that interest in theory is a casualty of the 20th-century fascination with method and fact. Though there is a great deal less of that grand or comprehensive theory that was a hallmark of 19th-century social philosophy and social science, there are still those persons occasionally to be found today who are engrossed in search for master principles, for general and unified theory that will assimilate all the lesser and more specialized types of theory. But their efforts and results are not regarded as successful by the vast majority of social scientists. Theory, at its best, today tends to be specific theory—related to one or other of the major divisions of research within each of the social sciences. The theory of the firm in economics, of deviance in sociology, of communication in political science, of attitude formation in social psychology, of divergent development in cultural anthropology are all examples of theory in every proper sense of the word. But each is, clearly, specific. If there is a single social science in which a more or less unified theory exists, with reference to the whole of the discipline, it is economics. Even here, however, unified, general theory does not have the sovereign sweep it had in the classical tradition of Ricardo and his followers before the true complexities of economic behaviour had become revealed.

Developmentalism. Developmentalism is another overall influence upon the work of the social sciences, especially within the past three decades. As noted above, an interest in social evolution was one of the major aspects of the social sciences throughout the 19th century in western Europe. In the early 20th century, however, this interest, in its larger and more visible manifestations, seemed to terminate. There was a widespread reaction against the idea of unilinear sequences of stages, deemed by the 19th-century social evolutionists to be universal for all mankind in all places. Criticism of social evolution in this broad sense was a marked element of all the social sciences, preeminently in anthropology but in the others as well. There were numerous demonstrations of the inadequacy of unilinear descriptions of change when it came to accounting for what actually happened, so far as records and other evidences suggested, in the different areas and cultures of the world.

Influences of national growth and development

Beginning in the late 1940s and the 1950s, however, there was a resurgence of developmental ideas in all the social sciences—particularly with respect to studies of the new nations and cultures that were coming into existence in considerable numbers. Studies of economic growth and of political and social development have become more and more numerous. Although it would be erroneous to see these developmental studies as simple repetitions of those of the 19th-century social evolutionists, there are, nevertheless, common elements of thought, including the idea of stages of growth and of change conceived as continuous and cumulative and even as moving toward some more or less common end. At their best, these studies of growth and development in the new nations, by their counterposing of traditional and modern ways, tell a good deal about specific mechanisms of change, the result of the impact of the West upon outlying parts of the world. But as more and more social scientists have recently become aware, efforts to place these concrete mechanisms of change into larger, more systematic models of development all too commonly succumb to the same faults of unilinearity and specious universalism that early-20th-century critics found in 19th-century social evolution.

Social-systems approach. Still another major tendency in all of the social sciences since World War II has been the interest in "social systems." The behaviour of individuals and groups is seen as falling into multiple interdependencies, and these interdependencies are considered sufficiently unified to warrant use of the word "system." Although there are clear uses of biological models and concepts in social-systems work, it may be fair to say that the greatest single impetus to development of this area was widening interest after World War II in cybernetics—the study of human control functions and of the electrical and mechanical systems that could be devised to replace or reinforce them. Concepts drawn from mechanical and electrical engineering have been rather widespread in the study of social systems.

In social-systems studies, the actions and reactions of individuals, or even of groups as large as nations, are seen as falling within certain definable, more or less universal patterns of equilibrium and disequilibrium. The interdependence of roles, norms, and functions is regarded as fundamental in all types of group behaviour, large and small. Each social system, as encountered in social-science studies, is a kind of "ideal type," not identical to any specific "real" condition but sufficiently universal in terms of its central elements to permit useful generalization.

Emphasis on pattern and interdependence

Structuralism and functionalism. Structuralism in the social sciences is closely related to the theory of the social system. Although there is nothing new about the root concepts of structuralism—they may be seen in one form or other throughout Western thought—there is no question but that in the present century this view of behaviour has become a dominant one in many fields. At bottom it is a reaction against all tendencies to deal with human thought and behaviour atomistically—that is, in terms of simple, discrete units of either thought, perception, or overt behaviour. In psychology, structuralism in its oldest sense simply declares that perception occurs, with learning following, in terms of experiences or sensations in various combinations, in discernible patterns or gestalten. In sociology, political science, and anthropology, the idea of structure similarly refers to the repetitive patternings that are found in the study of social, economic, political, and cultural existence. The structuralist contends that no element can be examined or explained outside its context or the pattern or structure of which it is a part. Indeed, it is the patterns, not the elements, that are the only valid objects of study.

What is called functionalism in the social sciences today is closely related to structuralism, with the term structural-functional a common one, especially in sociology and anthropology. Function refers to the way in which behaviour takes on significance, not as a discrete act but as the dynamic aspect of some structure. Biological analogies are common in theories of structure and function in the social sciences. Very common is the image of the biological organ, with its close interdependence to other organs (as the heart to the lung) and the interdependence of activities (as circulation to respiration).

Interactionism. Interaction is still another concept that has had wide currency in the social sciences of the 20th century. Social interaction—or, as it is sometimes called, symbolic interaction—refers to the fact that the relationships among two or more groups or human beings are never one-sided, purely physical, or direct. Always there is reciprocal influence, a mutual sense of "otherness." And always the presence of the "other" has crucial effect in one's definition of not merely what is external but what is internal. One acquires one's individual sense of identity from interactions with others beginning in infancy. It is the initial sense of the other person—mother, for example—that in time gives the child its sense of self, a sense that requires continuous development through later interactions with others. From the point of view of interactionist theory, all one's perceptions of and reactions to the external world are mediated or influenced by prior ideas, valuations, and assessments. Always one is engaged in socialization or the modification of one's mind, role, and behaviour through contact with others.

FUTURE OF THE SOCIAL SCIENCES

What has been covered in the preceding paragraphs may be the most that can be said within restricted compass about the social sciences of the 20th century without turning to the individual social sciences themselves and related disciplines (see below). The concern here has been with only those major contextual influences, tendencies of overall character, and dominant ideas or theories that the social sciences taken as a whole manifest in one degree or other.

Relation of social sciences to society

There is one final aspect of the subject that must be considered briefly, for how it is resolved will have much effect upon the future of the social sciences in the West. This is the relation of the social sciences to organized society, to government and industry, and other institutional centres of authority. At the present time, there is a significant and undoubtedly growing feeling among social scientists, especially younger ones, that the relationship has become altogether too close. The social sciences, it is said, must maintain their distance, their freedom, from bureaucratized government and industry. Otherwise they will lose their inherent powers of honest and dispassionate criticism of the ineffective or evil in society. Although there may be a certain amount of feeling ranging from the naïve to the politically revolutionary in such sentiments, they cannot be taken lightly, as is apparent from the serious consideration that is being given on a steadily rising scale to the whole problem of the relationship between social science and social policy.

Since the inception of the social sciences—since, indeed, the time when the universities in the West came into being for the express purpose of training professional men in law, theology, and medicine—man has properly sought, through knowledge, to influence social policy, taking this latter term in the widest sense to include not merely the policies of national government but of local government, business, professions, and so on. What else, it may be asked, are the social sciences all about if it is not to use knowledge to improve social life; and how else but through influencing of the major institutions can such improvement take place?

So much is true, comes the answering response. But in the process of seeking to influence the great agencies of modern power and function—of what is loosely called the Establishment—the social sciences may themselves become influenced adversely by the values of power and affluence to be found in these great agencies. They themselves may become identified with the status quo. What the social sciences should give, say the partisans of this view, is a continuation of the revolutionary or at least profoundly reformist tradition that was begun in the 18th century by the philosophers of reason who, detesting the official establishment of their day, sought on their own to transform it. What is today called objectivity or methodological rigour turns out to be, say these same partisans, acceptance of the basic values of reigning government and industry.

It is this essential conflict regarding the purposes of the social sciences, the relation of the social sciences to government and society, and the role of the individual social scientist in the society of the 20th century that bids fair at this moment to be the major conflict of the years ahead. How it is resolved may very well determine the fate of the social sciences, now less than two centuries old.

(R.A.N./Ed.)

Cultural anthropology

Etymologically, anthropology is the science of man. In fact, however, it is only one of the sciences of man, bringing together those disciplines the common aims of which are to describe man and explain him on the basis of the biological and cultural characteristics of the populations among which he is distributed and to emphasize, through time, the differences and variations of these populations. The concept of race, on the one hand, and that of culture, on the other, have received special attention; and although their meaning is still subject to debate, these terms are doubtless the most common of those in the anthropologist's vocabulary.

Early emphasis on "primitive" man

Anthropology, which is concerned with the study of human differences, was born after the Age of Discovery had opened up societies that had remained outside the technological civilization of the modern West. In fact, the field of research was at first restricted to those societies that had been given one unsatisfactory label after another: "savage," "primitive," "tribal," "traditional," or even "preliterate," "prehistorical," and so on. What such societies had in common, above all, was being the most "different" or the most foreign to the anthropologist; and in the early phases of anthropology, the anthropologists were always European or North American. The distance between the researcher and the object of his study has been a characteristic of anthropological research; it has been said of the anthropologist that he was the "astronomer of the sciences of man."

Anthropologists today study more than just primitive societies. Their research extends not only to village communities within modern societies but also to cities, even to industrial enterprises. Nevertheless, anthropology's first field of research, and the one that perhaps remains the most important, shaped its specific point of view with regard to the other sciences of man and defined its theme. If, in particular, it is concerned with generalizing about patterns of human behaviour seen in all their dimensions and with achieving a total description of social and cultural phenomena, this is because anthropology has observed small-scale societies, which are simpler or at least more homogeneous than modern societies and which change at a slower pace. Thus they are easier to see whole.

What has just been said refers especially to the branch of anthropology concerned with the cultural characteristics of man. Anthropology has, in fact, gradually divided itself into two major spheres: the study of man's biological characteristics and the study of his cultural characteristics. The reasons for this split are manifold, one being the rejection of the initial mistakes regarding correlations between race and culture. More generally speaking, the vast field of 19th-century anthropology was subdivided into a series of increasingly specialized disciplines, using their own methods and techniques, that were given different labels according to national traditions. The Table shows the terminology current in North America and in continental Europe.

Cultural and physical anthropology

Thus two large disciplines—physical anthropology and cultural anthropology—and such related disciplines as prehistory and linguistics now cover the program that originally was set up for a single study of anthropology. The two fields are largely autonomous, having their own relations with disciplines outside anthropology; and it is unlikely that any researchers today work simultaneously in the fields of physical and cultural anthropology. The generalist has become rare. On the other hand, the fields have not been cut off from one another. Specialists in the two fields still cooperate in specific genetic or demographic problems and other matters.

Prehistoric archaeology and linguistics also have notable links with cultural anthropology. In posing the problem of the evolution of mankind in an inductive way, archaeology contributed to the creation of the first concepts of anthropology, and archaeology is still indispensable in uncovering the past of societies under observation. In many areas, when it is a question of interpreting the use of rudimentary tools or of certain elementary religious phenomena, prehistory and cultural anthropology are mutually helpful. "Primitive" societies that have not yet reached the metal age are still in existence.

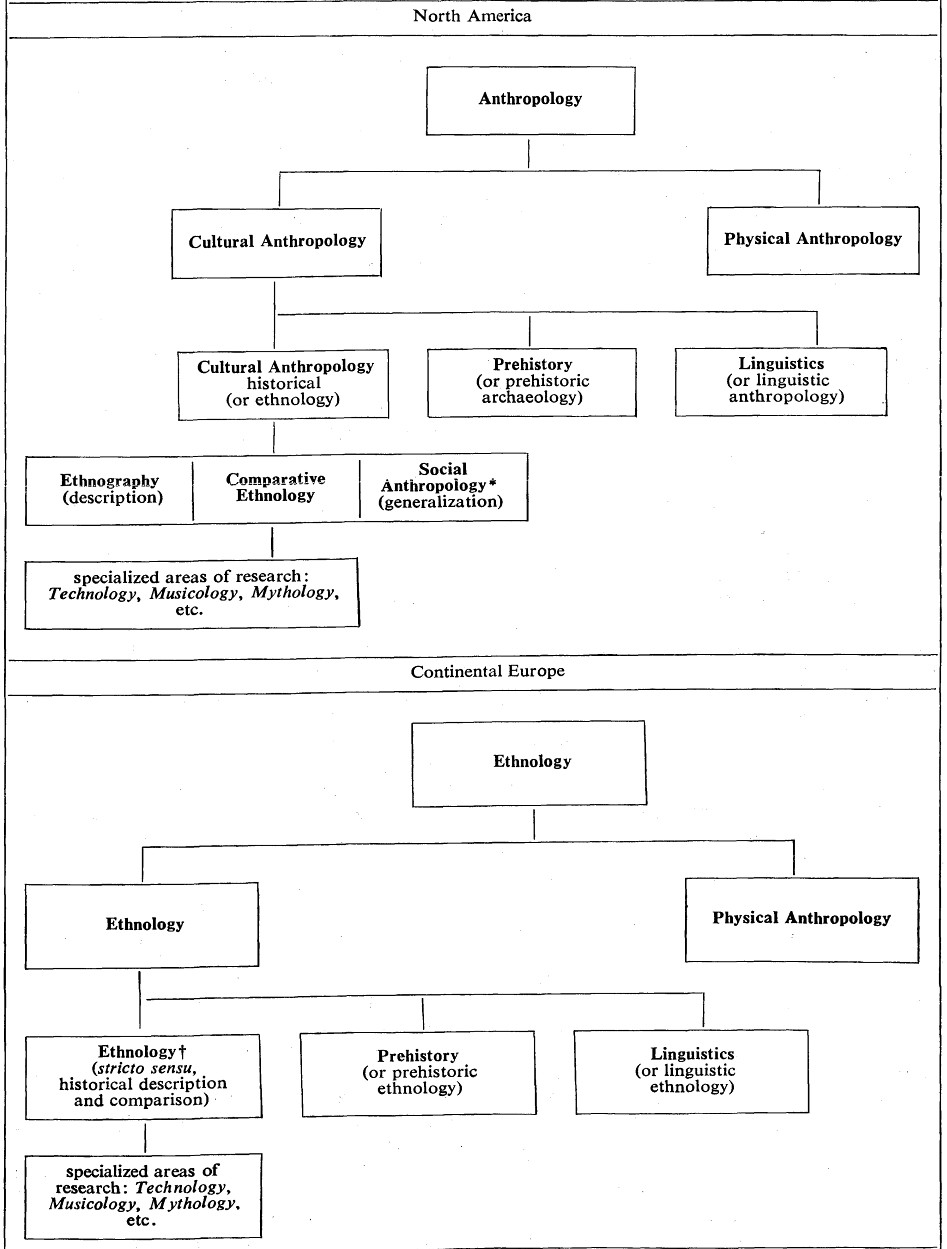
Relations between linguistics and cultural anthropology are numerous. On a purely practical level the cultural anthropologist has to serve a linguistic apprenticeship. He cannot do without a knowledge of the language of the people he is studying, and often he has had to make the first survey of it. One of his essential tasks, moreover, has been to collect the various forms of oral expression, including myths, folk tales, proverbs, and so forth. On the theoretical level, cultural anthropology has often used concepts developed in the field of linguistics: in studying society as a system of communication, in defining the notion of structure, and in analyzing the way in which man organizes and classifies his whole experience of the world.

Cultural anthropology maintains relations with a great number of other sciences. It has been said of sociology, for instance, that it was almost the twin sister of anthropology. The two are presumably differentiated by their field of study (modern societies versus traditional societies). But the contrast is forced. These two social sciences often meet. Thus, the study of colonial societies borrows as much from sociology as from cultural anthropology. And it has already been remarked how cultural anthropology intervenes more and more frequently in urban and industrial fields classically the domain of sociology.

Exchanges with other disciplines

There have also been fruitful exchanges with other disciplines quite distinct from cultural anthropology. In political science the discussion of the concept of the state and of its origin has been nourished by cultural anthropology. Economists, too, have depended on cultural anthropology to see concepts in a more comparative light and even to challenge the very notion of an "economic man" (suspiciously similar to the 19th-century capitalist revered by the classical economists). Cultural anthropology has brought to psychology new bases on which to reflect on concepts

Terminology of Anthropological Disciplines in North America and Continental Europe


North America
Anthropology
Cultural Anthropology
Physical Anthropology
Cultural Anthropology historical (or ethnology)
Prehistory (or prehistoric archaeology)
Linguistics (or linguistic anthropology)
Ethnography (description)
Comparative Ethnology
Social Anthropology* (generalization)
specialized areas of research: Technology, Musicology, Mythology, etc.
Continental Europe
Ethnology
Ethnology
Physical Anthropology
Ethnology† (stricto sensu, historical description and comparison)
Prehistory (or prehistoric ethnology)
Linguistics (or linguistic ethnology)
specialized areas of research: Technology, Musicology, Mythology, etc.

*The term social anthropology is used in Great Britain, either instead of ethnology or instead of cultural anthropology. †There is a tendency at the moment—in France, for example—to use the three terms describing the levels of research which are gradually establishing themselves in North American usage—*i.e.*, ethnography, etc. In eastern and central Europe the term ethnography is often used in the inclusive sense of ethnology.

of personality and the formation of personality. It has permitted psychology to develop a system of cross-cultural psychiatry, or so-called ethnopsychiatry. Conversely, the psychological sciences, particularly psychoanalysis, have offered cultural anthropology new hypotheses for an interpretation of the concept of culture.

The link with history has long been a vital one because cultural anthropology was originally based on an evolutionist point of view and because it has striven to reconstruct the cultural history of societies about which, for lack of written documents, no historical record could be determined. Cultural anthropology has more recently suggested to historians new techniques of research based on the analysis and criticism of oral tradition. And so "ethnohistory" is beginning to emerge. Finally, cultural anthropology has close links with human geography. Both of them place great importance on man either as he uses space or acts to transform the natural environment. It is not without significance that some early anthropologists were originally geographers.

HISTORICAL DEVELOPMENT OF CULTURAL ANTHROPOLOGY

All human societies have been curious about how their customs originated and what the differences between their own culture and that of neighbouring societies might mean. Thus, in a sense they have all constructed their own anthropologies. But the interpretations put forward, even when they were founded partly on accurate observation, most often remained on the level of myth. Embryonic scientific thought began to appear in only a limited number of centres of civilization: in the classical Mediterranean world, in China, in the medieval Arab world, and in the modern Western world. Only in the West, however, did various ideas converge to bring about the birth of scientific anthropology in the 19th century.

A characteristic common to all these centres of civilization was the control that they exercised over vast areas and the opportunity that they enjoyed—through their soldiers, merchants, pilgrims, and missionaries—to gather observations on a wide variety of populations. Such a gathering of data was necessary in order even to begin to understand how men adapted to their environments, how they used their various economic, social, and political institutions, and how mankind evolved from simple to complex societies. Historians and philosophers among the ancient Greeks, Arabs, and Chinese all asked such questions. To take only the example of western Europe, many pertinent questions were posed by the French philosophers Jean Bodin and Michel de Montaigne as early as the 16th century, by the English philosophers Thomas Hobbes and John Locke in the 17th, and by the French philosophers Montesquieu, Rousseau, and Voltaire in the 18th, to mention only those who are often placed among the precursors of modern anthropology.

Historical changes affecting the birth of anthropology

19th-century beginnings. Modern anthropology began to take shape before the middle of the 19th century because of a series of innovations in the Western world. The last great phase of the discovery of the world had begun at the end of the 18th century. At the same time, political and intellectual revolutions had facilitated the questioning of certain religious dogmas, thus opening the way to the discussion of hitherto half-forbidden subjects. The 19th century, therefore, soon saw a revival of interest in and study of the origin of man, the unity or plurality of the human species, and the fixity or mutability of animal species.

Thus, the science of anthropology developed as an outgrowth of contemporary studies of the classification of human races; of the comparative characteristics of human anatomy; of the history of human settlements; of the classification of languages and the comparison of grammars; of the comparison between primitive and ancient societies; and of the historical development of man's economy and industry. Finally, about 1840, a principle for the study of human facts was proposed: the concept of evolution. This was even before Charles Darwin had published his celebrated *Origin of Species* (1859). This concept, arising in strong debates, provided the starting point for anthropology.

Evolutionism. Almost to the end of the 19th century, evolutionism determined the complexion of the new science. A major task of cultural anthropology was thought to be that of classifying different societies and cultures and defining the phases and states through which all human groups pass—the linear interpretation of history. Some groups progress more slowly, some faster, as they advance from the simple to the complex, from the homogeneous to the heterogeneous, from the irrational to the rational. It suffices to quote an American anthropologist, Lewis Henry Morgan:

> As it is undeniable that portions of the human family have existed in a state of savagery, other portions in a state of barbarism, and still other portions in a state of civilization, it seems equally so that these three distinct conditions are connected with each other in a natural as well as necessary sequence of progress (*Ancient Society,* 1877).

Other quotations from a Scotsman, John F. MacLennan, or an Englishman, Edward B. Tylor, would take the same position.

Cultural anthropology, then, set out to analyze the totality of human culture in time and space. But by assuming a linear conception of history, it too often neglected the discontinuities and interferences of concrete history. Morgan, and particularly Tylor, however, sometimes felt the necessity of introducing the concept of the "diffusion," or spread, of cultural characteristics from one people to another—thus suggesting that characteristics could develop independently and converge and that a people could leap over "stages" of evolution by borrowing knowledge from others. Moreover, because it based itself on a theory that all mankind had a similar psychic outlook or that something called "human nature" was universal, anthropology also failed to take into account the fact that the same cultural trait can mean different things depending on the society in which it is found.

Impact of Marxist evolutionism

Marxism and the collectors. At the same time, in the second half of the 19th century another kind of evolutionism developed, that of Karl Marx and Friedrich Engels. Partly independent of anthropological evolutionism (Marx's *Critique of Political Economy* dates from 1859), partly linked to it (Engels' most important work appeared after Morgan's *Ancient Society* and made use of it), the Marxist theory laid stress on the causes of human evolution. A society was defined by its mode of production, on which its political, juridical, and ideological superstructures were allegedly based. These superstructures continued to exist after the mode of production had changed; and in the conflict that followed, this contradiction opened the way to a new type of society. Numerous anthropologists have taken the Marxist analysis into account, even if only to retain its historical view and to reject its economic determinism.

During this same period, especially toward the end of the 19th century, the tales of missionaries, traders, and travelling adventurers included an abundance of miscellaneous information that was collected in such works as Sir James Frazer's *Golden Bough* (1890) and Ernest Crawley's *Mystic Rose* (1902). These rather encyclopaedic collections of customs, religious and magical practices, and other curious data were read with relish by the intellectual community; the theories that accompanied the collections were equally appreciated by evolutionary-minded anthropologists, as the theories were meant to establish an evolutionary sequence of magical, religious, and scientific thought, using the data as evidence.

20th-century trends. By the beginning of the 20th century, many cultural anthropologists had already begun to turn toward what might be called a more pluralistic viewpoint. To account for the variety of societies and cultures and the broadening of the differences that separated them, they suggested taking the total circumstances of each human group into account by considering the whole of its history, the contacts that it had had with other groups, and the favourable or unfavourable circumstances that had weighed on its development. Such a view was distinguished by a marked relativism: each culture represented an original development, conditioned as much by its social as by its geographical environment and by the manner in which it used and enriched the cultural materials that came to it from neighbours or others (through "diffusion") or from its own creativity (through "invention" and "adaptation").

Boas and the culture history school. Cultural anthropology was also diversifying its concepts and its areas of research without losing its unity. Franz Boas, a German-born American, for example, was one of the first to scorn the evolutionist's search for selected facts to grace abstract evolutionary theories; and he inspired a number of students—Ruth Benedict, Alfred L. Kroeber, Margaret Mead, and Edward Sapir—to go out and seek evidence of man's behaviour among people in their natural environs, to go into the field to gather facts and artifacts and record observable cultural processes. He thus is known as the founder of the so-called culture history school, which for much of the 20th century dominated American cultural anthropology.

Emphasis on field work and observation

Beyond this emphasis on field work and first-hand observation, it may also be said that Boas inclined toward what was called functionalism or the functional approach—an approach based on sociological theories of the late 19th and early 20th centuries that tended to liken societies to living organisms or machines, with interdependent parts. In the words of Melville J. Herskovits, one of Boas' students,

> the *functional* view, attempts to study the interrelation between the various elements, small and large, in a culture. Its object is essentially to achieve some expression of the unities in culture by indicating how trait and complex and pattern, however separable they may be, intermesh, as the gears of some machine, to constitute a smoothly running, effectively functioning whole (from *Man and His Works,* 1948).

Boas insisted upon this method of considering any single culture as a whole. Finally, by emphasizing the importance of collecting life histories, he drew attention to the problems posed by connections between culture and personality.

Mauss and the "sociological" school. In a similar way, Marcel Mauss, in France, influenced the characteristic tendencies of a whole generation of European sociologists and cultural anthropologists, including Alfred Métraux and Claude Lévi-Strauss, and founded the Institute of Ethnology of the University of Paris; he also influenced such men as the noted British cultural (or social) anthropologists Bronislaw Malinowski and Arnold R. Radcliffe-Brown. In general it may be said that Mauss, like Boas, was insistent upon studying social phenomena as a system—but in a slightly different fashion. Like many others of his time he conceived of systems as self-regulating or equilibrium-seeking, composed of elements that operate to maintain the integration or adaptation of the system. Mauss gave impetus, in fact, to what was called structuralism or the structural approach, which focussed more on society as an indivisible social organism than on society as an interrelation of individuals (the functionalist's emphasis). Like Boas, Mauss also tried to twin culture and personality—that is, cultural anthropology and psychology.

The "grand diffusionists." The large and influential American school of "culture history" anthropologists led by Boas should not be confused with a distinct and smaller group of Austro-German diffusionists, led by Fritz Graebner and Wilhelm Schmidt, who constituted what has been called the "culture-historical" school in Europe. These latter, too, had rejected classical 19th-century evolutionism, but they were nevertheless inclined toward painting grand theories—principally the theory that out of a few ancient cultural centres or civilizations, born quite separately, there had developed all the array of cultures existing today. Diffusion, or the spreading of culture traits, in their view, was the prime force of human development, and all cultural development could be traced to a few inventive centres. Because they termed these original centres Kulturkreise, (or "cultural clusters"), they were also known as the Kulturkreise school of cultural anthropology. This kind of pseudo-history was carried to even greater lengths by a British group of diffusionists, led by Grafton Elliot Smith and William J. Perry, who even named a single fountainhead of all cultural development—Egypt.

The Kulturkreise school

Functionalism and structuralism. Some schools of research that began to develop between the two world wars more or less vigorously rejected the historical approaches, sometimes denying any interest in them whatever. According to the cultural functionalists, including the followers of Malinowski, the only way to explain facts was to define the function that they performed currently in a given culture. The aim of all cultural anthropological research, they held, should be to perceive the totality of a culture and the organic connection of all its parts. Consequently, comparison did not make sense: each culture was a unique reality. History, moreover, made no more sense; a culture was to be interpreted at one point in time, as if the age and the origin of the elements composing it were without importance. The only thing that counted was the function the elements performed *now.* Earlier cultural anthropologists had talked of "survivals," customs or other cultural traits that survived from out of the past though no longer with any real function or meaning. But Malinowski would say, "There are no survivals"; everything current, according to the functionalists, has *some* function.

Whereas the name of Malinowski is supremely associated with the school of functionalism, the name of Radcliffe-Brown is known as one of the most important proponents of present-day structuralism. Relying on the concepts of formal mathematics and linguistics, Radcliffe-Brown and other structuralists tried to determine whether in cultural anthropology it was possible to reveal that which "suggests the character of a system" beyond empirical reality and which "alone is the true object of science" (Lévi-Strauss). A structure is not a sum of social relations, which are only the primary material from which the observer extracts "structural models." A structure is a system of which the members of the society being studied are not aware or only partly so. The model that the cultural anthropologist constructs from the system is valid when the model's operation can account for all the observed facts. This exacting approach has proved particularly useful in studying kinship and marriage relations as well as myths. The difficulties of using this approach in other fields, as well as the fact that historical changes are difficult to include in this sort of static analysis, strengthen the objections that many workers in the field have raised against it.

Structure unperceived by society's members

Cultural psychology. One development of the interwar period led certain cultural anthropologists to speak of a new subdiscipline, cultural psychology, or ethnopsychology, which is based on the idea that culture conditions the very psychological makeup of individuals (as opposed to the older notion of a universal psyche or human nature). In the 1930s, for instance, in her studies of the American Southwest, Ruth Benedict found that the ways in which the Pueblo Indians thought and reasoned were strikingly different from the ways in which their immediate neighbours thought and reasoned, even though their geographical environment was virtually identical. Her conclusion was that each culture over the ages had evolved and given to its members a unique "psychological set" or orientation toward reality and that this set actually determined how the members saw and processed information from the environment. Culture, in effect, affects the ways in which the mind works.

Studies in culture and personality have developed in many directions. Research into forms of child rearing, for instance, have called in question the universality of Freudian propositions concerning parent–child relationships. There have been many studies of value systems, which give a culture what has been called its "configuration," or of the personality types prized or rejected by each culture, or of the "national characteristics" of certain modern societies. The results of these studies have, however, been uneven in quality.

Neo-Marxism and neo-evolutionism. Finally, certain theoretical tendencies of the 19th century came back into favour. For political reasons Soviet cultural anthropologists have conducted their research in the tradition both of Marxist analysis and of a fairly rigid evolutionism. Even their choice of subjects is sometimes linked to official ideology—as, for example, a program of religious anthropology aimed expressly at the "elimination of religious prejudice in the Russian population." Elsewhere, in France, for example, a brand of neo-Marxism has influenced a new generation of cultural anthropologists to concentrate on analyses of primitive economies. Classical evolutionism, meanwhile, has been revived in the United States by some cultural anthropologists who speak of "multilinear evolutionism" or many paths to modernization.

STATUS OF CONTEMPORARY CULTURAL ANTHROPOLOGY

The problem of making cultural anthropology "scientific"

It is true that cultural anthropology has not reached a state of complete coherence. This is clear from the persistence of divergent national traditions and from the way in which research can be impregnated with explicit or implicit ideologies. It is also true that different schools of thought coexist in the same country and that cultural anthropology is not therefore based on a unified body of concepts, whereas a science is defined above all as a homogeneous language for interpreting a specific level of reality. A "science" of culture would seem possible only if anthropologists could free themselves of ethnocentrism and produce concepts and other elements that were universal, objective, and theoretically significant. The functionalists think they have fulfilled these conditions. The structuralists challenge this and, in their turn, try to fulfill the conditions. Thus cultural anthropology—as opposed, for example, to linguistics—has developed only very partially a terminology independent of a national or private language. These limitations are still encountered by most of the social sciences. But cultural anthropology's primary aim—to permit cross-cultural comparability—makes the problem even more serious.

The new research and fieldwork. Cultural anthropology is undergoing a crucial test of another kind. Its traditional objects of study—"primitive" or "traditional" cultures—seem to be disappearing. Either they are dying out because they find it impossible to adapt themselves to a modern world or they are transforming under the direct or indirect influence of modern industrial societies. Moreover, those that do remain at a folk level often take exception to being placed among societies that are the subject of anthropological study, seeing this as a manifestation of condescension and a vestige of domination.

Much cultural anthropological research and study has entered the library or laboratory. One of the criticisms of Boas and others engaged in pure fieldwork was that they were collectors rather than systematizers. There is thus a considerable wealth of ethnographic data to be analyzed, collated, classified, and interpreted in order to be made useful. Files of information are being arranged in what are called Human Relations Area Files. More and more typologies are being constructed, typologies based on political systems or technology, or systems of kinship. In addition, new readings of the material are being attempted in the hope that mathematical formulations or models might be obtained. Also emerging is the study of insufficiently known societies by techniques of simulation.

Many cultural anthropologists refuse to turn to the laboratory and continue to do fieldwork, either among Western populations or among modernizing, formerly colonial populations. They are joined in this task by researchers native to those populations. For some anthropologists these field studies provide an opportunity for a true anthropological experiment, determining how people respond to modernizing influences and how elements of the old culture evolve into those of the new. Such anthropologists tend to reject the concept that social systems seek integration and "equilibrium." Instead they propose a more "dynamic" interpretation of traditional societies and emphasize the role played therein by tensions and conflicts.

Economic development and anthropology

In any case, at a time when the problems of development are among the primary cares of the world, a growing number of anthropologists are devoting themselves to research the results of which can be used in political policy and decision making—whether they are employed directly by interested governments, or lent by foreign governments or international organizations, or recruited by foundations for study and development.

Non-Western cultural anthropologists. A significant development in the latter half of the 20th century has been the emergence of more and more non-Western cultural anthropologists. Originally, cultural anthropology was a Western interest and endeavour, and it has continued to be dominated by Westerners. Even in non-Western countries where anthropology institutes and university departments have begun to multiply somewhat—as in Japan, India, and some Latin-American nations—cultural anthropologists have remained rather constricted. Japan is a good example. Cultural anthropology as an independent science there is still young, having arisen largely only since World War II; and most Japanese cultural anthropologists in the schools have had to be hybrid teachers, attaching themselves to sociology or social science departments and teaching sociology or some other related discipline in addition to cultural anthropology. Not only have cultural anthropology courses been few but also funds for field studies have been limited, so that there have been few lengthy and intensive studies; what research there has been has focussed largely on Japanese or other East or Southeast Asian communities. Furthermore, Japanese cultural anthropologists have shared a problem faced by many non-Western researchers, in that the native language in which they write has not been as readily accessible to foreigners as have been western European languages. "International communication," the Japanese cultural anthropologist Takao Sofue has noted, "has [thus] been seriously restricted with the result that Japanese scientists have been isolated from effective criticism from abroad" ("Social Anthropology in Japan," *American Behavioral Scientist,* 12:15–17, Jan.–Feb. 1969). It has also meant, of course, that they have not been sufficiently widely read abroad to make their influence felt. This problem, though, is not so serious in non-Western countries like India, where a European language constitutes a major language of scholarly communication.

Applied studies. From the cultural anthropologist's point of view, applied studies—that is, research meant to give practical aid and guidance to governments and other organizations—have in many ways been an undoubted gain. Concerned as they so often were with the effects of social change, applied studies offered the nearest approach to the controlled experiment in the social sciences. The specialized inquiries greatly deepened the knowledge of particular aspects of primitive society and culture, especially of economic and political organization, land tenure, and law. The scientific value of such research apart, work in the applied field also offered to many anthropologists the purely human satisfaction of aiding backward peoples in their struggle to meet and master the forces of Western civilization.

The concrete gains derived by colonial governments were more difficult to assess, partly because the officials were not bound to act upon the cultural anthropological findings and partly because the value of the findings was not always wholeheartedly accepted. Sometimes, it is true, the cultural anthropologist found himself embarrassed by the excessive confidence of his employers that he had the key to all problems. More often, the employers were inclined to question whether cultural anthropology was in fact as helpful and the information it provided as indispensable as enthusiasts would make it out to be. Some impatience was felt with the "academic" cultural anthropologist who would insist on comprehensive studies when only some specific information was asked for, or who seemed to deal in a complicated fashion, using complicated language, with issues that to the practical man appeared straightforward. To all this cultural anthropologists could reply that, though the knowledge they sought was not indispensable to government, it facilitated informed and smooth government.

But cultural anthropologists also had to face another, more disturbing criticism—that they overemphasized the importance of tradition and were hostile to modern development. Nor was this view limited to colonial administrators; educated Africans and Indonesians openly expressed their distrust of a science the primary interest of which was in "primitive" peoples and which might play into the hands of reactionaries and upholders of "colonialism."

If these objections did not promise too well for the future of applied cultural anthropology, cultural anthropologists themselves had grown more cautious. They came to fear that the applied work might entice too many of the younger cultural anthropologists away from general and theoretical research, so that the very progress of the discipline might be endangered. Conversely, the man fully committed to applied work, like the permanent government cultural anthropologist, would be in danger of losing touch with universities and academic centres, and hence with the advances achieved in his discipline. He would turn into a mere technician, perhaps still useful to his employers but no longer truly representing anthropological knowledge.

The theoretical researcher versus the technician

There were graver problems of an ethical nature. A change of roles is forced upon the cultural anthropologist when he is consulted on the best way to implement government policies. To be sure, he might see no cause for disagreeing with the policy, and the best way of imposing it might well be understood to be the one best serving the interests of the native peoples. Even so, the cultural anthropologist, in abandoning the point of view of the scientist, must pronounce upon the merits and demerits of particular courses of action and thus introduce value judgments. Nor will the issues always be clear-cut and uncontroversial; in that case the cultural anthropologist might have to take sides and argue from his own political and moral convictions. And if his recommendations had little chance against administrative considerations or the dictates of "higher policy," personal frustrations would be added to the dubiousness of his position.

On the other hand, if the cultural anthropologist presented his facts without adding recommendations or warnings, he would be furnishing information that might be put to uses with which he could not in good conscience agree. Or again, he might be tempted to restrict his advice to the most efficient means for achieving certain ends, dismissing the ends themselves, the policy to be implemented as not of his concern—which would hardly diminish his ethical commitment.

All these issues were widely and on occasion heatedly debated among cultural anthropologists. In an attempt to clear the air the Society for Applied Anthropology published in 1951 a carefully worded code of ethics. It appealed to the social conscience of the individual research worker and to his responsibility at all times to uphold the moral tenets of civilization—respect for the individual and for human rights and the promotion of human and social well-being. Not all cultural anthropologists were prepared to endorse this assumption of a moral mission on the part of the "disinterested" scientist. The dilemma, then, though vital for the future of applied cultural anthropology, remained unresolved. (Pa.M./Ed.)

Sociology

Sociology is a branch of the science of human behaviour that seeks to discover the causes and effects that arise in social relations among persons and in the intercommunication and interaction among persons and groups. It includes the study of the customs, structures, and institutions that emerge from interaction, of the forces that hold together and weaken them, and of the effects that participation in groups and organizations have on the behaviour and character of persons. Sociology is also concerned with the basic nature of human society, locally and universally, and with the various processes that preserve continuity and produce change.

It is social life that is distinctive in the regulation of behaviour in human beings; the human animal does not have such instincts as serve to guide the behaviour of lower animals, and he is therefore more dependent on social organization than is any other species. Institutionalized social forms therefore are assumed to play the major part in influencing human actions, and it is the task of sociology to discover how these forms operate on the person, as well as how they are established, develop, elaborate, interact with one another, and decay and disappear. Among the most important of such structures is the family, the subject of an important field of sociology. The peer group, the community, the economic and political orders, various voluntary associations, and special organizations such as the church and the military are of particular importance in this inquiry.

Relation of sociology to other disciplines

Though sociology can be considered as a part of the Western tradition of rational inquiry inaugurated by the ancient Greeks, it is specifically the offspring of 18th- and 19th-century philosophy and has been viewed as a reaction against the frequently nonscientific approaches of classical philosophy and folklore to social phenomena. It was for a time presented as a part of moral philosophy, which covered the subject matter that eventually also became the concern of the various social sciences that are now separate from moral philosophy. Some aspects of other fields remain of interest to the sociologist. Although psychology has traditionally centred its interest on the individual and his internal mental mechanisms, and although sociology has given its major attention to collective aspects of human behaviour, the two disciplines share the subfield of social psychology. The relation of sociology to social anthropology is even closer, and until about the first quarter of the 20th century the two subjects were usually combined in one department, differentiated mainly by the emphasis of the anthropologists on the sociology of preliterate peoples. Recently even this distinction has been fading, as social anthropologists have increasingly added studies of various aspects of modern society to their field of interest. Political science and economics had much of their early development in the practical interests of nations and for a time evolved separately from basic sociology; but recently in both fields an awareness of the potential utility of some infusion of sociological concepts and methods has brought relations closer. A somewhat similar situation has also been developing in respect to law, education, and religion and to a lesser extent in such contrasting fields as engineering and architecture.

Nineteenth-century sociology, influenced by the successes of biology and evolutionary theory, took an interest in resemblances between men and lower animals—in their having, for example, similar instincts—and also in the parallels between biological and social evolution. These interests have declined, but sociology continues to share with the other sciences some interest in ecology, behavioral genetics, and questions of fertility and mortality as they relate to population studies. There is also a conviction among sociologists that contact between physiology and sociology is necessary to avoid errors of ignorance in both fields.

HISTORICAL DEVELOPMENT OF SOCIOLOGY

Early major schools of thought. The founders of sociology spent decades almost exclusively in the process of finding a direction for their new discipline. In the course of this groping effort they tried several highly divergent pathways, some suggested by methods and contents of other sciences, others invented outright by the imagination of the scholar.

Social Darwinism and evolutionism. Darwinian evolutionary theory doubtlessly suggested a way in which a science of human behaviour could become academically respectable, and a line of creative thinkers, including Herbert Spencer, Benjamin Kidd, Lewis H. Morgan, E.B. Tylor, L.T. Hobhouse, and others, developed analogies between human society and the biological organism and introduced into sociological theory such biological concepts as variation, natural selection, and inheritance—evolutionary factors resulting in the progress of societies through stages of savagery and barbarism to civilization, by virtue of the survival of the fittest. Some writers also perceived in the growth stages of each individual a recapitulation of these stages of society. Strange customs were thus accounted for on the assumption that they were throwbacks to an earlier useful practice; an example offered was the make-believe struggle sometimes enacted at marriage ceremonies between the bridegroom and the relatives of the bride, reflecting an earlier bride-capture custom.

The notion of stages of evolution

Social Darwinism waned in the 20th century, but in its popular period it was used to justify unrestricted competition and a laissez-faire doctrine in order that the "fittest" would survive and that civilization would continue to advance.

Determinism: economic, environmental, biological. Except in the philosophy of Karl Marx (whose writings ranged over all the social science fields rather than specifically in sociology), the doctrine of economic determinism never gained a strong foothold in sociology. This was not a consequence of scholarly ignorance; sociologists of all periods have read Marx and have usually read such writers as the historian Charles A. Beard, who emphasized economic self-interest, and Werner Sombart, the German sociologist who had been a convinced Marxist in his early career. But there have been only some adapted reflections of these economic views in the writings of such sociologists as Franklin H. Giddings or Frank H. Hankins who viewed some political and religious doctrines as rationalizations of economic and social interests.

The human geographers—Ellsworth Huntington, Ellen Semple, Friedrich Ratzel, Paul Vidal de La Blache, Jean Brunhes, and others—were also read critically by sociologists but did not make a lasting major contribution to the mainstream of sociological thought, even though there are some who believe that the social morphology of Émile Durkheim, Maurice Halbwachs, and others—that is, their theories about the roles of individuals interacting in a social system—grew in part from this interest.

Aside from the interest in evolution, organismic analogies, and the instinct concept, sociologists have not found biological determination of value to them and have spent more energy in refuting it than in making use of it.

Early functionalism. Following the achievement of a consensus that there should be a place for a science of sociology, there emerged an international effort to define the distinctive character of the subject and especially to clarify its differences from psychology and biology, fields that had also begun to generalize about human behaviour. A Frenchman, Émile Durkheim (1858–1917), was prominent among scholars who considered this question; he argued that there can arise from various kinds of interaction among individuals certain new properties (*sui generis*) not found in separate individuals. These "social facts" as he called them—collective sentiments, customs, institutions, nations—call for study and explanation on a distinctly sociological level rather than on the level of individual psychology. Furthermore, the interrelations of the parts of a society were perceived as cohering into a unity, an integrated system with a life character of its own, exterior to the individual, and exercising constraint over his behaviour. This direction of causation, from group to individual (rather than the reverse as conceived by most biologists of the time) gave encouragement to the scholar of the new science. Some writers have designated such a view "functionalism," although the term has in recent years acquired some broader variations of meaning.

Durkeim's "social facts"

Durkheim also pointed out that groups could be held together on two contrasting bases: the sentimental attraction of similarities (mechanical solidarity), such as occurs in friendship groups and among relatives and neighbours, and the organization of complementary differences (organic solidarity), such as occurs in industrial, military, governmental, and other organizations that exist because they have tasks to perform. Other theorists of Durkheim's period, notably Henry Maine and Ferdinand Tönnies, made similar distinctions in different terms—*status* and *contract* (Maine) and *Gemeinschaft* and *Gesellschaft* (Tönnies)—and conceived of the major trend of civilization as an expansion of the latter and a relative decline of the former.

Some later anthropologists, especially Bronisław Malinowski and A.R. Radcliffe-Brown, developed a doctrine also called functionalism, based on the recognition of the interrelatedness of the parts of a society, in bonds so thoroughly interpenetrating that a change in any single element would tend to produce a general disturbance in the whole. This concept gained a following for a time among many social anthropologists, leading some to advocate a policy of complete noninterference with even the most objectionable practices in a preliterate society (such as headhunting) for fear that control might produce far-reaching disorganization.

William G. Sumner, in his *Folkways,* defined an institution as a "concept and a structure," meaning a purpose or function that is carried out by some systematic organization of persons. Much of the sociology of Max Weber consists of the analysis of societies in such terms. Georg Simmel, sometimes called the founder of the "formal school" of sociology, viewed society as a process ("something functional") that is real and not merely an abstraction, and he built on this idea a statement of sociology consisting of a systematic analysis of social forms.

Modern major directions of interest. The early schools of thought—each presenting a systematic formulation of sociology that implied possession of exclusive truth and that involved a conviction of the need to destroy rival systems—in time gave way to distinguishable directions of interest and emphasis that did not have to be considered inharmonious. These new directions have no dominant leaders and no clearly defined borderlines.

Functionalism and structuralism. Following the main contributions in the earlier theoretical formulations of Charles H. Cooley, such later authors as Pitirim A. Sorokin, Talcott Parsons, Robert Merton, Everett C. Hughes, and others have elaborated on the nature of organizations and their relation to the behaviour of persons and have attempted to build workable conceptualizations of very large social systems, nations, and societies. Sorokin designated his viewpoint as "integralist" and wrote at length about the civilization-cultures that in their balance of values and conditions could be viewed as entities that had distinguishable life cycles, with "ideational," "idealistic," and "sensate" stages marking their growth and decline, thus following a philosophy-of-history tradition shared by Edward Gibbon, Oswald Spengler, and Arnold Toynbee.

Cultural life cycles

Talcott Parsons has given attention to social systems in a more analytical way, inquiring into the conditions that each system must meet in order to survive (the "functional prerequisites"), the character of the standardized and stable interpersonal arrangements (structures) needed to make each system work, the relations to environmental conditions, problems of boundaries, the recruitment and control of members, and the like. Along with Robert Merton and others, he also worked on the classifications of such structures and on distinctions of function.

The subject matter and methods involved in such structural-functional analysis have indeed become so broad that some authors (such as Marion Levy) have held that it becomes synonymous with scientific analysis in general, or at least with scientific study of the nature of organization.

On a smaller scale, Kurt Lewin and his co-workers pursued somewhat parallel questions, investigating the nature of small groups, families, professional and military units, looking for arrangements and relationships of the parts of each person's "psychological life space" and of the interrelations of these to a "social space" or society's total range of action. The choice of such relatively small units for research made fruitful experimentation possible, and from Lewin's leadership grew the influential research movement that became known as group dynamics. Some writers have also applied the descriptive term *microfunctionalist* to this tradition.

Symbolic interactionism. Sociologists did not for long find the 19th-century instinctivist psychology congenial, and most of them also failed to appreciate the doctrines of classical or Watsonian behaviourism, which sought to be totally objective and experimental. One influential movement in social psychology, however, did take early root and eventually became the largest and most influential field in modern American sociology. In recent years it has become known as "symbolic interactionism," but it was under development for decades before it acquired a name.

Out of early ideas expressed by J. Mark Baldwin and William James, a group of three scholars, John Dewey, George H. Mead, and Charles H. Cooley, built the foundations of a psychology that was to become most useful to sociology. In brief, their contribution was to advance the theory that mind and self are not part of the innate

The concept of self

equipment of the human organism but arise in experience and are constructed in a social process—that is, in a process of interaction among persons in intimate, personal communication with one another. The self, or self-concept, as developed by Mead and others, is thus essentially an internalization of aspects of an interpersonal or social process. It exists in imagery and symbolization and is internalized and organized for each person out of his perception of how other persons conceive him. This self-concept, however inexact, fluctuating, and uncertain, nevertheless functions as a guide in social behaviour—that is, persons tend to act in order to preserve the existing or desired image of their self.

William I. Thomas, a sociologist and colleague of the philosopher Mead at the University of Chicago in the early years of the 20th century, regularly taught a course in social psychology based on Mead's conceptions. Thomas was succeeded in 1919 by Ellsworth Faris, himself a psychologist but later a member of the department of sociology, and through his work the tradition was further developed and brought into closer relation to the sociological tradition of Robert E. Park and Ernest W. Burgess, also at Chicago. In this tradition an interest in an appropriate methodology accompanied the growth of substantive knowledge; Thomas particularly emphasized the value of extensive use of personal documents, life histories, and autobiographies. In recent years interest in research on the self and self-conscious behaviour has spread widely, and is now participated in by psychologists, philosophers, and essayists, as well as by a movement within sociology called "ethnomethodology," which investigates areas of symbolic interaction by informal observation, reflection, and skilled interpretation, methods sometimes called *Verstehen* (understanding).

Modern determinism. Economic determinism reflects the interest that a few early sociologists took in views of Karl Marx, such as the idea that differentiation into social classes and conflict between these classes derive from economic factors and the belief that the political system is in large part a product of such social stratification. A residue of this kind of determinism is found among the self-proclaimed "Marxian sociologists." Perhaps the most widely read of these was C. Wright Mills, whose concept of a "power elite" has been extensively and critically examined, with varying resulting judgments on its utility. As Mills saw it, this elite constitutes an integrated ruling group of a capitalistic economic and military system, sometimes called the military-industrial complex, exercising arbitrary power in its own interests. This particular determinism is not supported by most existing objective research, which generally finds a far more pluralistic distribution of political power.

A contrasting view of class conflicts was advocated by Karl Mannheim, who saw the cleavages as ideologically produced, as divergences in modes of thought rather than as rational perception of economic interests. Since Mannheim hoped that such conflicts could be resolved, his doctrine should not be considered fully deterministic, but it did stimulate an effort to interpret the relations between ideas and actions that came to be known as the "sociology of knowledge."

Mathematical modelism. A variety of efforts has been made to describe and investigate behaviour mathematically, through measurement and counting and the use of mathematical models. This approach in part characterized the early "sociometry" of J.L. Moreno (although its meaning has greatly drifted and broadened in recent years), the "field theory" of Kurt Lewin, and the investigations by George K. Zipf, John Q. Stewart, and others into the relations of rank and size of political units, the frequency of word use in language, and other simple arithmetic relations. Some of the concepts of game theory, first introduced into economics by its inventors, John von Neumann and Oskar Morgenstern, have also penetrated into sociology. Also the rapidly expanding use of computers has in recent times encouraged the development of various kinds of simulation of behaviour. Some investigations of complex interaction patterns have been carried out by devising games with rules to fit the problem and persons to execute the roles. When specified rules become highly detailed and complex, the outcome may be sought through the use of a computer; thus the game is converted into a simulation. Sociologists have participated, along with other social scientists, in the creation of such simulations of various political and military processes. Extension of these techniques into a variety of interaction processes is to be expected.

METHODOLOGICAL CONSIDERATIONS IN CONTEMPORARY SOCIOLOGY

Much of 19th-century sociology was devoid of systematic method, but late in the period the proliferation of schools of thought, based on speculative sociologies, made evident the need for ways of obtaining verifiable knowledge. Early attempts were crude and unfruitful; such broad surveyors as Charles Booth, who produced a monumental series on London, relied mainly on the gathering of masses of facts. Frédéric Le Play in France made extensive studies of family budgets. Herbert Spencer and others assembled vast stores of observations made by other persons, using these to illustrate and support generalizations already formulated.

Early exploitation of statistical materials, such as officially recorded rates of births, deaths, crimes, and suicides, provided only a moderate advance in knowledge, because this approach was too capable of supporting preconceived ideas. Among the most successful of this type of study was research on suicide by Émile Durkheim, whose successors in France and elsewhere developed the methodology a considerable way toward scientific adequacy.

After the turn of the century, interest in, and the determination to achieve, a sociological methodology grew steadily. The *Methodological Note,* constituting the greater part of a volume in W.I. Thomas and Florian Znaniecki's *Polish Peasant in Europe and America* (5 vol., 1918–20), has been recognized as an important advance, not so much in methodology as in committing sociologists to the task of achieving it.

Inductive research

Significant advances toward scientific effectiveness occurred at the University of Chicago in the 1920s. Under the stimulation of Robert E. Park, Burgess, and their colleagues a series of studies of the metropolis was conducted. The spirit was inductive, and hypotheses were discovered in rather than imposed on gathered information. Large numbers of students took part in the effort and contributed to both methods and findings. A conspicuous part of the effort consisted of mapping locations of various phenomena: land uses, residences of population categories (racial, ethnic, and occupational), residences of persons who commit various types of crimes or suicide, families becoming divorced or broken through desertion, and so forth. But along with such information on spatial distributions, data were sought by other means, including participant observation in groups and communities, gathering of life histories and case studies, assembly of relevant historical information, study of the life cycles of social movements and sects, and the like. Attention was explicitly given to the improvement of methodology in all of these efforts, to an extent approximately equal to the attention given to substantive findings. Here for the first time was developed a large-scale cooperative effort in which theory, methodology, and findings evolved together in an inductive process. The influence of this development at Chicago spread rapidly about the United States and in time influenced sociology almost everywhere it was studied in the world.

Statistics. Statistical methods were introduced into sociology from other sciences, and virtually from the start, sociologists have found statistical measures of relationship of great value. Karl Pearson's "coefficient of correlation," for example, has been a popular as well as important statistical concept for the measurement of cause-and-effect relationships among continuous variables. This method reveals the *degree* of causal connection between two variables, though not necessarily the *nature* of the connection. In sociology there are types of data that are relevant to causal inquiry but do not have the characteristics that qualify for the Pearsonian coefficient. Thus, much development work has been done to provide other measures of association involving, for example, rankings of groups or individuals or qualitative comparisons (such as whether males and females differ systematically in specified qualities).

Factor analysis, also based on an elaboration of Pearsonian correlation, performs another valuable service to sociology. If there are a large number of variables causally intertwined in a complex way, it is possible that these variables can be reduced to a small number of factors. Fifty different tests of mental ability, for instance, may be in fact 50 different mixtures of only seven or eight dimensions of mental ability. Factor analysis involves reducing such variables to a more limited number of common factors and determining the relative importance of each factor in the original variables. The process has its imperfections and the computations are laborious, but the availability of computers has overcome the latter disadvantage, and in recent years the technique has increased in use.

These statistical methods and many others are applicable to all branches of sociology and are increasingly fruitful in transforming sociology into science (see also STATISTICS). In general, the growth of statistical methods has been so rapid that the invention of new techniques has outstripped the ability of scholars to find data worthy of the devices. Thus the rate of progress in the near future may depend to a large extent on improvement in satisfactory data gathering and measurement. Methodologies of data gathering are in fact of major interest in sociology. Techniques of observation—of persons, groups, organizations, communities—have been extensively developed. Important for the same purpose are the various means of quantifying these observations, including scales of various kinds, sociometric techniques that make interrelations subject to statistical analysis, content analysis of written materials, and classification of cross-cultural information.

Experiments. Experimental methods, once believed to be inapplicable to sociological research, were extensively applied by psychologists, first on individuals and later on groups. By the 1930s some psychologists—notably Kurt Lewin and his colleagues and also Muzafer Sherif—found means of conducting experiments on social interaction. Sociologists soon followed their example and in time a number of laboratories for such research were established; Robert F. Bales, at Harvard, has made systematic observations on interaction in small, artificial groups and has produced clear and useful results, confirmed in other laboratories. Experiments are also conducted in classrooms, in summer camps, in formal organizations, and elsewhere. In general the success of experimentation has been greatest in simple situations in which the number of variables is limited. Complex experiments, however, are possible in some circumstances, and the design of complex formal experiments is becoming a developed art in a variety of fields, including sociology.

Harvard interaction studies

Data collection. Within the main categories of research methods there are many special problems for which techniques have been devised. Data collection, for example, is effected in many different ways, from unstructured observation, essentially methodless, to sophisticated measurement through special instruments. Some of the basic problems of data collection concern such matters as the most efficient use of terminology, the definitions of units to be measured, and the classifications to be used. In general it is necessary to consider the nature of a specific problem in order to choose the most appropriate unit. For example, in a study of the relation of the size of a city to the cost of operating its local government, the proper unit might well be the population residing within its political boundaries. If the research question, however, is the relation of city size to any of a number of forms of social disorganization, it may be more fruitful to recognize that sociologically the significant unit would include much or all of the settled areas outside the city limits.

In the fields of social differentiation and occupational mobility the matter of definition of specific occupations is critical. If persons are asked in a questionnaire to state their occupation, the usual response is to give only one occupation, and this one is sometimes vaguely defined and made obscure by the tendency to give a euphemistic answer. Persons change occupations; some have more than one; some might claim an occupation that they merely aspire to. The art of obtaining useful answers to such important questions involves carefully designed questions adapted to the specific purposes of the study. General classifications, intended for a variety of studies, have limited utility.

In the process of gathering research data for sociology there are occasional obstacles to direct observation. In such cases indirect indicators may provide crude but useful substitutes. For example, alcoholic consumption in a small village in which the beverage is supposed to be prohibited may be estimated by a count of empty bottles in trash receptacles, or perhaps in the town dump. Library book circulation has been used to estimate the use of television in a community in which withdrawals of books of fiction declined, while nonfiction withdrawals remained as before.

Questionnaires

Questionnaires are convenient for obtaining information from large numbers of respondents but involve many methodological problems. Wording of questions must of course be intelligible to uneducated and uninterested persons, must have standard meanings to persons of varying backgrounds, must avoid topics that arouse resistance and refusal to complete the questionnaire, and must avoid being too complex or difficult so that returns are insufficient or constitute a biased sample. Since it is known that slight alterations in the wording of questionnaire items may produce considerable variations in the pattern of responses, the precise wording becomes a matter of some art as well as science. A similar effect occurs in the order of items, since some may suggest or influence responses on later ones.

Interviewing

Similar issues are involved in data gathering through interviewing. It is necessary to control such variables as the appearance, manner, and approach of the interviewer, the specific manner in which questions are asked, ways of avoiding interviewer influence on the responses, and the tendency of some respondents to refuse to answer questions or to discontinue the interview. To meet the problems of resistance on sensitive subjects and inarticulateness about some feelings, various indirect or projective devices may be employed so that a respondent in answering one question provides information he may not realize he is giving about other questions.

Questionnaires and interviews may be so arranged that the patterns of responses form a scale, converting qualitative variations into measures available for statistical treatment. An early scaling method, devised in the late 1920s by a psychologist, L.L. Thurstone, is still widely used in sociology. It is formed in the following way: a list of questionnaire items is presented to a number of judges who independently relist the items in the order in which they consider them important or of interest. From their decisions are selected items on which there is satisfactory agreement of scale value.

Scaling may also be provided by statements to which a respondent is asked whether he "strongly approves," "approves," is "undecided," "disapproves," or "strongly disapproves." Or the quantitative differences may be introduced through a logical sequence of preference answers—for example, whether the respondent would admit a particular category of person (a) to close kinship by marriage, (b) to his club as a personal chum, (c) to employment in his occupation, (d) to citizenship in his country. Here it is assumed that the later answers imply more desired social distance.

Sociometry

A method or class of methods called sociometry has been under development since its introduction in the middle 1930s by J.L. Moreno. The essence of the method is the collection and tabulation of information about various types of interaction among members of groups of small or moderate size. The interaction may be either actual behaviour or merely anticipated or desired behaviour, and it may consist of preferences for various kinds of association with other persons, such as having them as friends, sitting with them, working with them, and the like. The information may be collected by observation of real behaviour or by interviews or questionnaires with specific items regarding personal choices. After the information is gathered, it is sometimes put in the form of a sociogram, consisting of names of persons enclosed in circles or squares distributed over an area and connected with lines and arrows that indicate both detail of choices and general patterns of relationships. A person receiving many choices is readily seen as the target end of many lines and is sometimes referred to as a "star." A person completely unchosen has no lines pointing toward his name and is called an isolate. Further investigation of persons typed in this fashion may be made by statistical methods, case studies, or otherwise. Overall, it can be said that various improvements and elaborations of the basic sociometric approach have been made, and the method is now less distinctively separate from other social psychology research than it was originally.

Ecological patterning. Ecological methods in sociology were first developed in connection with research on the characteristics of the metropolis, especially in regard to features of a nonsocial character, such as the patterns resulting from the distribution and movements of populations and institutions in the general process of struggling for advantage. A conspicuous part of most early urban studies consisted of mapping such distributions. The patterns of land values, of locations of various types of businesses and industries, of ethnic categories of the population, and of types of behaviour (delinquency and crime, vice, family disorganization, mental disorders, etc.) were all shown to be interrelated in a general urban ecology. This fact was then shown to be related to many aspects of behaviour of city people, and valuable contributions were made to such general sociological topics as social differentiation, migration and vertical mobility, and social disorganization.

In recent years sociological ecology has broadened in meaning and in the elaboration of methods. One modern approach, known as ecosystem theory, consists of tracing general patterns of flow of materials, energy, and information into a system and their transformation during the flow through the system, among other things.

Problems of bias. Since most sociological knowledge is based on the study of samples from some larger universe of items, the possibilities of major errors from sampling bias constitute a methodological issue. Where biases cannot be controlled, the direction and extent may sometimes be estimated, but elimination of biases through use of quotas—or, when possible, random methods—yields the best results. This can be done, for example, by first randomly selecting a number of definable regions and metropolitan areas, then selecting randomly from each such area certain urban blocks and rural segments, then further selecting from these segments certain dwelling units, and finally selecting from the dwelling units the specific persons to constitute the sample.

Search for error

In every stage of the process of discovery in sociology there are possibilities of error, and recognition of these is a part of the progress of sociological methodology. There is continuous creation of technical devices to reduce such errors and to estimate the amount of error that has not been eliminated.

National methodological preferences. All the methods described above are widely used, but their relative popularity in various nations is somewhat related to both the nature of the financial support of research and the field of national interest. Where agricultural problems are of major interest, rural sociology and community studies that can be conducted inexpensively by one or a few investigators are popular. In France, Italy, and several other European nations, industrial sociology is understandably important, much of it based on case studies of industries and the experiences of workers. Sociology in Great Britain, the Scandinavian countries, and Japan covers most of the fields mentioned above.

The broad methodological concepts have varied somewhat according to the country and according to the subfield of sociology. Early in the century there was presumed to be a general difference between the sociologies of European countries and the sociologies of the United States—the former appearing to prefer broad sociological theory based on philosophical methods and the latter showing more inclination toward induction and empiricism. Such differences have declined steadily in recent times, and what differences remain may be in part a result of the differential financing of expensive research.

In the U.S.S.R. and nations under its influence there is much emphasis on the concepts and methods of Marxist sociology, which has only a small following elsewhere. A more important methodological issue divides basic scientific sociology from applied sociology; scholars interested in applied sociology tend to deprecate the methods and findings of the scientific sociologists as being either irrelevant or supportive of an objectionable status quo. Issues of ethics have also in recent years been raised, particularly in regard to observations and experiments in which the privacy of subjects may be felt to be invaded.

STATUS OF CONTEMPORARY SOCIOLOGY

Professional status. The Greek philosophers and the line of European philosophers in the succeeding centuries throughout Western civilization discussed much of the subject matter of sociology without thinking of it as a distinct subject. In the early 19th century all the subject matter of the social sciences was discussed under the heading of moral philosophy. Even after Auguste Comte introduced the word *sociologie* in 1838, the matter was combined with other subjects for some sixty years. Not until the universities undertook a commitment to the subject could a person make a living as a full-time sociologist. This commitment had first to be made by scholars of other fields, of which history was a principal early sponsor.

As early as 1876, at the new Johns Hopkins University, some of the content of sociology was taught in the department of history and politics. In 1889 at the University of Kansas, the word appeared in the title of the department of history and sociology. In 1890 at Colby College, a historian, Albion Small, taught a course called sociology, as did Franklin H. Giddings in the same year at Bryn Mawr College. But the first real commitment to the creation of a field of sociology took place in 1892 at the new University of Chicago, where newly arrived Albion Small asked for and received permission to create a department called sociology—the first such in the world. In the following year or two, departments in the subject were founded at Columbia, Kansas, and Michigan and very soon afterward at Yale, Brown, and many other universities. By the late 1890s nearly all of the educational institutions in the United States either had departments of sociology or offered courses in the subject.

First university department of sociology

In 1895 the *American Journal of Sociology* began publication at the University of Chicago, in time to be followed by a large number of journals in many other countries. Ten years later the *American Sociological Society* was organized, also to be followed in time by a large number of national, regional, international, and special sociological organizations. These quickly institutionalized the subject and have continuously served to guide its directions and to establish, very roughly, its boundaries. Eventually in 1949 the International Sociological Association was established under the sponsorship of UNESCO, and Louis Wirth (1897–1952) of the University of Chicago was elected its first president.

The rapid growth in numbers of full-time sociologists, along with growth of publications, allowed the content of the discipline to expand rapidly. By 1970 there were more than a dozen important sociological journals and an indefinite number of minor journals in the U.S., as well as a considerable number in other nations. Research grew throughout the 20th century at an accelerated pace, especially since the 1920s, partly as a consequence of strong financial support from foundations, government, commercial sources, and private gifts. Along with this there came a flourishing of research institutes, some affiliated with university departments and some independent. A small but increasing number of full-time sociologists gain their livelihood through full-time research independent of universities.

Similar developments have occurred in various other parts of the world, with variations resulting from special conditions in each case. In France, where Auguste Comte and later Émile Durkheim gave early impetus to sociology, there was early development in many fields of the subject. The two World Wars slowed the development, but after 1945 a strong revival of interest in sociology took place, during which the French government established a number of institutes in the social sciences at the level of institutes in the natural sciences, including several in Paris for sociological research—notably the Centre d'Études Sociologiques, the Institut National d'Études Démographiques, and the Maison des Sciences de l'Homme. These institutes receive government funds and employ many full-time sociologists, some of them among the prominent scholars in the nation. French universities have been somewhat more conservative; the Sorbonne, for example, had in 1970 only one chair officially assigned to sociology. The new University of Nanterre, however, established a department with four professorships. A rich amount of research publication has been produced in France since World War II, particularly in general sociology, theory, methodology, social psychology, industrial sociology, and the sociology of work.

German sociology had a strong base in the late 19th century and afterward, and the writings of Ferdinand Tönnies, Max Weber, Georg Simmel, and others were influential in all parts of the world. By the early 1930s, however, official Nazi hostility had impeded its development and by the time of the Second World War had destroyed it as an academic subject in Germany. Immediately after the war a new generation of scholars, aided by visiting sociologists, imported the new empirical research methods and began the development of a style of German sociology much different from the earlier theoretical and philosophical traditions. At the University of Frankfurt the Institut für Sozialforschung (social research), established by private financing before the war, was revived and has stimulated much research production. West German universities remained conservative for a time, but two newly created universities—the Free University of Berlin and the University of Constance—made sociology one of their major subjects. By 1970 most West German universities had at least one chair in sociology. National needs received special emphasis, including administrative research of use to planning, studies of unemployment, youth problems, and delinquency. A significant amount of research is also published in such fields as rural sociology, political sociology, and the family.

Sociology in Great Britain

In Great Britain, despite the early prominence of Herbert Spencer and L.T. Hobhouse, sociology was little regarded by leading universities until recent years. Before World War II Britain excelled in anthropology, especially in the study of nonwhite societies of the empire. Sociology concentrated on studies of the poor, and much of it was undertaken by persons whose affiliation was similar to that of social workers in the U.S. The major prewar sociology department, at the London School of Economics, had the objective more of social reform than scientific objective research. In the postwar period, however, a considerable revival of sociology took place; Oxford and Cambridge recognized the subject by creating positions for sociologists, and various new universities established chairs and departments. Significant work in Britain has been done in such fields as population and demography, sociology of organization, and general sociology. The Tavistock Institute of Human Relations in London has become world famous and concentrates on human relations in the family, the work group, and organizations.

A parallel growth took place in Canada, Australia, and New Zealand. Canada, with some apparent reluctance, allowed itself to be much influenced by American sociology and has recently built many new departments with sociologists trained in the U.S.

The Scandinavian countries have also to a considerable extent adopted the methods and some of the content of American sociology, and the subject has had rapid development in many of the universities and in research institutes, some of which are connected with universities. There is also a considerable amount of interchange between sociologists in these countries.

Japan has a record of much sociological activity dating back to the 1870s. The Japanese Sociological Society (Nippon Shakai Gakkai), headquartered at the University of Tokyo, was founded in 1923; and by 1960 there were about 150 universities and colleges with courses in the subject. In the early period sociology was nearly all imported; Comte and Spencer, and later Giddings and Gabriel Tarde, were their important theorists. After the Second World War there were rapid changes in sociology in Japan, with empirical research methods largely replacing the earlier philosophical style. Importations from American sociology became abundant. Popular among these were industrial sociology, educational sociology, public opinion research, and the study of mass communications.

Sociology in the Soviet Union was long held back by the perceived incompatibility of the subject with Marxist theory. In recent years, however, it has been permitted to develop, and sociological institutes and chairs of sociology are increasing. By 1970 the Soviet Sociological Association had more than a thousand members. Leading research interests have been such subjects as labour productivity, education, crime, and alcoholism. There remains an apparent tendency to avoid issues that might imply conflict with Marxist thought.

Nations under the influence of the Soviet Union have also from time to time been inhospitable to sociology, but the strong interest of younger scholars has made possible some relaxation of this opposition, and in recent years there has been considerable progress of sociology in Yugoslavia, Czechoslovakia, Hungary, and Poland, with occasional setbacks in some areas, such as Czechoslovakia.

In Israel the dominant department of sociology is at the Hebrew University in Jerusalem, where there are also several research institutes. Israeli sociology maintains continuous close contacts with American sociology, and many of the leading Israeli sociologists have had training or teaching experience in the U.S. Among the specialties in Israel are research in methodology, communication, criminology, and the collective settlements (*kibbutzim*) in which new forms of custom and social organization are observed while under development.

The passing of the Fascist regime in Italy and the relative liberalization in Spain have opened the door to sociology, and academic chairs and research institutes are gradually increasing in these countries. Of particular interest are studies of industrial efficiency and social mobility. The general conservatism of universities, however, may constitute a retarding influence for some time to come.

In Latin America objective sociology has been much resisted, partly because it has been viewed as a threat to the political and social order, but also because of meagre financial support of research and the low salary level of professors, many of whom must supplement their earnings in the practice of law, in civil service, and other occupations. In the 1960s, however, the number of full-time chairs increased and a number of research institutes, some financed by U.S. funds, were established. Political instability in some countries remains a major hindrance, and in such countries able scholars continue to be forced from their university positions from time to time.

Little by little sociology is penetrating into some of the developing nations. A number of African universities have formed departments, and the subject is gaining in importance in the Philippines, India, Indonesia, and Pakistan.

Scientific status. It is evident that sociology has not achieved triumphs comparable to those of the several older and more heavily supported sciences. A variety of interpretations have been offered to explain the difference—most frequently, that the growth of knowledge in the science of sociology is more random than cumulative. The true situation appears to be that in some parts of the discipline—such as methodology, ecology, demography, the study of social differentiation and mobility, attitude research, and the study of small-group interaction processes, public opinion, and mass communication—there has in fact taken place a slow but accelerating accumulation of organized and tested knowledge. In some other fields the expansion of the volume of literature has not appeared to have had this property. Critics have attributed the slow pace to a variety of factors—the appetite of sociologists for neologisms and jargon, a disposition for pseudo-

Criticisms of sociology

quantification, and excessive concern with imitation of the methods of natural sciences, overdependence on data from interviews, questionnaires, and informal observations. All these shortcomings can be found in contemporary sociology, but none is characteristic of all areas. In general there has been progress toward efficient terminology and methods and toward more satisfactory data, and conclusions are increasingly based on the harmonious mixture of research methods applied to varied and repeated studies, and therefore are less dependent on the strength of one particular methodological device.

Bias, in more than one direction, is sometimes presumed to be a chronic affliction of sociology. This may arise in part from the fact that the subject matter of sociology is familiar and important in the daily life of everyone, so that there exist many opportunities for the abundant variations in philosophical outlook and individual preferences to appear as irrational bias. Thus critics have expressed disapproval of the sociologists' skepticism on various matters of faith, of their amoral relativism concerning customs, of their apparent oversimplifications of some principles, and of their particular fashions in categorization and abstraction. But skepticism toward much of the content of folk knowledge is a characteristic of all science, and relativism can be interpreted as merely an avoidance of antiscientific ethnocentrism. Furthermore, abstraction, categorization, and simplification are necessary to the advancement of knowledge, and no one system satisfies everyone.

The dispute about the main purpose of sociology, whether it works to understand behaviour, or to cause social change, is a dispute found in every pursuit of scientific knowledge, and such polarization is far from absolute. Persons differ in the degree to which they regard the value of science as an intellectual understanding of the cosmos or as an instrument for immediate improvement of the human lot. Since even the "purest" scientist conceives of his work as benefiting mankind, the issue narrows to a difference in preference between an *ad hoc* attack on immediate human problems and a long-run trust that basic knowledge, gathered without reference to present urgencies, is even more valuable. Sociologists differ on this issue; in some countries there is much pressure toward early practicality of results; in others, including the United States, the larger number of scholars and the principal sociological associations have shown preference for "basic science." In very recent times, however, there has emerged a radical movement among students in various countries involving advocacy of complete commitment to action on current political and social problems.

A degree of polarization has also arisen over the proper strategy for research—whether research should take its directions from the needs of society and mankind or from the evolving theoretical corpus of sociology. In nations that allow academic freedom such disputes are usually of low intensity, because each scholar selects his research interests on any basis he prefers, including that of personal taste. In this way presumably the motivation of the investigator is maximized.

Objectivity and value-neutrality

Sociologists most interested in action express impatience at the claims of others who prefer to separate their research from personal values. Much of the dispute prevails only because the two sides argue past each other. There can be wide agreement that no human being is without personal values, that research forced to confirm a particular set of values is not good science, and that there can be scientific issues toward which a particular investigator is value-neutral. In research that is susceptible to contamination by the values of the worker, it is generally possible to minimize the damage by employing methodological devices that help to insulate the scientist from his wishes for a particular outcome—such devices as objective observational techniques and measurement methods, independent and blind analysis of results, and so forth.

Current trends. It would appear that the growth of sociology will accelerate in the visible future. Among present trends suggesting this likelihood are the increase in public appreciation of the subject, the expansion of available funds for both teaching and research, the steady reduction of sectarian opposition to inquiry into social institutions, the improvement in research methods and methods for gathering data that qualify for modern statistical treatment, and the growth of acceptance and support from scientists in other fields. There are possible factors that could inhibit such growth, such as some forms of extreme nationalism and internal conflict, but such conditions so far have impeded development only locally and temporarily.

Furthermore, it appears likely that public interest in the development of sociological knowledge will increase as a consequence of rising awareness of its promise for human safety and welfare. As the expansion of civilization, with its advanced science and technology, progressively conquers the natural hazards that afflict preliterate and preindustrial peoples and diminishes such threats as natural catastrophes, famine, and disease, a wide range of new problems emerges. These are not the menaces of an impersonal nature, but dangers that arise from imperfection in human behaviour, particularly in organized human relations. Wars have shown a tendency to become larger and ever more destructive, and the causes, though far from being understood, clearly lie, in large measure, in the complexities of social organization, in the interaction of great corporate national bodies. There appears to be little hope that politics, unaided by social science among other disciplines, could reverse this trend.

Domestic problems within nations, regions, cities, and towns appear also to become increasing sources of human troubles. There is a general rise in the severity of ethnic hostilities, and of internal conflicts between generations, political factions, and other divisions of the populations. There are also threats to human welfare from various forms of general social disorganization, reflected in the spread of pockets of poverty, crime, vice, political corruption, and family disorganization. In recent times the threats of overpopulation and potential destruction of the ecological environment have added a further reason for public alarm. Contemporary sociology obviously does not yet provide the solutions, but what prospects of human survival there are depend a great deal on the increase of the applicable knowledge of various social sciences, including sociology.

Emerging subfields. Because human behaviour observes no limits in its directions, it is possible for sociologists to extend their inquiries accordingly. The expansion of sociological interests thus has involved some penetration of adjacent traditional academic fields, such as political science, economics, anthropology, psychology, communications, speech, and to some extent even physiology and zoology. Fields within traditional sociology have also broadened their content, producing such expanded subjects as ecology and comparative sociology. Not all this extension is new, however, since much of the 19th century sociology was also very broad, especially the cosmic sociology of one worker, Lester F. Ward, who conceived sociology as the science of sciences, properly covering and organizing all knowledge.

Applications of sociology

Applications of sociology also appear to be spreading in a variety of directions, and here the possibilities seem unlimited. Sociologists aid industries in obtaining more efficient production; they help unions to increase their power; they organize rebellions of young persons, reform disorganized villages, counsel persons and families, and give or sell services to a wide variety of consumers. To what extent these applied activities will continue to spread will doubtless depend on their effectiveness relative to other means of gaining the same effects.

There is also an expansion of sociology into other than practical applications; for example, there is mathematical sociology, in which mathematical models of social behaviour are developed without systematic observations of behaviour. These efforts are not directed toward immediate human use, but may have value as bases for comparison with real behaviour and thus aid explanation of behavioral causes. A mathematical model of a completely just theoretical process of social mobility, for example, could be useful as a standard for comparing actual mobility at different times and in different nations.

Emerging methodologies. After the easiest sociological questions have been answered, the further progress of research requires ever greater effort and cost, and the proportion of discoveries by individual investigators declines as the necessity for larger teamwork research expands. This foreshadows increasing complexity of the organization of research, as has already taken place in older sciences. Large-scale research in sociology is made possible, and perhaps inevitable, by the availability of expensive computers, elaborate techniques of multivariate analysis, and the storage of information in the form of data banks and the like.

The strongest methodological emphasis in the near future is likely to be on the processes of rigorous testing of generalizations that now appear to be of strategic value in the general structure of sociological knowledge. Complete surprises in the field of human behaviour are less likely than in other sciences, since most of the possible human situations have been familiar in folk knowledge as well as in academic sociology. But the subject contains many inconsistent principles, and few of these have been put to a definitive test, partly from lack of adequate methodology and to some extent from shortage of funds and scientific manpower.

Emerging roles for sociologists. In general the principal employment of sociologists has been in educational institutions, but recently, in various countries, there has been an increasing penetration into other fields of activity. Sociologists, particularly in earlier decades, have been involved in various organized agencies devoted to social work. They have also participated in government work at various levels, from the lower bureaucratic ranks all the way to high administrative responsibility, and in the case of Thomáš Masaryk, former president of Czechoslovakia, to the highest office of a nation. In the United States sociologists have been extensively employed in the Bureau of the Census; the Bureau of the Budget; the Institutes of Health; various other sections of the Department of Health, Education, and Welfare; and the office of the president, where they have made contributions to policy.

Other directions of sociological activity include the roles of consultant, social critic, and social activists and even revolutionaries. When the activity diverges far enough from traditional academic sociology, it may cease to be regarded as sociological, but it appears likely that sociologists will continue to spread their activities over the ever-widening region of national or global concern, in the name of their science or otherwise. (R.E.L.F.)

RELATED FIELDS

Social psychology. Social psychology is the scientific study of the behaviour of individuals in their social and cultural setting. Though the term may be taken to include the social activity of laboratory animals or those in the wild, the emphasis here is on human social behaviour.

Once a relatively speculative, intuitive enterprise, social psychology has become an active form of empirical investigation, the volume of research literature having risen rapidly after about 1925. Social psychologists now have a substantial volume of observation data covering a range of topics; the evidence remains loosely coordinated, however, and the field is beset by many different theories and conceptual schemes.

Early impetus in research came from the United States, and much work in other countries has followed U.S. tradition, though independent research efforts are being made elsewhere in the world. Social psychology is being actively pursued in the United Kingdom, Canada, Australia, Germany, The Netherlands, France, Belgium, Scandinavia, Japan, and the Soviet Union. Most social psychologists are members of university departments of psychology; others are in departments of sociology or work in such applied settings as industry and government.

Scope of social psychology

Much research in social psychology has consisted of laboratory experiments on social behaviour, but this approach has been criticized in recent years as being too stultifying, artificial, and unrealistic. Much of the conceptual background of research in social psychology derives from other fields of psychology. While learning theory and psychoanalysis were once most influential, cognitive and linguistic approaches to research have become more popular; sociological contributions have also been influential.

Social psychologists are employed, or used as consultants, in setting up the social organization of businesses and psychiatric communities; some work to reduce racial conflict, to design mass communications (*e.g.,* advertising), and to advise on child rearing. They have helped in the treatment of mental patients and in the rehabilitation of convicts. Fundamental research in social psychology has been brought to the attention of the public through popular books and in the periodical press.

Research methods. Laboratory experiments, often using volunteer students as subjects, omit many features of daily social life. Such experiments also have been criticized as being subject to bias, since the experimenters themselves may influence the results. Research workers who are concerned with more realistic settings than with rigour tend to leave the laboratory to perform field studies, as do those who come from sociological traditions. Field research, however, also can be experimental, and the effectiveness of each approach may be enhanced by the use of the methods of the other.

Many colleges and universities have a social-psychology laboratory equipped with observation rooms permitting one-way vision of subjects. Sound and video recorders and other devices record ongoing social interaction; computing equipment and other paraphernalia may be employed for specific studies.

Social behaviour is understood to be the product of innate biological factors resulting from evolution and of cultural factors that have emerged in the course of history. Early writers (*e.g.,* William McDougall, a psychologist) emphasized instinctive roots of social behaviour. Later research and writing that tended to stress learning theory emphasized the influence of environmental factors in social behaviour. In the 1960s and 1970s field studies of nonhuman primates (such as baboons) drew attention to a number of similarities with human social behaviour, while research in cultural anthropology has shown that many features of human social behaviour are the same regardless of the culture studied. It is coming to be a widely accepted view that human social behaviour seems to have a biological basis and to reflect the operation of evolution as in the case of patterns of emotional expression and other nonverbal communication, the structure of language, and aspects of group behaviour.

Much research has been done on socialization (the process of learning from a culture), and learning has been found to interact with innate factors. An innate capacity for language, for example, makes it possible to learn a local language. Culture consists of patterns of behaviour and ways of organizing experience; it develops over the course of history as leaders and innovators introduce new elements, only some of which are retained. Many aspects of social behaviour can be partly accounted for in terms of their history.

Social perception. In some laboratory experiments, subjects watch stills or moving pictures, listen to tape recordings, or directly observe or interact with another person. Subjects may be asked to reveal their social perception of such persons on rating scales, to give free descriptions of them, or to respond evaluatively in other ways. Although such studies can produce results that do not correspond to those in real-life settings, they can provide useful information on the perception of personality, social role, emotions, and interpersonal attitudes or responses during ongoing social interaction.

The effect of cultural stereotypes

Research has been directed to how social perception is affected by cultural stereotypes (*e.g.,* racial prejudice), by inferences from different verbal and nonverbal cues, by the pattern of perceptual activity during social interaction, and by the general personality structure of the perceiver. The work has found practical application in the assessment of employees and of candidates for positions.

There has also been research on the ways in which perception of objects and people is affected by social factors such as culture and group membership. It has been shown, for example, how coins, colours, and other physical cues are categorized differently by people as a result of their group membership and of the categories provided by language. Other studies have shown the effect of group pressures on perception.

Interaction processes. The different verbal and nonverbal signals used in conversation have been studied, and the functions of such factors as gaze, gesture, and tone of voice are analyzed in social-interaction studies. Social interaction is thus seen to consist of closely related sequences of nonverbal signals and verbal utterances. Gaze has been found to perform several important functions. Laboratory and field studies have examined helping behaviour, imitation, friendship formation, and social interaction in psychotherapy.

Among the theoretical models developed to describe the nature of social behaviour, the stimulus–response model (in which every social act is seen as a response to the preceding act of another individual) has been generally found helpful but incomplete. Linguistic models that view social behaviour as being governed by principles analogous to the rules of a game or specifically to the grammar of a language have also attracted adherents. Others see social behaviour as a kind of motor skill that is goal-directed and modified by feedback (or learning), while other models have been based on the theory of games, which emphasizes the pursuit and exchange of rewards and has led to experiments based on laboratory games.

Small social groups. All small social groups do not function according to the same principles, and, indeed, modes of social activity vary for particular kinds of groups; *e.g.,* for families, groups of friends, work groups, and committees.

Early research on social groups

Earlier research was concerned with whether small groups did better than individuals at various tasks (*e.g.,* factory work), while later research has been directed more toward the study of interaction patterns among members of such groups. In the method known as sociometry, members nominate others (*e.g.,* as best friends) to yield measures of preference and rejection in groups. Others have studied the effects of democratic and authoritarian leadership in groups and have greatly extended this work in industrial settings. In research on how people respond to group norms (*e.g.,* of morality or of behaviour), most conformity has been found to the norms of reference groups; *i.e.,* to such groups as families or close friends that are most important for people. The emergence and functioning of informal group hierarchies, the playing of social roles (*e.g.,* leader, follower, scapegoat), and cohesiveness (the level of attraction of members to the group) have all been extensively studied. Experiments have been done on processes of group problem solving and decision making, the social conditions that produce the best results, and the tendency for groups to make risky decisions. Statistical field studies of industrial work groups have sought the conditions for greatest production effectiveness and job satisfaction.

Social organizations. Such organizations as businesses and armies have been studied by social surveys, statistical field studies, field experiments, and laboratory experiments on replicas of their social hierarchies and communication networks. Although they yield the most direct evidence, field experiments present difficulties, since the leaders and members of such organizations may effectively resist the intervention of experimenters. Clearly, efforts to try out democratic methods in a dictatorship are likely to be severely punished. Investigators can study the effects of role conflict resulting from conflicting demands (*e.g.,* those from above and below) and topics such as communication patterns in social organizations. Researchers also have studied the sources of power and how it can be used and resisted. They consider the effectiveness of different organizational structures, studying variations in size, span of control, and the amount of power delegation and consultation. In factories, social psychologists study the effects of technology and the design of alternative work-flow systems. They investigate methods of bringing about organizational change; *e.g.,* in the direction of improving the social skills of people and introducing industrial democracy.

Changes in approach since 1900

Ways of looking at working organizations have changed considerably since 1900. Classical organization theory was criticized for its emphasis on social hierarchy, economic motivation, division of labour, and rigid and impersonal social relations. Later investigators emphasized the importance of flexibly organized groups, leadership skills, and job satisfaction based on less tangible rewards than salary alone. There has been a rather uneasy balance in the industrial social psychologist's concern with production and concern with people.

Personality. It is evident that there are individual differences in social behaviour; thus, people traditionally have been distinguished in terms of such personality traits as extroversion or dominance (see PERSONALITY). Some personality tests are used to predict how an individual is likely to behave in laboratory discussion groups, but usually the predictive efficiency is very small. Whether or not an individual becomes a leader of a group, for example, is found to depend very little on what such personality tests measure and more on his skills in handling the group task compared with the skills of others. Indeed, the same person may be a leader in some groups and a follower in others. Similar considerations apply to other aspects of social behaviour, such as conformity, persuasibility, and dependency. Although people usually perceive others as being consistent in exhibiting personality traits, the evidence indicates that each individual may behave very differently, depending on the social circumstances.

Socialization. The process by which personality is formed as the result of social influences is called socialization. Early research methods employed case studies of individuals and of individual societies (*e.g.,* primitive tribes). Later research has made statistical comparisons of numbers of persons or of different societies; differences in child-rearing methods from one society to another, for example, have been shown to be related to the subsequent behaviour of the infants when they become adults. Such statistical approaches are limited, since they fail to discern whether both the personality of the child and the child-rearing methods used by the parents are the result of inherited factors or whether the parents are affected by the behaviour of their children.

Problems in the process of socialization that have been studied by experimental methods include the analysis of mother–child interaction in infancy; the effects of parental patterns of behaviour on the development of intelligence, moral behaviour, mental health, delinquency, self-image, and other aspects of the personality of the child; the effects of birth order (*e.g.,* being the first-born or second-born child) on the individual; and changes of personality during adolescence. Investigators have also studied the origins and functioning of achievement motivation and other social drives (*e.g.,* as measured with personality tests).

Several theories have stimulated research into socialization; Freudian theory led to some of the earliest studies on such activities as oral and anal behaviour (*e.g.,* the effect of the toilet training of children on obsessional and other "anal" behaviour). Learning theory led to the study of the effects of rewards and punishments on simple social behaviour and was extended to more complex processes such as imitation and morality (*e.g.,* the analysis of conscience).

The self. Such concepts as self-esteem, self-image, and ego-involvement have been regarded by some social psychologists as useful, while others have regarded them as superfluous. There is a considerable amount of research on such topics as embarrassment and behaviour in front of audiences, in which self-image and self-esteem have been assessed by various self-rating methods. The origin of awareness of self has been studied in relation to the reactions of others and to the child's comparisons of himself with other children. Particular attention has been paid to the so-called identity crisis that is observed at various stages of life (*e.g.,* in adolescence) as the person struggles to discern the social role that best fits his self-concept.

Factors in the origin of predispositions of individuals

Attitudes and beliefs. Research into the origins, dynamics, and changes of attitudes and beliefs has been carried out by laboratory experiments (studying relatively minor effects), by social surveys and other statistical field studies, by psychometric studies, and occasionally by field experiments. The origins of these socially important predispositions have been sought in the study of parental attitudes, group norms, social influence and propaganda, and in various aspects of personality. The influence of personality has been studied by correlating measured attitudes with individual personality traits and by clinical studies of cognitive and motivational processes; so-called authoritarian behaviour, for example, has been found to be deeply embedded in the personality of the individual. Early research based on statistical analyses of social attitudes revealed correlations with such factors as radicalism–conservatism. Later research on consistency provided extensive laboratory evidence of consistency but little evidence of it in actual political behaviour (*e.g.,* in attitudes on different political issues).

Research on attitude change has studied the effects of the mass media, the optimum design of persuasive messages, the effects of motivational arousal, and the role of opinion leaders (*e.g.,* teachers and ministers). Research has been carried out into the origins, functioning, and change of particular attitudes (*e.g.,* racial, international, political, and religious), each of which is affected by special factors. Attitudes toward racial minority groups, for example, are affected by social conditions, such as the local housing, employment, and the political situation; political attitudes are affected by social class and age; and religious attitudes and beliefs strongly reflect such factors as inner personality conflict.

Various specialties in social psychology. Many social psychologists are concerned with such aspects of public opinion (social survey) research as the design of standardized interviews and questionnaires. Forms of questions have been devised to compensate for errors that arise from the efforts to respond in a socially approved manner; some are designed to detect lying. Mass communications have been devised on the basis of research into persuasion. Use is also still made of Freudian symbolism and theory.

Research into the causes of mental disorders has shown the importance of social factors in the family and elsewhere. Mental patients often show deficiencies in social performance that may be the cause of other symptoms. Many social psychologists hold that social factors may also apply to such disorders as schizophrenia, which also seem to have hereditary and chemical bases. There has been a corresponding growth in the use of various kinds of social therapy in psychiatry (*e.g.,* group therapy, therapeutic communities, and social-skills training).

Considerable research has been devoted to industrial productivity, absenteeism, labour turnover, accidents, and job satisfaction. Factors that have been found to be important include the style of supervision and management, the size and composition of working groups, the technology and the work-flow systems, the span of control, and other features of the organizational structure. Research results point strongly toward the advantages of a less rigid hierarchical structure of authority, with more delegation of authority and consultation, training in supervisory skills, small and cooperative work teams, and interesting and varied work.

A major application of research in social interaction and group behaviour is in training in social skills, as in the T-groups, or sensitivity training, noted above. Role playing with video-tape playback and training in the imitation of other persons who serve as behavioral models are used in teaching people new skills. Actual training on the job has the advantage that there is no gap between the training and the work itself. All of these methods have been shown to be effective, depending on the job and the teacher. Social-skills training has been given successfully to industrial managers and supervisors, social workers and clergymen, interviewers, public speakers, mental patients, and juvenile delinquents.

A great deal of research has been done on factors underlying racial prejudice, but the understanding thus obtained has not had much effect upon the social problems involved. Similarly, the causes of delinquency and crime have been extensively studied, but it is not feasible to manipulate the factors influencing crime, such as genetic factors, methods of upbringing, and inequalities of opportunity. Social psychology has made some contribution to education; sociometry is quite widely practiced as a means of grouping children, and evidence is growing about the optimum styles of teacher behaviour. (M.Ar./Ed.)

Criminology. Criminology is the scientific study of the nonlegal aspects of crime (including juvenile delinquency). In its wider sense, embracing penology, it is thus the study of the causation, correction, and prevention of crime—seen from the viewpoints of such diverse disciplines as ethics, anthropology, biology, ethology (the study of character), psychology and psychiatry, sociology, and statistics. Whereas the traditional legal approach to crime focusses on the action of crime and the protection of society, criminology focusses on the person of the criminal and the essential interests of the individuals of whom society consists. Whereas criminal law has been a relatively conservative force, often slow to change even where change has seemed imperative, criminology as a part of the developing social sciences of the past hundred years has been a revolutionary force—its object being not to replace the legal system in dealing with crime and punishment but to supplement it, making it less rigid and more sympathetic to approaches wider than strictly legal ones.

Criminology and criminal law compared

Without denying the value of "pure research," one must point to criminology and particularly penology as primarily practical subjects or "applied" disciplines. This practical value of criminological research can make itself felt in several ways. Its accumulated findings can give judges, prosecutors, lawyers, probation officers, and prison officials better understanding of crime and criminals, leading hopefully to more effective and humane sentencing and methods of treatment. Criminological research and knowledge can be equally at the disposal of legislators and administrators to assist in their task of reforming the law and improving penal and reformatory institutions. Essentially this purveyance of information represents a neutral role for criminologists; they garner the facts, and the various governmental officials decide for themselves what kind of practical conclusions to draw from the facts. Increasingly, however, some criminologists—like their counterparts in such fields as the atomic sciences—are demanding that scientists fully shoulder the moral and political responsibilities for their discoveries and for the use made of them instead of leaving vital decisions entirely to their governments. Thus some criminologists, for instance, insist upon actively campaigning against capital punishment, given the facts as they see them. Opponents of this activist role, on the other hand, contend that penological arguments are not sufficient but must be weighed along with political, social, religious, and moral arguments and that this all-round consideration should be left to responsible political bodies. The view does not deny the right of criminologists to express their opinions as ordinary citizens and voters; it does contend, nevertheless, that a government of officials responsive to the popular will, however fallible it may be, is less dangerous than a "government by experts."

Another question involving the scope and functions of criminology is whether or not it should extend to the study of crime detection, involving such measures as photography, toxicology, fingerprint study, and the like. In several countries, notably Austria and Belgium, and at the school of criminology of the University of California at Berkeley, this so-called criminalistics has long been an important branch of criminological teaching and research, and the distinguished *Journal of Criminal Law, Criminology, and Police Science* (U.S.) devotes much of its space to criminal investigation. Actually, the only reason for excluding it from criminology is perhaps the expense of staff and equipment, which can be better borne by police colleges and similar specialized institutions. On the other hand, in recent decades criminology has undergone an important and perfectly legitimate extension of its territory by devoting much attention to so-called victimology—the study of the victim of crime, his relations to the criminal, and his role as a potential causal factor in crime.

The role of criminalistics and victimology

Although the exclusion of criminalistics makes it easier to locate criminology on the map of scientific studies, its origin in, its close relations to, and its partial dependence on so many other disciplines result in considerable diversity and confusion regarding its proper place in the academic curriculum. Universities in continental Europe, when they do not ignore criminology altogether, tend to treat it as part of legal education; even where its principal teachers are not lawyers. In Great Britain the only existing Institute of Criminology is part of the law faculty of Cambridge University; in other schools criminological research and teaching are usually divided between departments of sociology or social administration, law faculties, and institutes of psychiatry. In South America the anthropological and medical elements predominate, and in the United States, criminology, with a few notable exceptions, forms an established section of departments of sociology.

Given this situation in which criminology is submerged in other fields, it is not surprising that most teachers and researchers in criminology regard themselves first as sociologists, psychologists, lawyers, or whatever and only secondarily as criminologists. Their education contributes to this status; although a number may have pursued some criminological studies in their undergraduate years, criminology is largely a postgraduate discipline, at least in terms of major concentration for students.

This floating character of criminology weakens its position and tends to lend doubt to its claim to scientific status. Nevertheless, other disciplines—such as psychology, psychiatry, history, sociology, and social anthropology—have gone through similar birth pangs and, even after having achieved more or less assured positions, still face challenges to their claim to being scientific disciplines. The answer lies perhaps in historian H.R. Trevor-Roper's remark, "there are sciences and sciences." If the results of research can be viewed relatively, it is possible to perceive science in the criminologist's systematic application of sound research methods and his development of a body of facts from which he interprets general trends on a subject of real importance to mankind.

Historical development. The origins of criminology are generally dated from the late 18th century, when those imbued with a spirit of humanitarianism began questioning the cruelty, arbitrariness, and inefficiency of criminal justice and prison systems. From this period arose the so-called classical school of criminology, composed of such reformers as the Italian Marchese di Beccaria and the Englishmen Sir Samuel Romilly, John Howard, and Jeremy Bentham, all of whom may be said to have sought penological and legal reform rather than criminological knowledge per se—that is, knowledge about crime and criminals. Their principal aims were to mitigate legal penalties and subject judges to the principle of *nulla poena sine lege* or "due process of law" and also to reduce the application of capital punishment and humanize penal institutions. In all this they were moderately successful, but in their desire to make criminal justice "just," they tried to construct rather abstract and artificial equations between crimes and penalties, thereby forgetting the personal characteristics and needs of the individual criminal. Moreover, the object of punishment was seen as being primarily retribution, with deterrence occupying second place, and reformation lagging far behind.

Positivist school

By the second half of the 19th century these deficiencies, together with the influential teachings of the French sociologist Auguste Comte, had prepared the ground for the positivist school, which sought to bring a scientific neutrality into criminological studies. Instead of assuming a moral stance that focussed on measuring the criminal's "guilt" and "responsibility," the positivists attempted a morally neutral and social interpretation of crime and its treatment. Their leading figure, Cesare Lombroso (1836–1909), professor of psychiatry and anthropology at the University of Turin, sought through firsthand observation and measurement of prison inmates to determine the characteristics of criminal types. Some of his investigations led him into anthropometric interpretations—for example, his oft-criticized deduction of the "born criminal" with cranial, skeletal, and neurological malformations—but largely he and other positivists helped to introduce the ideas that crime has multiple causes and that most criminals are not born criminal but are shaped by their environmental upbringing and associations. With the positivists, therefore, the emphases in criminology had turned to experimental case studies and to preventive and rehabilitative measures. Without the upheaval caused by the positivists not only criminological research in the modern sense but also the present-day alternatives to capital punishment and old-fashioned imprisonment such as probation, suspended sentence, fines, and parole, inadequate as some of them are, would have been unthinkable.

Today, nevertheless, the feeling is widespread that the battle of ideas fought by the classical and positivist schools has not yet produced the secure foundations on which the criminology and the penal systems of the future can be built. Thus a third school, the postwar movement of "social defense," also originating in Italy, has tried to combine their best features and eliminate their excesses. This school disapproves of any rigid typology of criminals and stresses the uniqueness of human personality; it refuses the "scientism" of the positivists in favour of a strong belief in moral values—most importantly in balancing the rights of criminals and the rights of society. The school still speaks with too many voices, however, to be conclusively labelled.

Independent of these debates between schools, however, are the advances accomplished by such great figures as the statisticians Adolphe Quetelet (1796–1874) and André Michel Guerry (1802–66), the sociologists Gabriel Tarde (1843–1904) and Émile Durkheim (1858–1917), and of course Sigmund Freud, all of whom introduced a wealth of new ideas into the old problems of the social and individual characteristics of human, including criminal, behaviour.

Forming and testing a hypothesis

Modern trends. The objectives of criminological research are sometimes said to be threefold: descriptive, causal, and normative. The descriptive aspect consists of the collection of relevant and reliable facts, together with their interpretation. The collection does not begin as a random and meaningless running after whatever phenomena happen to rouse the researcher's interest. Rather it is preceded by some hypothesis or "hunch," an assumption about what the researcher expects to find. The hypothesis "organizes" his inquiry. As the collection of facts proceeds, he may find either that his hypothesis is correct or that it requires revision or abandonment and thus the development of a new hypothesis to guide further research. The history of criminology reflects this perennial revision and renewal of inquiry, this continuous process of abandoning seemingly well-established theories in favour of new ones. Lombroso's theory of the "born criminal," the theory that all crime (or at least all economic crime) is due to poverty, and the theory that all juvenile delinquents come from broken homes had all to be drastically revised.

The causal aspect involves relating the effects of one body of facts on another. Although regarded nowadays with some suspicion or indifference (even in the natural sciences), the search for causes is not being dispensed with altogether. So long as one does not jump to causal conclusions when arriving at statistical correlations or does not pressure "facts" into some proof of a popular theory, theories of causation can be useful in planning for the alleviation of crime and criminality.

The normative aspect, however, is more decidedly suspect. Research aimed at formulating so-called laws governing criminological phenomena has been thus far futile and does not look promising. What is sometimes regarded as a "law" has in reality been a mere trend. To diagnose statistical trends may be useful so long as their possible ephemeral nature is recognized, but trends are not laws, and there is thus but little scope for the normative approach in criminology.

Cross-disciplinary and cross-cultural approaches

Criminology, as suggested earlier, is cross-disciplinary and indeed draws methods or techniques from both the natural and the social sciences and must continually take heed of developments in other fields. It also depends increasingly on cross-cultural approaches; there have been recent statistical (though admittedly controversial) comparisons of "delinquent generations" in England, New Zealand, Denmark, and Poland; various studies of the sociological and statistical aspects of homicide; and studies of the ecological significance of "criminal areas" and of the possibility of predicting criminality.

In common with other disciplines, criminology must face such distinctions as between pure and applied research and between statistical and intuitive ways of thinking, but what is almost unique to criminological research is its intense involvement with society and its difficulty in achieving "detachment." Not only do society's biases toward crime and punishment influence a criminologist's choice and execution of research but also he is dependent on the willing cooperation of governmental departments and other public authorities to secure essential raw material or data. There are only a few limited areas—such as adolescent delinquency and gang activities—in which research can be pursued privately without resort to official help.

Seen in the light of the previous remarks on the history of criminology, the development of criminological research can be divided into three stages: the prescientific, lacking any theoretical basis and largely identifiable with the work of the classical school; the semiscientific, possessing theories and hypotheses but without scientifically sound techniques and largely characteristic of the efforts of the positivist school; and the scientific, hopefully trying to repair earlier deficiencies and developing or improving the techniques described below.

Statistics. Often serving as the initial step in any research and regarded by some researchers, perhaps incorrectly, as the one and only reliable technique, the collection and interpretation of statistics for social and criminological purposes began in Europe early in the 19th century. The reputed "father" of this criminological method is the Belgian astronomer Adolphe Quetelet, who is perhaps best remembered for his famous "law," developed from the French and Belgian statistics, which showed that crime in any given country remains fairly constant over the long term (short-term fluctuations being insignificant). When he qualified himself, however, by saying that the volume and kind of crime were constant only so long as society's social, economic, and political conditions remained unchanged, he deprived the "law" of much of its significance for a rapidly changing modern world.

The manner and extent of data collection today differ considerably from country to country or, in federal unions like the United States, even from state to state or province to province. They differ in how often data are collected and published, in what items are given importance, in the choice between complete listings and sample surveys, in the ratio between governmental and private research, and so forth. Such far-reaching differences, together with the differences in law and its administration and in popular views and habits, have made it so far impossible to devise a meaningful system of international criminal statistics. Generally, however, there is increasingly less tendency to collect any and all data, regardless of reliability or practical value, and to concentrate on limited, reliable data involving matters of agreed upon importance.

Comparison of police and court statistics

A noteworthy distinction to be made is between police statistics and court statistics. Police data are nearer to the event but perhaps less reliable, and they usually describe the crimes only, not the criminals. The data from the courts, being based on convictions, do deal with the persons involved but include only the material brought forward by the prosecution and the defense. All criminal statistics indeed depend entirely on human factors such as the willingness of private individuals and officials to prosecute, on the popularity or unpopularity of the criminal laws at issue, and so forth. The figures also usually fail to rate very clearly the gravity of individual cases; except for such broad categories as "petty" and "grand," theft is theft regardless of the value of the objects stolen. Only recently have more detailed categories been attempted for criminological research.

Case studies. Also called the individual "case history," the case study concentrates on the career or life of one individual or group of individuals and is the method used primarily, though not exclusively, by psychologists, psychiatrists, and psychoanalysts. If well done, such histories can give deep insights into the personalities and motives of criminals, but the method does have shortcomings. Although the volume of case histories has grown large, their reliability is sometimes suspect—partly because of a criminal's natural reluctance to expose himself completely and partly because of the nature of the publication of case histories. Their publication is comparatively rare; professional ethics often forbid the exposure of details given confidentially, and those studies actually published may be too few to be typical and may even on occasion be designedly selective because of an investigator's wish to prove a theory.

Closely related to case studies are autobiographies and other books written by ex-prisoners, but in spite of their considerable human and scientific interest, they do suffer from even greater disadvantages, chiefly questionable objectivity. Sociologists have also contributed important studies of individuals in their social environments.

Typologies. The typological method involves classifying offenses, criminals, criminal associations, criminal areas, or whatever according to some criteria of relatedness or similarity. Thus there have been attempts to dichotomize criminals as either "normal" or "abnormal," "habitual" or "professional," or to form a continuum of criminals from the "insane" at one extreme through various career criminals, petty offenders, and white-collar criminals to "organized" or "professional" criminals at the other extreme. The typological method is less impersonal and heterogeneous than the statistical method and less individual or specific than the case study. Developed mainly in Germany and Austria and more recently in the United States, the method has been disputed; psychiatrists and psychoanalysts have especially questioned its value, primarily because it attempts to reduce complex phenomena to simple terms and tends to ignore important individual differences. Nevertheless, employed with restraint, the method is indispensable as a bridge between the two extremes, and it is in fact often used unknowingly by both statisticians and case students.

Experimental methods. A controlled experiment involves taking two closely related situations or groups, subjecting one of them to a specific change and comparing the subsequent characteristics of both. In the past, so-called experiments by judicial, penal, and reformatory institutions were not really controlled or even experimental in the scientific sense, for public agencies, at least in theory, are bound by the idea of justice to give equal treatment to equals, not one kind of treatment to one group and another kind to another group. Thus, generally speaking, most controlled experiments must be left to universities and other private bodies, and indeed the need for strict control and variable treatment has been recently accepted in such researches as Harvard University's Cambridge-Somerville Youth Study, which sought the effects of counselling on "pre-delinquent" boys.

Problems of crime prediction

Prediction studies. Criminological prediction—not unlike actuarial prediction used by insurance companies—is intended to forecast, usually in percentages, the future conduct of persons under certain conditions. Based on statistics or case histories or both, the predictions attempt to indicate probabilities—how any specific individuals or groups are likely to be affected by certain conditions or treatments. Thus, for example, various categories of criminals are listed as likely to be recidivistic.

The techniques involved in constructing prediction tables are too complicated to be discussed in brief; they have been developed and refined in the past 40 years mainly by Sheldon and Eleanor Glueck of Harvard University and also by several other authors in various countries. Statistical prediction by itself can never be conclusive; it must be subjected to rigorous validation for any individual or group, and even then it can merely show certain probabilities, which should be used in penal decisions only with the greatest caution and along with the lessons of experience derived from other sources. Nevertheless, the method can be valuable in supplementing the inevitably limited personal experience of judges and administrators, and indeed in recent decades prediction research has probably nowhere in the social sciences been more popular and urgent than in criminology.

Action research. Action research, which is often contrasted with experimental research, consists of drawing upon the observations of field workers and other persons directly involved with delinquents, potential delinquents, or prisoners. Thus, for example, have social workers attempted to help slum children and adolescents with their problems and at the same time studied their delinquent behaviour, related it to their environment, and evaluated the results of the youth clubs or other services offered. The chief values of action research are that it aims at practical results through collaboration with fieldworkers, tries to build a bridge between theoretical and practical work, and may well dispense with formal hypotheses and simply aim at preventive tactics. Its best known and perhaps most successful example has so far been Clifford Shaw's Chicago Area Project, which, in close cooperation with the famous ecological studies of the University of Chicago, has tried to enlist suitable local people to deal with the social problems of their area.

Sociological research. Sociological research involves various methods—general surveys and personal interviews, as well as statistical, case-study, typological, experimental, and predictive techniques—and thus the purpose of classifying sociological research separately is chiefly to signal its focus or fields of interest. Mainly, criminology derives benefits from three fields of sociological study: (1) social institutions, involving such things as different conceptions of, or attitudes toward, property held by various societies or groups or the different effects of mass media on crime; (2) social groups, involving such things as the influence of juvenile gangs or criminal subcultures on individual criminal behaviour or the influence of prejudice on certain racial, national, or religious minorities; and (3) ecology, involving the study of different geographical areas and their rates and kinds of crime. Clifford Shaw's ecological studies of Chicago have been especially revealing in analyzing urban areas that either "breed" crime or "attract" crime. (H.M./Ed.)

Economics

No one has ever succeeded in neatly defining the scope of economics. Economists used to say, with Alfred Marshall, the great English economist, that economics is "a study of mankind in the ordinary business of life; it examines that part of individual and social action which is most closely connected with the attainment and with the use of the material requisites of wellbeing"—ignoring the fact that sociologists, psychologists, and anthropologists frequently study exactly the same phenomena. Another English economist, Lionel Robbins, has more recently defined economics as "the science which studies human behaviour as a relationship between (given) ends and scarce means which have alternative uses." This definition—that economics is the science of economizing—captures one of the striking characteristics of the economist's way of thinking but leaves out the macroeconomic approach to the subject, which is concerned with the economy as a whole.

What economists do

Difficult as it may be to define economics, it is not difficult to indicate the sort of questions that economists are concerned with. Among other things, they seek to analyze the forces determining prices—not only the prices of goods and services but the prices of the resources used to produce them. This means discovering what it is that governs the way in which men, machines, and land are combined in production and that determines how buyers and sellers are brought together in a functioning market. Prices of various things must be interrelated; how does such a "price system" or "market mechanism" hang together, and what are the conditions necessary for its survival?

These are questions in what is called "microeconomics," the part of economics that deals with the behaviour of such individuals as consumers, business firms, traders, and farmers. The other major branch of economics is "macroeconomics," in which the focus of attention is on aggregates: the level of income in the whole economy, the volume of total employment, the flow of total investment, and so forth. Here the economist is concerned with the forces determining the income of a nation or the level of total investment; he seeks to learn why full employment is so rarely attained and what public policies should be followed to achieve higher employment or more stability.

But these still do not exhaust the range of problems that economists consider. There is also the important field of "development economics," which examines the attitudes and institutions supporting economic activity as well as the process of development itself. The economist is concerned with the factors responsible for self-sustaining economic growth and with the extent to which these factors can be manipulated by public policy.

Cutting across these three major divisions in economics are the specialized fields of public finance, money and banking, international trade, labour economics, agricultural economics, industrial organization, and others. Economists may be asked to assess the effects of governmental measures such as taxes, minimum-wage laws, rent controls, tariffs, changes in interest rates, changes in the government budget, and so on.

The size of the profession

In the 19th century, economics was the hobby of gentlemen of leisure and the vocation of a few academicians; economists wrote about economic policy but were rarely consulted by legislatures before decisions were made. Today, there is hardly a government, international agency, or large corporation that does not have its resident economist. According to an estimate of the National Science Foundation (U.S.), for instance, there were 11,000 economists in the United States in 1966. Clearly, much depends on how one defines the job of an economist: the list of the National Science Foundation is confined to persons whose chief competence is in any one of the recognized economic specialities. Of the 11,000 professional economists, about 4,500 were employed as teachers of economics; the rest worked in various research or advisory capacities, either for themselves, for industry, or for government. This leaves out of account many others employed in accounting, commerce, marketing, and business administration; they may think of themselves as economists, but their professional expertise falls within other fields. There are perhaps another 10,000 economists in the rest of the world—their numbers have never been counted. It would be reasonable to estimate the total number of professional economists in the world in 1970 at 20,000, a number that was apparently growing at about 5 percent per year. There were about 75 English-language journals in economics and another 25 in various foreign languages, with new ones appearing every year. This implies the publication of about 1,500 scientific papers per year, not to mention the 700 new books on economics published every year. This is indeed "the age of economists," and the demand for their services seems insatiable.

HISTORICAL DEVELOPMENT OF ECONOMICS

The work of Adam Smith

The effective birth of economics as a separate discipline may be traced to the year 1776, when the Scottish philosopher Adam Smith published *An Inquiry into the Nature and Causes of the Wealth of Nations.* There was, of course, economics before Adam Smith: the Greeks made significant contributions and so did the medieval scholastics; from the 15th to the 18th century, an enormous pamphlet literature appeared that developed the implications of economic nationalism, a body of thought now known as "mercantilism"; for a brief period in the 18th century the French "physiocrats" developed a fairly sophisticated economic model; and several other 18th-century figures can compete with Smith for the title of "first economist." Nevertheless, Adam Smith wrote the first full-scale treatise on economics and, by his magisterial influence, founded what later generations were to call the "English School of Classical Political Economy."

Analysis of the market. The *Wealth of Nations,* as its title suggests, is essentially a book about economic development and about policies that promote or hinder development. In its practical aspects it is an attack on the protectionist doctrines of the mercantilists and a brief for free trade. But in the course of attacking "false doctrines of political economy," Adam Smith was led to analyze the workings of a free-enterprise system as a governor of human activity. In a competitive market each individual, being one among many, can exert only a negligible influence on prices; each must take prices as they come and is free only to vary the quantities bought and sold at given prices; yet the sum of all individuals' separate actions determines prices. The "invisible hand" of the market, as Adam Smith was fond of saying, assures a social result that is independent of individual intentions and thus creates the possibility of an objective science of economic behaviour. Adam Smith believed that he had found, in the competitive market, an instrument capable of converting "private vices" (like selfishness) into "public virtues" (like maximum production). But this is only true if the competitive system is embedded in an appropriate legal and institutional framework, an insight that Adam Smith developed at length but that was largely forgotten by later generations. Within this great tome on the theme of rich and poor nations was contained a simple theory of value (or prices), a crude theory of distribution, an even cruder theory of international trade, and a primitive theory of money; but with all their imperfections, these were the building blocks of classical and modern economics. The book's very fecundity gave it strength because it left so much for disciples to tidy up.

The later classical economists

Construction of a system. David Ricardo's *Principles of Political Economy and Taxation* (1817) was, in one sense, simply a critical commentary on the *Wealth of Nations;* in another sense, it gave an entirely new twist to the developing science of political economy. Ricardo invented the concept of the "economic model," a tightly knit logical apparatus consisting of a few strategic variables, an apparatus that was capable of yielding, after a bit of manipulation, results of enormous practical import. At the heart of the Ricardian system is the notion that economic growth must sooner or later be arrested, owing to the rising cost of growing food on a limited land area. An essential ingredient of this argument is the Malthusian principle—enunciated in Thomas Malthus' *Essay on Population* (1798)—that population tends to increase up to the limits set by the existing supply of food, thus holding down wages. As the labour force increases, extra food to feed the extra mouths can be produced only by extending cultivation to less fertile soil or by applying capital and labour to land already under cultivation (with diminishing results because of the so-called law of diminishing returns). Although wages are held down, profits do not rise proportionately because tenant farmers outbid each other for superior land. The chief beneficiaries of economic progress, therefore, are the landowners.

Since the root of the trouble, according to Ricardo, is the declining yield of wheat per unit of land, one obvious solution is to import cheap wheat from other countries. Eager to show that Britain would benefit from specializing in manufactured goods and exporting them in return for food, Ricardo hit upon the "law of comparative costs" as proof. He assumed that, within countries, labour and capital are free to move in search of the highest returns; between countries, however, they are not. In these circumstances, Ricardo showed, the benefits of trade are determined by a comparison of costs *within* each country, rather than by a comparison of costs *between* countries. It pays a country to specialize in the production of those goods that it can produce *relatively* more efficiently and to import everything else; although India may be able to produce everything more efficiently than England, India is nevertheless well advised to concentrate its resources on textiles, in which its efficiency is relatively greater, and to import British capital goods. The beauty of the argument is that if all countries take full advantage of the territorial division of labour, total world output is certain to be larger than it will be if some or all countries try to become self-sufficient. Ricardo's law became the fountainhead of 19th-century free-trade doctrine, which would have been enough, if he had said nothing else, to give him a place in the economists' pantheon.

Ricardo's influence

The influence of Ricardo's treatise was felt almost as soon as it was published, and for over half a century the Ricardian system dominated economic thinking in Britain. In 1848 John Stuart Mill's restatement of Ricardo's thought in his *Principles of Political Economy* brought it new authority. After 1870, however, most economists turned their backs on the range of problems that had concerned Ricardo and began to re-examine the foundations of the theory of value; that is, they became interested in the theory of why goods exchange at particular prices, so that for a while they devoted almost all of their efforts to the problem of resource allocation under conditions of perfect competition.

The three generalizations of Marxism

Marxism. A few words must first be said, however, about the last of the classical economists, Karl Marx. The first volume of *Das Kapital* appeared in 1867; the second and third after his death, in 1883 and 1894. For a generation, therefore, the competitive market theorists jostled with the followers of Marx. By 1900 the intellectual battle was over, and thereafter professional economists largely lost interest in Marx. Despite the Russian Revolution, despite what amounts to official endorsement of Marxism in one-third of the world, and despite the lingering influence of Marx's ideas, Marxian economics has been moribund ever since Marx's death in 1881. If Marx may be called "the last of the classical economists," it is because to a large extent he found his economics not in the real world but in the teachings of Smith and Ricardo. They had espoused a "labour theory of value," which holds that products exchange roughly in proportion to the labour costs incurred in producing them. Marx worked out all the logical implications of this theory and added to it "the theory of surplus value," which rests on the axiom that human labour alone creates all value and hence constitutes the sole source of profits. It is an axiom in the sense that it cannot be established in terms of the theory itself: it must be imported from without. To say that an economist is a Marxian economist is in effect to say that he shares the value judgment that it is socially undesirable for some people in the community to derive their income merely from the ownership of property. Since few professional economists in the 19th century accepted this ethical postulate and most were indeed inclined to find some social justification for the existence of private property and the income derived from it, Marxian economics fell on deaf ears. The Marxian system, moreover, culminated in three great generalizations: the tendency of the rate of profit to fall, the growing impoverishment of the working class, and the increasing severity of business cycles, of which the first is the linchpin of all the others. Marx's exposition of the "law of the declining rate of profit" is invalid; with it all of Marx's other predictions fall to the ground. In addition, Marxian economics had little to say on some of the practical problems that are the bread and butter of economists in any society. This is enough to suggest why Marxian economics failed to make many converts among academic economists. Marxists will reply that the reason is simply that academic economists have always been "lackeys of the capitalist class." Perhaps so, but the fact remains that Marx has had virtually no effect on modern economic thought.

The marginalists. The marginal revolution was essentially the work of three men: Stanley Jevons, an Englishman; Carl Menger, an Austrian; and Léon Walras, a Frenchman. Their contribution was the replacement of the labour theory of value by the marginal utility theory of value; their explanation of prices began with the behaviour of consumers in choosing among increments of goods and services (see ECONOMIC THEORY). The idea of emphasizing the marginal or last unit proved in the long run to be more significant than the introduction of utility. It was the consistent application of marginalism that marks the true dividing line between classical theory and modern economics. The classical political economists saw the economic problem as that of predicting the effects of changes in the quantity of capital and labour on the rate of growth of national output. The marginal approach, however, focussed on the conditions under which these factors tend to be allocated with optimal results among competing uses—optimal in the sense of maximizing consumers' satisfactions.

Throughout the last three decades of the 19th century, the English, Austrian, and French contributors to the marginal revolution largely went their own way. The Austrian school dwelt on the importance of utility as the determinant of value and vehemently attacked the classical economists as completely outmoded. A brilliant second-generation Austrian economist, Eugen von Böhm-Bawerk, applied the new ideas to the determination of the rate of interest, putting his stamp for all time on capital theory (see ECONOMIC THEORY). The English school, led by Alfred Marshall, sought a reconciliation with the doctrines of the classical writers. The classical authors, Marshall argued, concentrated their efforts on the supply side in the market; marginal utility theory was concerned with the demand side, but prices are determined by both supply and demand, just as a pair of scissors cuts with both blades. Marshall, seeking to be practical, applied his "partial equilibrium analysis" to particular markets and industries.

The leading French marginalist was Léon Walras, who carried the approach furthest by describing the economic system in general mathematical terms. For each product there is a "demand function" that expresses the quantities of the product that consumers demand as depending on its price, the prices of other related goods, the consumers' incomes, and their tastes. For each product there is also a "supply function" that expresses the quantities producers will supply as depending on their costs of production, the prices of productive services, and the level of technical knowledge. In the market, for each product there is a point of "equilibrium"—analogous to the equilibrium of forces in classical mechanics—at which a single price will satisfy both consumers and producers. It is not difficult to analyze the conditions under which equilibrium is possible for a single product. But equilibrium in one market depends on what happens in other markets (a "market" in this sense being not a place or location but a complex of transactions involving a single good), and this is true of every market. There are literally millions of markets in a modern economy, and therefore "general equilibrium" involves the simultaneous determination of partial equilibria in all markets. Walras' efforts to describe the economy in this way led the historian of economic thought Joseph Schumpeter to call his work "the Magna Carta of economics." Walrasian economics is undeniably abstract, but it provides an analytical framework for incorporating all of the elements of a complete theory of the economic system. It is not too much to say that nearly the whole of modern economics is Walrasian economics. Certainly, modern theories of money, of employment, of international trade, and of economic growth are all Walrasian general equilibrium theories in a simplified form.

The neoclassical era

The years between the publication of Marshall's *Principles of Economics* (1890) and the Great Crash in 1929 may be described as years of reconciliation, consolidation, and refinement. The three national schools gradually coalesced into a single mainstream. The theory of utility was reduced to an axiomatic system that could be applied to the analysis of consumer behaviour under various circumstances, such as a change in income or price. The concept of marginalism in consumption led eventually to the idea of marginal productivity in production, and with it came a new theory of distribution in which wages, profits, interest, and rent were all shown to depend on the "marginal value product" of a factor. Marshall's concept of "external economies and diseconomies" was developed by his leading pupil, Arthur Pigou, into a far-reaching distinction between private costs and social costs, thus laying the basis of welfare theory as a separate branch of economic inquiry. There was a gradual development of monetary theory, which explains how the level of all prices is determined as distinct from the determination of individual prices, notably by the Swedish economist Knut Wicksell. In the 1930s the growing harmony and unity of economics was rudely shattered, first by the simultaneous publication of Edward Chamberlin's *Theory of Monopolistic Competition* and Joan Robinson's *Economics of Imperfect Competition* in 1933 and then by the appearance of John Maynard Keynes's *General Theory of Employment, Interest and Money* in 1936.

The critics. Before going on, it is necessary to take note of the rise and fall of the German Historical school and the American Institutionalist school, which levelled a steady barrage of critical attacks on the orthodox mainstream. The German historical economists, who had many different views, basically rejected the idea of an abstract economics with its supposedly universal laws; they urged the necessity of studying concrete facts in national contexts. While they gave impetus to the study of economic history, they failed to persuade their colleagues that their method was invariably superior. The institutionalists are more difficult to categorize. "Institutional economics," as the term is narrowly understood, refers to a movement in American economic thought associated with such names as Thorstein Veblen, Wesley Clair Mitchell, and John R. Commons. These writers had little in common aside from their dissatisfaction with the abstract theorizing of orthodox economics, its tendency to cut itself off from the other social sciences, and its preoccupation with the automatic market mechanism. They failed to develop a theoretical apparatus that would replace or supplement the orthodox theory. This may explain why the phrase "institutional economics" has become little more than a synonym for "descriptive economics." The hope that institutional economics would furnish a new interdisciplinary social science proved stillborn. (This is perhaps not surprising, because it was by abstracting purely economic forces from the totality of social interactions that economics got so far ahead of the other social sciences in theoretical rigour.) Although there is no longer an institutionalist movement in economics, the spirit of institutionalism is alive in such works as the Harvard economist John Kenneth Galbraith's *The Affluent Society* (2nd ed., 1969) and *The New Industrial State* (1967).

Forms of imperfect competition

Returning to the innovations of the 1930s, the theory of monopolistic or imperfect competition remains somewhat controversial to this day. The older economists had devoted all their attention to two extreme types of market structure, that of "pure monopoly," in which a single seller controlled the entire market for one product, and that of "pure competition," characterized by many sellers, highly informed buyers, and a single, standard product. The theory of monopolistic competition gave recognition to the range of market structures that lie between these extremes, including (1) markets having many sellers with "differentiated products," employing brand names, guarantees, and special packaging that cause consumers to regard the product of each seller as unique; (2) "oligopoly," markets dominated by a few large firms; and (3) "monopsony," markets with a single monopolistic buyer and many sellers (see ECONOMIC THEORY). The theory produced the powerful conclusion that competitive industries in which each seller has a partial monopoly because of product differentiation will tend to have an excessive number of firms, all charging a higher price than they would if the industry were perfectly competitive. Since product differentiation—and the associated phenomenon of advertising—seems to be characteristic of most industries in developed capitalist economies, the new theory was immediately hailed as injecting a healthy dose of realism into orthodox price theory. Unfortunately, its scope was not great enough. It failed to provide a satisfactory theory of price determination under conditions of oligopoly. In advanced economies many of the manufacturing industries are oligopolistic. The result has been to leave a somewhat undigested lump at the centre of modern price theory, a constant reminder of the fact that economists still lack an adequate explanation of the conditions under which the giant firms of rich countries conduct their affairs.

Keynesian economics. The second major breakthrough of the 1930s, the theory of income determination, was primarily the work of one man—John Maynard Keynes. Keynes asked questions that in some sense had never been asked before; he was interested in the level of national income and the volume of employment rather than in the equilibrium of the firm or the allocation of resources. It was still a problem of demand and supply, but "demand" here means the total level of effective demand in the economy, and "supply" means the nation's capacity to produce. When effective demand falls short of productive capacity, the result is unemployment and depression; when it exceeds the capacity to produce, the result is inflation. The heart of Keynesian economics consists of an analysis of the determinants of effective demand. If one ignores foreign trade, effective demand consists essentially of three spending streams: consumption expenditures, investment expenditures, and government expenditures, each of which is independently determined. Keynes attempted to show that the level of effective demand so determined may well exceed or fall short of the physical capacity to produce goods and services: that there is no automatic tendency to produce at a level that results in the full employment of all available men and machines. This fundamental implication of the theory came as something of a shock to exponents of the traditional economics who had been inclined to take refuge in the assumption that economic systems tend automatically to full employment. By keeping his attention focussed on macroeconomic aggregates, like total consumption and total investment, and by a deliberate simplification of the relations between these economic variables, Keynes achieved a powerful model that could be applied to a wide range of practical problems. His system subsequently underwent considerable refinement—some have said that Keynes himself would hardly have recognized it—and became thoroughly assimilated into the body of received doctrine (see ECONOMIC THEORY). Still, it is not too much to say that Keynes is perhaps the only economist to have added something really new to economics since Walras and perhaps since Ricardo.

Keynesian economics as conceived by Keynes was entirely "static"; that is, it did not involve time as an important variable. But a disciple of Keynes, Roy Harrod, soon developed a simple macroeconomic model of a growing economy; in 1948 he published *Towards a Dynamic Economics,* launching an entirely new speciality, "growth theory," which absorbed the attention of an increasing number of economists.

Postwar developments. In the 25-year period following World War II, economics was so totally transformed that those who studied it before the war might as well have lived in another world. First of all, there was an enormous increase in the use of mathematics, which came to permeate virtually every branch of economics. Previously, few economists had made use of mathematics other than differential and integral calculus. Matrix algebra became important with the advent of "input–output analysis," an empirical method of reducing the technical relations between industries to a manageable system of simultaneous equations; it was an attempt to put quantitative flesh on the bones of a general equilibrium model of the economy. A closely related phenomenon was the development of "linear programming" and "activity analysis," which opened up a whole host of industrial problems to numerical solution and introduced economists for the first time to the mathematics of "inequalities" rather than exact equations. Likewise, the emergence of "growth economics" promoted the use of difference and differential equations.

New mathematical tools

Hand in hand with the spread of mathematical economics went an increasing sophistication of empirical work under the rubric of "econometrics," comprising a combination of economic theory, mathematical model building, and statistical testing of economic predictions. The development of econometrics had an impact on economics in general, since those who formulated new theories began to cast them in terms that allowed them to be empirically tested.

The postwar years also saw a renewal of interest in the underdeveloped countries. Economists became aware that they had too long neglected "an inquiry into the causes of the wealth of nations." There was also a conviction that economic planning of one variety or another was needed to close the gap between the rich and poor countries. Out of these concerns came the field of development economics. Regional economics, urban economics, health economics, and the economics of education are other offshoots of the mainstream since 1945.

The postwar tendencies in economic thought were best exemplified, not by the emergence of new techniques or by the addition of new parts to the economics curriculum but by the disappearance of divisive "schools," by the increasingly standardized professional training of economists all over the world, and by the transformation of the science from a rarefied academic exercise to an operational discipline geared to practical advice. This transformation brought prestige to the profession but also a new responsibility: now that economics really mattered, economics had to reckon with the conflict that so often exists between analytical precision and economic relevance.

The question of relevance was at the centre of a "radical critique" of economics that developed along with the campus revolts of the late 1960s. The radical critics declared that economics had become a defense of the status quo and that its practitioners had joined the power elite. The marginal techniques of the economists, ran the argument, were profoundly conservative in their bias, and encouraged a piecemeal rather than a revolutionary approach to social problems; likewise, the tendency in theoretical work to ignore the everyday context of economic activity amounted in practice to the tacit acceptance of prevailing institutions. The critics said that economics should abandon its claim of being a value-free social science and address itself to the great questions of the day—those of civil rights, poverty, imperialism, and nuclear war—even at the cost of analytical rigour and theoretical elegance.

It is true that the study of economics encourages a belief in reform rather than revolution; economics as a science does not provide enough certitude for any thoroughgoing reconstruction of the social order. It is also true that most economists tend to be deeply suspicious of monopoly in all forms, including state monopolies, and to favour competition between independent producers as a way of diffusing economic power. Finally, most economists prefer to be silent on large questions if they have nothing to offer beyond the expression of personal preferences: most economists as economists are fundamentally concerned with the professional standards of their discipline, and this may mean in some cases frankly conceding that economics has as yet nothing very interesting to say about these questions. (It does not mean, however, that they desire to justify the status quo.) Yet the radical critique of modern economics was not to be lightly dismissed. The radical economists were posing issues that were important. At the very least it could do the economic researcher no harm to think about the social and political relevance of his work.

METHODOLOGICAL CONSIDERATIONS IN CONTEMPORARY ECONOMICS

The role of value judgments

Economists are sometimes confronted with the charge that their discipline is not a science. Human behaviour, it is said, cannot be analyzed with the same objectivity as the behaviour of atoms and molecules. Value judgments, philosophical preconceptions, and ideological biases must interfere with the attempt to derive conclusions that are independent of the particular economist espousing them. Moreover, there is no laboratory in which economists can test their hypotheses.

This argument raises issues for all of the social sciences. Only a very general reply can be given here. Economists are wont to distinguish between "positive economics" and "normative economics." Positive economics seeks to establish facts: Will a subsidy to butter producers lower the price of butter? Will a rise in wages in the automobile industry reduce the employment of automobile workers? Will devaluation improve the balance of payments? Does monopoly foster technical progress? Normative economics, on the other hand, is not concerned with matters of fact but with questions of policy, of "good" or "bad": Should the goal of price stability be sacrificed to that of full employment? Should income be taxed at a progressive rate? Should there be legislation in favour of competition?

Positive economics in principle involves no judgments of value; its findings may be as impersonal as those of astronomy and meteorology, two natural sciences that are also denied the advantage of conducting laboratory experiments. As the British philosopher David Hume argued 200 years ago, there is no logical way to deduce "ought" from "is" or prescriptions from descriptions; all statements of fact are ethically neutral. In that sense a value-free economics is possible (at least in principle): if economics is about the application of means to achieve given ends, there would seem to be no reason why one cannot analyze the allocation of means to achieve *any* end. This is not to deny that most of the interesting economic propositions involve the addition of definite value judgments to a body of established facts, that ideological bias creeps into the very selection of the questions that economists investigate, that what is a means from one point of view may be an end from another, nor even that much practical economic advice is loaded with concealed value judgments, the better to persuade rather than merely to advise. This is only to say that economists are human. The commitment of economists the world over to the ideal of value-free positive economics (or to the candid declaration of personal values in normative economics) serves as a defense against the attempts of special interests to bend the science to their own purposes. The best assurance against bias on the part of any particular economist is the criticism of other economists. The best protection against special pleading in the name of science is the professional standards of scientists.

Importance of statistics

Methods of inference. But how, one may ask, are facts established in a science that cannot conduct experiments? In essence, the answer is: by means of statistical inference. Economists typically begin by describing the area under study according to what they feel to be important. Then they construct a "model" of the real world, deliberately repressing some of its features and emphasizing others; they abstract, isolate, and simplify, thus imposing order on a world that at first glance appeared disorderly. Having evolved an admittedly unrealistic representation of the real world, they then manipulate the model by a process of logical deduction, arriving eventually at some prediction or implication that is of general significance. At this point, they return to the real world to see whether or not the prediction is borne out by observed events.

But the observable events that are available to test a theory never exhaust the population of all such events: they are merely a sample of it. This raises the problem of statistical inference; namely, what can be inferred about a population from a sample of the population? The theory of statistical inference is simply an agreed-upon procedure for making such inferences, but in the nature of the case it never succeeds in removing all elements of judgment from an inference. Thus the empirical truths of economics are invariably surrounded by a band of doubt, and economists speak of them as "probable" or "likely"; they are propositions in which economists have "a certain degree of confidence" because it is unlikely that they could have come about by chance.

It follows that judgments are at the heart of both positive and normative economics. It is easy to see, however, that judgments about "degrees of confidence" and "statistical levels of significance" are of a totally different order from those that crop up in normative economics. When men say that every individual should be allowed to spend his income as he likes, that people should not be free to control material resources and to employ others, or that governments must offer relief for the victims of inexorable economic forces, they are making the kind of value judgments that laymen have in mind when they accuse economists of producing personal preferences in the guise of scientific conclusions. There is no room for such value judgments in positive economics.

Testing theories. In the past some economists tended to claim too much for their propositions. Economic models were said to be based on fundamental axioms and premises about economic behaviour that were absolutely true a priori because they were derived from an examination of one's own economic behaviour. Since the theorems of the model were deduced from these axioms by the laws of logic, the theorems also were held to be true a priori. Economic models did not need to be confronted with empirical evidence.

The problem of validity

This extreme apriorist position may be contrasted with the ultraempiricist view, which holds that one must begin and end with observable facts; the latter approach, however, has never appealed to more than a small minority of economists. In the middle ground between these two sharply opposing views is the methodological position that has found increasing acceptance among modern economists. It argues that one must test the predictions or conclusions of a model but without worrying too much about the realism of its premises, axioms, or assumptions. Most assumptions in economic theory cannot be tested directly. For example, there is the famous assumption of price theory that businessmen strive to maximize profits. Attempts to find out whether they do, by asking them, usually fail; after all, businessmen are no more fully conscious of their own motives than other people are. A logical approach would be to observe businessmen in action. But that would require knowing what sort of action is associated with profit maximizing, which is to say that one would have drawn out all the implications of a profit maximizing model. Thus one would be testing an assumption about business behaviour by comparing the predictions of a theory of the firm with observations from the real world.

This is not as easy as it sounds. Since the predictions of economics partake of the nature of probability statements, there can be no such thing as a conclusive, once-and-for-all test of an economic hypothesis. The science of statistics cannot prove any hypothesis, it can only fail to disprove it. A theory that survives a statistical test is not true as such; it is only provisionally true because it has so far resisted all attempts to falsify it. Attempts to falsify economic hypotheses never yield unambiguous results, and hence economic theories tend to survive until they are falsified repeatedly with new or better data. This is not because they are *economic* theories but because the attempt to compare predictions with outcomes in the social sciences is always limited by the rules of statistical inference.

It is not remarkable that competing theories exist to explain the same phenomena, with economists disagreeing as to which theory is to be preferred. While virtually all economists today agree that theories should be submitted to empirical testing and that the theory is to be preferred that allows predictions that conform, in a probabilistic sense, most closely to observable events, this precept can be very difficult to apply in practice. There have been periods in the history of economics when there was overwhelming agreement in the profession as to which models or theories were "true." But a period of consensus may be followed by a generation of doubt until a new departure is made that succeeds in producing a new consensus. In this, economics is not very different from physics.

Much has been written about the doubtful accuracy of economists' predictions. Of course, economists cannot foretell the future as such; only soothsayers do that. Economists can foretell the effects of specific changes in the economy, but they are better at predicting the direction than the actual magnitude of events. When economists predict that a tax cut will raise national income, one may be confident that the prediction is accurate; when they predict that it will raise national income by a certain amount in three years, however, the forecast is likely to miss the mark. The reason is that most economic models do not contain any explicit reference to the passage of time and hence have little to say about how long it takes for a certain effect to make itself felt. Short-period predictions generally fare better than long-period ones, in part because most economic models rely on propositions about the plans and intentions of economic agents, whereas the data on which the theories are tested are derived from past events. This is disappointing, but it does not mean that economics is not a science.

The firm and the household

Microeconomics. Since Keynes, economic theory has been of two kinds: macroeconomics—or the study of the determinants of national income—and the traditional microeconomics. The latter approaches the economy as if it were made up only of business firms and households (ignoring governments, banks, charities, trade unions, and all other economic institutions) interacting in two kinds of markets—product markets and markets for productive services, or factor markets. Households appear as buyers in product markets and as sellers in factor markets, where they offer men, machines, and land for sale or hire. Firms appear as sellers in product markets and as buyers in factor markets. In each type of market, price is determined by the interaction of demand and supply, and the problem of microeconomic theory is to say something meaningful about the forces that make up demand and supply.

Theory of choice. At first it appears that all one can say is that everything depends on everything else. But firms and households do not behave in a vacuum. Firms face certain technical constraints in producing goods and services, and households have definite preferences for some products over others. It is possible to express the technical constraints facing business firms by writing down a series of "production functions," one for each firm. A production function is simply a kind of equation that expresses the fact that the output of a firm depends on the quantity of inputs it employs and, in particular, that inputs can be technically combined in different proportions to produce a given level of output. A production engineer could calculate, on the basis of existing technical knowledge, the largest possible output that could be produced with every possible combination of inputs and in this way could define a boundary to the range of production possibilities open to a firm. By itself this does not tell how much the firm will produce or what mixture of products it will make or what combination of inputs it will adopt: these depend on the prices of products and the prices of inputs (or "factors of production"), which have yet to be determined. One may assume that the firm is motivated in a particular way: it wants to maximize profits, which are defined as the difference between the sales value of its output and the money outlays required to obtain its inputs. It will, therefore, select that combination of inputs that minimizes the costs of producing any given quantity of output and will select from the range of possible combinations of products that combination that maximizes its revenues. This is to say that it always tries to move along its production function, along the edge of the boundary of technical possibilities. But where it ends depends, in part, on the demand for its products. This leads to the part played by households in the system.

Households are endowed with definite "tastes" that can be expressed in a series of "utility functions," one for each household. A utility function is an equation like a production function, expressing the fact that the pleasure or satisfaction that households derive from consumption depends on the products that they purchase and on the various ways in which they combine these products in consumption to yield a given level of satisfaction. The utility function need not be specified in the same detail as a production function. One may think of it as a general description of the household's preferences between all the paired alternatives with which it will be confronted. Here, too, it is necessary to assume something about motivation to make any progress: the assumption is that households seek to maximize satisfaction, distributing their given incomes among available consumer goods in such a way as to derive the largest possible "utility" from consumption. Their incomes, however, remain to be determined.

Supply and demand curves

The purpose of production functions in economic theory is to provide an anchor in the bedrock of technology from which to derive the "supply curves" of firms in product markets and the "demand curves" of firms in factor markets. Similarly, the purpose of utility functions is to provide an anchor in subjective "tastes" from which to derive the "demand curves" of households in product markets and the "supply curves" of households in factor markets. All of these demand and supply curves express the quantities demanded and supplied as a function of prices, not because price is the only determinant of economic behaviour but because the purpose is to have a theory of price determination. Much of economic theory is devoted to showing how various production and utility functions, coupled with certain assumptions about behaviour, lead to demand and supply curves in which the quantity demanded is inversely related and the quantity supplied positively related to price. The figure depicts these relationships (curves would be just as suitable as straight lines).

Not all demand and supply curves look alike. The essential point is that most demand curves are negatively inclined, while most supply curves are positively inclined. This may seem a modest result for a great deal of effort, but the argument has powerful implications. The participants in a market will be driven automatically to the price at which the two curves intersect; this price p is called the "equilibrium" price or "market-clearing" price because it is the only price at which supply and demand are equal. If it is a market for butter, any change in the production function of dairy farmers or in the utility function of butter consumers or in the prices of cows, grassland, and milking equipment or in the incomes of butter consumers or in the prices of nondairy products that consumers buy can be shown to lead to definite changes in the equilibrium prices of butter and in the equilibrium quantity of butter produced. Better still, the effects of a government ceiling on the price of butter or of a tax on butter producers or of a price-support program for dairy farmers can be predicted with almost perfect certainty. As a rule, the prediction will refer only to the direction of change (the price will go up or down); but if the demand and supply curves of butter can be defined in quantitative terms on the basis of past data, one may be able to predict the actual magnitude of the change.

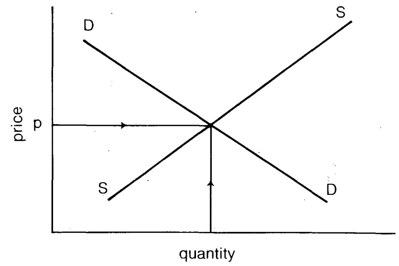


Illustration of the relationship of price to supply (S) and demand (D).

Theory of allocation. This analysis of the behaviour of firms and households is to some extent symmetrical: all economic agents are conceived of as ordering a series of attainable positions in terms of an entity they are trying to maximize. For a firm these attainable positions are essentially input combinations; for a household they are product combinations. From the maximizing point of view, some combinations are better than others; the best combination is called the "optimal" or "efficient" combination. The rule for efficient, optimum allocation may now be stated baldly: an optimum allocation is one that equalizes the returns of the marginal or last unit to be transferred between all the possible uses. In the theory of the firm, an optimum allocation of outlays among the factors of production implies that the "marginal physical product" of an additional dollar devoted to hiring the services of any one of the factors is the same for all factors; the so-called law of eventually diminishing marginal productivity, a property of a wide range of production functions, ensures that such an optimum exists. In the theory of consumer behaviour an optimum situation obtains when the consumer has distributed his given income in such a way that the "marginal utility" of each additional dollar spent on any of the products purchased is equal for all products; the "law of eventually diminishing marginal utility," a property of a wide range of utility functions, ensures that such an optimum exists. These are merely particular examples of the "equimarginal principle," which is not only at the core of the theory of the firm and the theory of consumer behaviour but also underlies the theory of money, of capital, and of international trade. In fact, the whole of microeconomics is nothing more than the spelling out of this principle in ever wider contexts.

Law of diminishing marginal productivity

The equimarginal principle is, of course, applicable to any decision that involves alternative courses of action. Economics furnishes a technique for thinking about decisions, whatever their character and whosoever makes them. Military planners may, for example, consider a variety of weapons in the light of a single objective, that of damaging an enemy; some of the weapons are effective against the enemy's army, some against the enemy's navy, and some against his air force; the problem is to find an optimal allocation of the defense budget, one that equalizes the marginal contribution of each type of weapon. But defense departments rarely have single objectives; along with maximizing damage to an enemy there may be another objective, such as minimizing losses from attacks. In that case, more than the equimarginal principle is needed for a decision; it is necessary to know how the department ranks the two objectives in order of importance, since different rankings will imply different optima. But a ranking of objectives is simply a utility function or a preference function.

In other words, when an institution pursues multiple ends, decisions about how to achieve them require a weighting of the ends. Every decision involves a "production function"—a statement of what is technically feasible—and a "utility function"; the equimarginal principle is then invoked to provide an efficient, optimal strategy. This applies just as well to the running of hospitals, churches, and schools as to the conduct of a business enterprise, to the location of an international airport as well as to the design of a development plan for an underdeveloped country. This is why economists crop up in what seem to be the most unlikely places, advising on activities that are obviously not being conducted for economic reasons.

The different approach of macroeconomics

Macroeconomics. There is, however, an approach to economics in which the foregoing considerations do not apply. That is the field known as macroeconomics. In macroeconomics one is concerned with the aggregate outcome of individual actions. Keynes's "consumption function," for example, which relates aggregate consumption to national income, is not built up from individual consumer behaviour; it is simply an empirical generalization. The focus is on income and expenditure flows rather than on markets. Purchasing power flows through the system from business investment to consumption, but it leaks out at two places in the form of personal and business savings. Counterbalancing the savings are investment expenditures in the form of new capital goods, houses, and so forth, which constitute a source of new injections of purchasing power in every period. Since savings and investments are carried out by different people for different motives, there is no reason why "leakages" and "injections" should be equal in every period. If they are not equal, national income, the sum of all income payments to the factors of production, will rise or fall in the next period. When planned savings equal planned investment, income will be at an equilibrium level, that is, a level at which it can sustain itself; when the plans of savers do not match those of investors, the level of income will go on changing until the two do match. One can complicate this simple model by making investment a function of the interest rate; by introducing the government budget, the money market, labour markets, imports and exports, foreign investment; and so forth. But all this is far removed from the problem of resource allocation and from the maximizing behaviour of individual economic agents.

The result is a kind of intellectual schizophrenia in which the techniques of microeconomics do not carry over fully into macroeconomics and vice versa. This is widely held to be an unsatisfactory state of affairs; economists have in recent years sought to build a bridge between the individual consumer and the overall consumption function and between the individual investor and the behaviour of aggregate investment. Nevertheless, the bridge remains incomplete, and the student of economics must be prepared to work with two boxes of tools.

FIELDS OF CONTEMPORARY ECONOMICS

The following is a bird's-eye view of the main fields of contemporary economics.

Public finance. Since the time of Ricardo economists have been concerned with the incidence of taxes, that is, with determining who it is that really pays a tax. If a corporation faced with a profits tax reacts by raising its prices, it may succeed in making the consumer pay the tax; on the other hand, if sales decline as a result of the rise in price the firm may have to lay off some of its workers, and the burden of the tax will be shared by consumers, wage earners, and shareholders. This simple example shows how complex may be the actual incidence of a tax. A large part of the literature of public finance in the 19th century was devoted to such problems.

Taxes and government expenditures

Keynesian economics brought new dimensions to public finance: the older preoccupation with tax incidence gave way to the analysis of the impact of government expenditures on the level of income and employment (see GOVERNMENT FINANCE). It was some time, however, before economists realized that they lacked a theory of government expenditures, that is, a set of criteria for determining what activities should be supported by governments and what should be the relative expenditure on each. One of the most exciting recent developments in the field of public finance is the attempt to devise such criteria. Decisions on public expenditures have proved to be susceptible to much of the traditional analysis of microeconomics. In the 1960s there developed a technique known as "cost–benefit analysis," which tries to appraise all of the economic costs and benefits, direct and indirect, of a particular activity so as to decide how to distribute a given public budget most effectively between different activities. This technique has been applied to everything from the construction of hydroelectric dams to the control of tuberculosis. Its exponents hope that the same type of analysis that has proved so fruitful in the past in analyzing individual choice may also succeed with problems of social choice.

Money. One of the oldest, widely accepted functions of government is control over the supply of money. The dramatic effects of changes in the quantity of money on the level of prices and the volume of economic activity were recognized and thoroughly analyzed in the 18th century, and monetary economics has ever since constituted one of the principal branches of economics. In the 19th century a complex and somewhat crudely formulated tradition grew up known as the "quantity theory of money," which held that any change in the supply of money can only be absorbed by variations in the general level of prices (the purchasing power of money). In consequence, prices will tend to change proportionately with the quantity of money in circulation. As the growth of fiat paper money gave governments increasingly effective control over the stock of circulating media, the quantity theory of money supplied an apparently simple rationale for the management of the economy: all that was needed to prevent inflation or deflation was to vary the quantity of money in circulation inversely with the level of prices.

Keynes's attack on monetary policy

One of the targets of Keynes's attack on traditional thinking in his *General Theory of Employment, Interest and Money* was this quantity theory of money. Keynes produced a different theory of the demand for money that implied that the impact of a change in the stock of money on the level of national income is weak and at best indirect; the effect on prices is virtually nil, he maintained, at least in economies with heavy unemployment such as prevailed in the 1930s. He put his emphasis instead on government budgetary and tax policy and direct control of investment. As a consequence, economists lost faith in monetary management and came to regard monetary policy as more or less ineffective in controlling the volume of economic activity.

In the 1960s there was a remarkable revival of the older view, at least among a small but growing school of American monetary economists. They accepted much of Keynesian economics but argued that the effects of fiscal policy are unreliable unless the quantity of money is regulated at the same time. They refurbished the quantity theory of money and tested the new version on a variety of data for different countries and for different time periods, leading to the broad conclusion that the quantity of money does matter (see MONEY).

In the late 20th century the controversy was still raging. It is notable that this debate, unlike previous debates in the history of monetary economics, was characterized by disputes over empirical findings—that is, it was focussed on the testable character of different monetary theories rather than on the manner of their formulation. Progress was made slower by the political overtones of the controversy: in some countries, belief in the efficacy of monetary policy had become a kind of litmus test of political conservatism. Nevertheless, a reconciliation between Keynesians and quantity theorists needed only some agreement as to the magnitude of monetary forces and the degree of stability of the demand for money. Monetary economics seemed at last to be coming of age as an empirical discipline.

International economics. The foundations of international economics were firmly established in the 19th century. The subject has consisted ever since of two distinct but connected parts: (1) the "pure theory of international trade," which seeks to account for the gains obtained from trade and to explain how these gains are distributed among countries, and (2) the "theory of balance-of-payments adjustments," which analyzes the workings of the foreign exchange market, the effects of alterations in the exchange rate of a currency, and the relations between the balance of payments and the level of economic activity.

In modern times, the Ricardian pure theory of international trade has been reformulated by the American Paul Samuelson, improving on the earlier work of two Swedish economists, Eli Heckscher and Bertil Ohlin. The so-called Heckscher-Ohlin theory explains the pattern of international trade as determined by the relative land, labour, and capital endowments of countries: a country will tend to have a relative cost advantage in goods requiring the intensive use of the country's relatively abundant factor of production (thus land-rich Canada exports wheat) and to import goods requiring the intensive use of the country's relatively scarce factor (thus capital-poor Canada imports American automobiles). This theory absorbs Ricardo's law of comparative costs but goes beyond it in linking the pattern of trade to the economic structure of trading nations. It implies that foreign trade is a substitute for international movements of labour and capital—which raises the intriguing question of whether or not foreign trade may work to equalize the prices of all factors of production in all trading countries. Whatever the answer, the Heckscher-Ohlin theory provides a model for analyzing the effects of a change in trade on the industrial structures of economies and, in particular, on the distribution of income between factors of production. Much of the recent effort of specialists in international economics has gone toward refining the Heckscher-Ohlin model and testing it on an ever wider range of empirical evidence.

The economics of labour and industry

Labour. Like monetary and international economics, labour economics is an old economic speciality. It gets its raison d'être from the peculiarities of labour as a commodity. Labour itself is not bought and sold; rather, its services are hired and rented out. But since people cannot be disassociated from their services, various nonmonetary considerations play a role in the sale of labour services as contrasted with the sale of machine time or the rental of land. Yet, the bulk of the literature in labour economics was until recently concerned solely with the demand side of the labour market. Wages, the textbooks all said, were determined by the "marginal productivity of labour," that is, by the relationships of production and by consumer demand. If the supply of labour came into the picture at all, it was merely to allow for the presence of trade unions; unions could only raise wages by limiting the supply of labour.

After a long period of neglect, the supply side of the labour market began, in the 20th century, to attract the attention of economists. First, attention shifted from the individual worker to the household as a supplier of labour services; the increasing tendency of married women to enter the labour force and the wide disparities and fluctuations observed in the rate that females participate in a labour force drew attention to the fact that an individual's decision to supply labour is not independent of the size, age structure, and asset holdings of the household to which he or she belongs. Second, the new concept of "human capital"—that people make capital investments in their children and in themselves by incurring the costs of education and training, the costs of searching for better job opportunities, and the costs of migration to other labour markets—has served as a unifying explanation of the diverse activities of households in labour markets. In this way, capital theory (see ECONOMIC THEORY) has become the dominant analytical tool of the labour economists, replacing or supplementing the traditional theory of consumer behaviour. The economics of training and education, the economics of information, the economics of migration, the economics of health, and the economics of poverty are some of the by-products of this new perspective. A field that was at one time regarded as rather cut-and-dried has taken on new vitality.

Labour economics, old or new, has always regarded the explanation of wages as its principal task, including the factors determining the general level of wages in an economy and the reasons for wage differentials between industries and occupations. Wages are influenced by trade unions; the impact of their activities is of increased importance at a time when most governments manage the economy with one eye on the unemployment statistics. The prewar fears of chronic unemployment gave way to the postwar fears of chronic inflation at or near levels of full employment. In response to this a vast literature sprang up after 1945 analyzing the inflationary pressures stemming from both the supply side and the demand side of labour markets. Whether prices were being pushed up by the labour unions ("cost push") or pulled up by excess purchasing power ("demand pull") became the issues in this long debate on inflation, a controversy that is, of course, intimately related to the quarrels in monetary economics mentioned earlier.

Industrial organization. The principal concerns of this field are the structure of markets, public policy toward monopoly, the regulation of public utilities, and, of late, the economics of technical change.

The monopoly problem, or, more precisely, the problem of the maintenance of competition, does not fit well into the received body of economic thought. Economics started out, after all, as the theory of competitive enterprise, and even today its most impressive theorems require the assumption of numerous small firms, each having a negligible influence on price. Yet the typical market structure of manufacturing today is that of oligopoly—competition among the few—and some industries are dominated by firms so large that their annual sales volume exceeds the national income of the smaller countries of western Europe. It is tempting to leap to the conclusion that oligopoly is deleterious to economic welfare, on the ground that it leads to the misallocation of resources. But some economists, notably Joseph Schumpeter, have argued that economic growth and technical progress are brought about not through free competition but through large firms and the destruction of competition. According to this view, monopoly has its origin in the need of business firms to protect themselves from the risks associated with the introduction of new products, new techniques, and new methods of marketing. The giants, therefore, compete not in price but in successful innovation, and this kind of competition has proved more effective for economic progress than the more traditional price competition described in orthodox textbooks of economic theory.

Merits of monopoly and competition

Although this thesis smacks of *post hoc ergo propter hoc* ("after this, therefore because of this")—giant firms have prospered in rapidly growing economies; therefore, growth is due to giant firms—it makes the merits of "trust busting" and cartel dissolution somewhat less compelling. The question is what sort of competition is socially most desirable. If each of four or five large firms in an oligopolistic industry finds it necessary to compete in terms of the quality of its products and its research or by means of better technology and superior merchandising, the performance of the industry may well be more satisfactory than if it were reorganized into a price-competitive industry. But if the four or five giants settle down to a quiet life and concentrate their rivalry on sales promotion techniques, the verdict must be less favourable. One cannot, it seems, draw facile conclusions about the competitive results of different market structures; it is necessary to approach the monopoly problem with a generous dose of pragmatism.

The reason why there is so much uncertainty in the economic discussion of policies toward big business is the lack of a general theory of oligopoly. There are dozens of special theories applying to special cases, but there is no single, organizing framework with testable implications about the behaviour of oligopolists in general.

Agriculture. Farming has long provided economists with their favourite example of a perfectly competitive industry. But with increasing government regulation of agriculture, it also provides striking examples of the effects of price controls, income supports, output ceilings, and marketing cartels. Agricultural economics commands attention wherever governments wish to stimulate farming or to protect farmers—which is to say, everywhere.

Agricultural economists generally have been closer to their subject matter than other economists. In consequence, more is known about the technology of agriculture, the nature of farming costs, and the demand for agricultural goods than is known about any other industry. The student of economics who wants to learn how to estimate a production function or a demand curve is well advised to go to the literature on agricultural economics.

The underdeveloped countries have furnished a new laboratory for agricultural economics. Many such countries have a "subsistence agriculture," in which the farmer produces mainly for his own family consumption and brings to market only what is left over. Subsistence farmers are only tenuously linked to the money economy. They are more reluctant than commercial farmers to take the risks entailed in experimenting with new seeds, fertilizers, and farming methods: given the vagaries of the weather, they prefer to operate on the basis of the worst weather that can be expected rather than the best or even the average. While the pessimism of subsistence farmers is perfectly rational, it makes the task of predicting their response to new prices or new methods doubly difficult. The economist must also recognize that every agricultural area presents its own production, processing, and marketing problems. To suggest changes that will raise agricultural productivity in these circumstances is no easy task, and agricultural economists often find themselves warning governments not to intervene too vigorously, since much intervention in the past has been shown to be wrong or inappropriate.

The problem of economic development in Africa, Asia, and Latin America centres on the agricultural sector; one of the abiding questions is how far industrialization can proceed without there first being an agricultural revolution. For this reason, if for no other, agricultural economics has a large future.

Growth and development. The study of economic growth and development is not a single branch of economics but falls, in fact, into two quite different fields. The two fields—"growth" and "development"—employ different methods of analysis and are indeed addressed to two distinct types of inquiry.

Development economics is easy to describe. It is one of the three major subfields of economics, the other two being microeconomics and macroeconomics. Development economics resembles economic history in that it seeks to explain the changes that take place in economic systems with the passage of time.

The subject of economic growth is not so easy to characterize. It is the most technically demanding field in the whole of modern economics, impossible to grasp for anyone who lacks differential calculus. Its focus is the properties of equilibrium paths, rather than equilibrium states. One makes a model of the economy and puts it into motion, requiring that the time paths described by the variables be self-sustaining in the sense that they continue to be related to each other in certain characteristic ways. Then one can investigate the way economics might approach and reach these steady-state growth paths from given starting points. Beautiful and frequently surprising theorems have emerged from this experience, but as yet there are no really testable implications nor even definite insights into how economies grow.

Harrod–Domar growth model

Growth theory began with the work of Roy Harrod in England and Evsey Domar in the United States. Their joint product has been known ever since as the Harrod–Domar model. Keynes had shown that new investment has a multiplicative effect on income and that the increased income generates extra savings to match the extra investment, without which the higher income level could not be sustained. One may think of this as being repeated from period to period, remembering that investment, apart from raising income disproportionately, also generates the capacity to produce more output that cannot be sold unless there is more demand, that is, more consumption and more investment. That is all there is to the model. It contains one behavioral condition—that people tend to save a certain proportion of extra income, a tendency that can be measured. It contains one technical condition—that investment generates additional output, a fact that can be established. And it contains one equilibrium condition—that planned saving must equal planned investment in every period if the income level of the period is to be sustained. Given these three conditions, the model generates a time path of income and even indicates what will happen if income falls off the path.

More complex models have since been built, incorporating different saving ratios for different groups in the population, technical conditions for each industry, definite assumptions about the character of technical progress in the economy, monetary and financial equations, and much more.

Mathematical economics. Differential calculus has long been the traditional tool of mathematical economics. Many economic problems, particularly in microeconomics, take the form of maximizing some variable (such as profits) subject to a constraint (such as the production function), for which calculus supplies the simplest technique. Traditionally it was applied to problems in comparative statics. These problems include so-called endogenous variables, the values of which are determined within the model, as well as constants that originate outside the model and are called "exogenous variables" or "parameters." The object is to discover the effects of changes in one or more of the parameters upon the equilibrium situation. The latter is a situation in which all of the endogenous variables are simultaneously in a state of rest. If the value of some of the parameters is changed, the result is a new equilibrium state.

The application of mathematical methods

Much economic analysis, even when it is expressed in words, is simply the method of comparative statics. But comparative statics has its limitations: it tells the investigator *where* the system will arrive, but it does not tell him *when* it will arrive or what will happen along the way; and it cannot tell him whether, once driven out of the way, it will ever get back to its destination. In other words, comparative statics ignores the process of adjustment from the old equilibrium state to the new one, and it entirely neglects the time element in that adjustment process. The study of this process of adjustment over time is called "economic dynamics," and one may think of it as the economics of disequilibria.

Just as differential calculus is the mathematics of comparative statics, difference and differential equations are the ideal tools for handling dynamic problems. Difference equations deal with time as a discrete variable—changing only from period to period—whereas differential equations treat time as a continuous variable; the choice between them is simply one of convenience. They enable one to ask such questions as: if the system is pushed out of equilibrium, perhaps because one of the parameters of the model has changed, will economic forces drive it toward a new equilibrium position or away from one, will the time path described by the endogenous variables be steady or fluctuating, and if fluctuating, will the movements be damped down or will they increase and become explosive?

Economic dynamics is one of the newer developments of mathematical economics, and often it falls short of the ambitious demands made on it. Dynamic models, for example, are typically formulated in terms of linear equations, not because the world is linear but because nonlinear equations can be very difficult to solve. Likewise, the coefficients of difference and differential equations are usually taken to be constants, again for the sake of making the mathematics of the analysis manageable. This means that if the economic environment changes as the model runs its course, its predictions will be false. An abiding danger in all mathematical economics is the tendency to adopt economic assumptions for the sake of mathematical convenience. The way to meet this danger is for economists to acquire enough mathematical sophistication so that they will not be dazzled by displays of mathematical technique.

Econometrics. Like mathematical economics, econometrics is something economists do rather than a special area of interest. Econometrics refers to the study of empirical data by statistical methods, the purpose of which is the testing of hypotheses and the estimation of relationships suggested by economic theory. Whereas mathematical economics considers the purely theoretical aspects of economic analysis, econometrics attempts to falsify theories that are expressed in explicit mathematical terms. But frequently the two go together.

The classic technique for estimating an economic relationship is that of "least squares," which is a method of fitting a trend line to a scatter of observations that minimizes the square of the deviations of the observed points from the line. To take a simple example: the Keynesian theory assumes that consumers' expenditures depend principally on income; one may interpret this to mean that consumption depends *only* on income and then test the hypothesis by trying to fit a trend line to a series of observations of income and consumption over a period of time. In so doing, one is really saying that the observations that fall to either side of the line are due either to errors in measuring the variables or to errors in specifying the relationship between consumption and income. It is essential to the method of least squares that these "errors" be randomly distributed or at any rate distributed in known ways. When this condition is violated, least squares estimates are unreliable. It is sometimes difficult to tell with economic data just how the errors are randomly distributed, and it is precisely for this reason that an econometrician is needed rather than an ordinary statistician.

Problem of error distribution

A still more significant trend in recent econometrics is the tendency to move from single-equation estimates (such as the relationship between consumption and income) to systems of simultaneous equations. While consumption depends on income, income also depends on consumption; this kind of interdependence requires two equations rather than one. More generally, most economic variables are the result of demand and supply forces that simultaneously determine quantities and prices. To estimate a demand curve for butter from a single-equation regression (by relating the price of butter to the quantities of butter consumed, the incomes of consumers, and the prices of near substitutes for butter) is likely to produce a biassed answer because the price of butter is also influenced by supply conditions in the dairy industry. This creates the so-called identification problem, namely, the question of whether it is possible to identify a demand curve or a supply curve from observed price–quantity data. The use of simultaneous equation models to estimate economic relationships is by now perhaps the best way of distinguishing econometrics from economic statistics.

The foregoing discussion covers only nine major branches of economics. There are many other fields in economics, including economic history, comparative economic systems, business cycles, economic forecasting, national income accounting, managerial economics, business finance, marketing, the economics of natural resources, economic geography, consumer economics, and regional economics.

(M.Bg.)

Political science

Political science is most generally understood to mean the systematic study of government processes by the application of scientific methods of analysis. More narrowly and more traditionally, it has been thought of as the study of the state and of the organs and institutions through which the state functions. In most countries, political science is thought to be a single discipline, but the plural form has been used in France, as in the name of the École Libre des Sciences Politiques (now Institut d'Études Politiques de l'Université de Paris), founded in 1871—although there is also an Association Française de Science Politique. Speculation about political subjects is not unknown in ancient non-Western cultures, but most students agree that the roots of political science are to be found in the earliest sources of Western thought, especially in the works of Aristotle, who is recognized by many as the founder of political science.

Political science and political philosophy

Although political science may be distinguished from political philosophy, the distinctions are unsatisfactory inasmuch as they lack categorical rigour. In the most usual distinction, political philosophy is thought to be concerned primarily with the study of political ideas, often within the context of their times. It is strongly normative in its thrust and disposition and rationalistic in its method. Political science, however, concerns itself with institutions and behaviour, eschews normative judgments as much as possible, and attempts to derive principles from objective facts with as much quantification as the evidence will allow. Political philosophy thus speculates about the place and order of values, the principles of political obligation (why men should or should not obey political authority), and the nature of such terms as right, justice, and freedom. Political science, on the other hand, seeks to establish by observation (and, if possible, by measurement) the existence of uniformities in political behaviour and to draw correct inferences from these data. The stated differences between political philosophy and political science are less than is sometimes supposed, for the most empirical scholar in both the social and natural sciences makes use of unproved postulates, hunches, and intuitions; and the most rationalistic philosopher employs conceptions that embody empirical statements.

Some theorists, however, who believe it possible to develop a completely value-free science of politics insist that the distinction between political philosophy and political science is not faint but vivid. It is their opinion that only a few hundred philosophers and political theorists have ever contributed to systematic speculation about and study of politics. It was said in 1966 by one of the exponents of this view that "probably two out of every three political scientists who have ever lived are alive and practicing today." In this view, political philosophy is believed to be addicted to obscurity and opacity of statement—an effort to bespeak the unspeakable—and it is thought that its task should be a more modest one, namely, to grope with the grammar of philosophical statements. This would require political philosophers to put their intelligence to the elucidation of the language of politics and to expose the difficulties placed by language in the consideration of matters of fact. One of those interested in this approach was the English political philosopher T.D. Weldon in *The Vocabulary of Politics* (1953), but few have followed his initiative.

Scientific status of the subject

The question as to whether political science is a science is largely inconsequential, because the problem is primarily one of definition. If the term science is to be applied to any body of systematically organized knowledge based on facts ascertained by empirical methods and described by as much measurement as the material allows, then political science is a science, just as are the other social disciplines. If, on the other hand, the term science is to be limited to those disciplines in which the scholar can control the materials to be studied and can perform experiments that others can reproduce under the same conditions and in which predictability is possible, then the label is less appropriate, although not entirely misapplied. The American political economist Thorstein Veblen denied that political science was anything more than a "taxonomy of credenda," and, more recently, a British writer, Bernard Crick, has said that the hope of creating an artificial science of politics on natural principles, although not originally and uniquely American, has been largely generated and sustained by two aspects of the American culture—an agreement on liberal doctrine that has made politics less a matter of serious doctrinal splits than of mere disagreement among partisans of the same creed, and a general national preoccupation with technology.

HISTORICAL DEVELOPMENT OF POLITICAL SCIENCE

Early trends. The origins of contemporary political science are to be found in the enthusiasm for the creation of social science that was widespread in the 19th century, an enthusiasm stimulated by the rapid growth of the natural sciences. It might be said that one starting point for the development of modern political science is the work of the Comte Henri de Saint-Simon, a notable Utopian Socialist, who in 1813 suggested that morals and politics could become "positive" sciences; that is, disciplines whose authority to command belief would rest not upon subjective preconceptions but upon objective evidence. With him worked the mathematician and philosopher Auguste Comte, the two collaborating in the publication in 1822 of the *Plan of the Scientific Operations Necessary for the*

Reorganization of Society, which argued, among much else, that politics would become social physics and that the purpose of social physics was to discover unchanging laws of progress. Out of this collaboration emerged the law of the three stages through which knowledge had to pass—the theological, the metaphysical, and the positive—that Comte was to establish as the theme of the science of social physics, a study he came to name sociology. An intellectual connection between political science and sociology was thus early established in schemes of political and social regeneration and reform, although political science was thought to be limited to only one form of association in society, namely, the state. Comte thought that the principal methods for the study of social phenomena were observation, experiment, and abstraction. Although one might have thought that politics could not be an experimental science, Comte was of the view that political experimentation did take place whenever there was a change in the life of the state, intended or not. It must be said, however, that even on this account there is no close similarity to experimentation in chemistry and physics, in which all the variables can be controlled.

In the search for more objective methods of inquiry into political and other social phenomena in the 19th century, contributions to the explanation of the state were supplied by several new intellectual disciplines. Because political science deals with some aspects of human behaviour, for example, it is closely allied to other social sciences that also deal with human behaviour. Long before the development of scientific inquiries in the 19th century, numerous theories of the state had drawn inspiration from the human being as model, as in the *Policraticus* of John of Salisbury (1159), in which the physiology of the body and that of the state are compared; or in *The Republic* of Plato, in which the elements of the human personality prefigure the class structure of the state; and in Rousseau's *Contrat Social,* in which the political order is animated by a general will (will being a human attribute). The positivism of the 19th century, however, brought new approaches to the study of the state, although the older ones continued to coexist with them. Among those following Comte was the Polish-born sociologist Ludwig Gumplowicz, who built a sociology on Comtean foundations but who owed much also to Darwin, to the social Darwinist Herbert Spencer, and to contemporary anthropology. In Gumplowicz' view, the earliest forms of group life were small hordes bound by consanguinity, which developed into matriarchies and patriarchies. He supposed the existence of a social-evolutionary process characterized by conflicts between autonomous groups and by conflicts of interest within those groups. The product of this process was the state, founded on force and maintained by power.

Sociological analyses of the state

Gumplowicz and several other 19th-century political sociologists anticipated 20th-century concerns in political science with the significance of groups, the nature of interests, the role of parties as interest groups, and the social context within which political events occur.

Still another 19th-century writer with some precedental connection to the political science of the 20th was the Italian Vilfredo Pareto. Although he lived and wrote in both the 19th and 20th centuries, he may be counted in the earlier period because of his advocacy of the "logicoexperimental" approach to sociology, which involved observation and logical inference. Pareto had no direct influence on the development of political science, but in two respects his sociology had implications for the developing discipline. First, his was a psychological sociology, and much of his concern was with the influence of beliefs, attitudes, opinions, and sentiments in shaping social life. This anticipates the 20th-century approach of many political scientists who regard psychology as the most important adjunct of the scientific study of politics. Second, Pareto thought of society as a system always tending toward equilibrium, and the conception of politics as a system was to mark much of academic political science after World War II (see below). Also to be mentioned in the context of important 19th-century sociological theories is the work of a Swedish political scientist, Rudolf Kjellen, whose systematic treatment of the state as a fusion of organic and intellectual moral elements in an ambience of geographical determinism led to a theory of politics that he termed geopolitics.

Juristic influences. In addition to sociology, another discipline with which the development of political science has been historically related is law. A close connection between the state and law was made in the 16th century in the French political philosopher Jean Bodin's theory of sovereignty, which supposed the necessary existence in each state of an authority to make the law. This theory gave rise to numerous juristic theories of the state, especially among German publicists of the 19th century who sought to define the nature of federation and empire. Although the doctrine of sovereignty made most sense as a statement of the power of monarchs to make the law in simple unitary (nonfederal) systems, earnest writers forced the facts of federation and empire to fit the theory and often ignored facts that did not. The Bodinian view invested sovereignty with attributes of singularity and omnipotence—that is, sovereignty was viewed as being indivisible and absolute—but theoretical efforts were made in Germany before 1871, notably by Georg Waitz, to establish federal theory on a division of sovereignty between the centre and the member states. After 1871, the theory of Max von Seydel, namely, that since sovereignty is indivisible, it cannot be divided and must rest in a single location—either in the constituent members of a federal system or in the centre—supplanted that of Waitz. Georg Jellinek, an Austrian, attempted to resolve the paradox that so-called sovereign states are in fact limited by constitutions and laws and by membership in the family of nations by arguing that the restrictions are autolimitations. That is to say, since by definition there is no power above the state, if in fact there are limitations it must be the state that created them. The state therefore continues to be omnipotent. In France, juristic theories of the state also persisted strongly because the teaching of political science was conducted in the law schools as "constitutional law." Efforts were made by some professors, however, to broaden the legal approach, notably by Léon Duguit, who attempted a sociological positivistic treatment of juridical rules for the limitation of the state, and Maurice Hauriou, who contributed a theory of institutions. In England the emergence of political science as a subject was recognized in the establishment of the London School of Economics and Political Science in 1895 and by the founding of a separate chair of politics at Oxford in 1912.

The doctrine of sovereignty

Developments in the United States. The enthusiastic development of social sciences in the 19th century, stimulated as it had been by the rapid growth of the natural sciences, reinforced an existing interest in politics in the United States and created a generation of distinguished American political scientists. There had, in fact, been much interest in the teaching of political subjects in American colleges and universities well before the 19th century. Political science in the United States, however, free of any connection with moral philosophy, fusion with history, or submergence in political economy, may be said to date from 1880, when John W. Burgess, after studying at the École Libre des Sciences Politiques, succeeded in establishing a separate school of political science at Columbia University. Although political-science faculties increased in numbers after 1900, the growth was uneven, and in some major institutions separate departments were not created until after World War I.

The development of American political science in the last quarter of the 19th century was influenced by the experience of numerous scholars who had done graduate work at German universities in which political science was taught as *Staatswissenschaft* ("science of the state") in an ordered, structured, and analytical organization of concepts, definitions, comparisons, and inferences. To modern readers the work of these men often seems somewhat formalistic and institutional in tone and focus. It did represent, however, an effort to establish an autonomous discipline, separate from history, moral philosophy, and political economy. Some of the new American political scientists, such as Woodrow Wilson and Frank Goodnow, also showed in their writing an awareness of such new intellectual currents as theories of evolution and turned their attention to an examination of American institutions that, in the United States Constitution, had originally been based on the admiration of much of the 18th-century world for the harmonious perfections of mechanics. For Wilson, it was to be Darwin, not Newton, who would provide the inspiration for a transformation in American political science from the study of static institutions to the study of social facts, more truly in the positivist temper, less in the analytic tradition, and more oriented toward factual realism.

Bentley's early work on group processes

Little-noticed at the time he wrote *The Process of Government* (1908), Arthur F. Bentley was to have a celebrated influence on the development of American political science in the 1930s and the 1950s. Although his influence was not immediately felt, he sounded certain themes that became the orthodoxy of the new science of politics after World War II. First, he rejected all metaphysics and normative formulations as "spooks" and "mind-stuff" and insisted that the proper study of politics was observable fact, in imitation of the natural sciences. Second, his basic concept was the "group," rejecting thereby all previous formulations of the subject matter of political science that centred on the "state." Third, the "raw material of government," he thought, was the activity of men and the processes through which this activity flowed in legislation, administration, and adjudication. Behaviour and process were to become the focus of much of the interest in political science in the 1950s, after Bentley's single work on politics was given new currency just before the outbreak of World War II. He was considerably before his time in presaging the end of preoccupation with institutional and descriptive analysis. He acknowledged the leadership of Gumplowicz, who, in his view, had "taken the most important step toward bringing out clearly the nature of the group process," discarding the "individual as a causal factor in society," and insisting "that all social movements are brought about by group interaction."

Although the effort of Bentley to develop an objective, value-free analysis of politics had no initial consequence, other movements toward this goal were more immediately successful. The principal impetus was provided by what became known as the Chicago school in the mid-1920s and thereafter. The leading figure in this movement was Charles E. Merriam, who in 1925 published *New Aspects of Politics,* a book that argued for a reconstruction of method in political analysis, urged the greater use of statistics in the aid of empirical observation and measurement, and postulated that out of the converging interests of politics, medicine, psychiatry, and psychology might come "intelligent social control." The basic political datum for Merriam at this stage of his thinking was "attitude"; hence his reliance upon the insights of psychology for a better understanding of politics. These ideas were not entirely new, since Graham Wallas, an Englishman, had said in *Human Nature in Politics* (1908) that a new political science should be based upon quantitative methods and that serious attention should be given to the psychological elements ("human nature") in political activity, including nonrational acts and the exploitation in political life of subconscious nonrational inferences. The American political scientist and journalist Walter Lippmann had expressed much the same view in *Public Opinion* (1922). One of those in the Chicago group who carried the connection between politics and psychology quite far was Harold Lasswell, in his *Psychopathology and Politics* (1930). In *Power and Personality* (1948) he fused the Freudian categories of the earlier work with subsequent writings on power.

Influential works of Merriam and Lasswell

These two leading expositors of the Chicago school, Merriam and Lasswell, both published books at about the same time that gave a central place to the phenomenon of power in the empirical study of politics. Merriam published *Political Power* in 1934 and Lasswell *Politics: Who Gets What, When, How* in 1936. Merriam undertook to show how power came into being, to describe what he called the credenda, miranda, and agenda of authority (which he tended to equate with power), the techniques of power holders, the defense available to those over whom power is wielded, and the dissipation of power. Lasswell's 1936 work was a naturalistic description of "influence and the influential." Although both were cast in the empirical mode, the second was more successful in this regard than the first, which tended to be abstract and rhetorical. A truly empirical work of the Chicago school that had considerable significance in the development of academic political science was Charles E. Merriam and Harold F. Gosnell's work, published in 1924, on *Non-Voting, Causes and Methods of Control,* which used sampling methods and survey data. Since then, certainly one of the most successful achievements in empirical political science has been the study of voter behaviour and election results. Although members of the Chicago school insistently professed an interest in value-free political science, they were characterized by two normative predilections—their acceptance of the values of the democratic system and their earnest attempts to improve it through their writings.

By 1945 political science in the United States was much more than the concern for institutions, law, formal structures of public government, procedures, and rules that it had earlier been. There was by then also a considerable body of writing on processes—on the dynamics as well as the statics of public governance. There were works on pressure groups and lobbies, on the "invisible government" operating behind public authority, on actual bureaucratic processes as distinguished from the rules of administrative procedure, on bosses and political parties, and on ethnic influences in the behaviour of the electorate. All of these works signified a turning away from formality and the development of a growing interest in the factual realities of political behaviour.

Developments in Europe. Outside the United States the development of political science was less quantitative and behavioral. In France the main tradition in the study of political institutions was still legal rather than sociological. Constitutional law and history were abundant, and politics in the main was viewed from legal perspectives, though not entirely so. As early as 1913 one writer, André Siegfried, had introduced the geographical and historical study of elections in *Tableau politique de la France de l'Ouest sous la troisième République,* and the development of French sociology in the works of Émile Durkheim and others had applications in political science.

In England, although there was little development of a quantitative science of politics, a substantial contribution was made toward the formulation of political philosophies that centred attention upon the group basis of politics. Particularly notable were the works of J.N. Figgis, G.D.H. Cole, Ernest Barker, and Harold Laski. In Sweden, Herbert Tingsten in his work *Political Behaviour: Studies in Election Statistics* (1937) gave currency in the title to what was to be the main development in political science after World War II.

METHODOLOGICAL CONSIDERATIONS IN CONTEMPORARY POLITICAL SCIENCE

Behavioralism. In American political science since the end of World War II, the behavioral persuasion has been the dominant one. A former president of the American Political Science Association has attributed the rapid development of the behavioral approach to six causes: the inspiration of the Chicago school; the immigration to the United States in the 1930s of large numbers of European scholars (particularly Germans) with backgrounds in European sociology, who stressed the relevance of sociology to politics; the movement of many political scientists into administrative and political positions during World War II; the influence of foundation support in the encouragement of research in political behaviour; the increasing development of the survey method in certain political studies, such as voter behaviour; and the missionary work of the Social Science Research Council under leadership sympathetic to behavioralism.

Etiology of the behavioral approach

Although the term behavioralism has been freely used in political-science writings, there is in fact confusion as to whether it is a field of study, a method, or an approach. One American political scientist, Heinz Eulau, in *The Behavioral Persuasion in Politics* (1963), has said that the behavioral persuasion "is concerned with what man does politically and the meanings he attaches to his behavior," and he has suggested that researchers cannot afford to get tangled up in problems of definition. Another American, Robert Dahl, has said that it is a mood or even "the scientific outlook." The term behavioral, then, may be merely a term having distinctiveness, weight, and value for a certain time only, since it seems primarily to signify that phase in the quarter century after World War II during which there was a significant revival of interest in empirical studies in politics, a movement strong enough to establish at least a partnership with the traditional approaches, although some of its advocates have gone so far as to say that their science has made traditional approaches outdated.

Systems analysis. Contemporary political science does not display as much coherence as earlier modes of inquiry (coherence was one of their principal virtues), but it does exhibit a refreshing complexity, of both concept and method. Although much effort has been spent on the production of empirically derived evidence in a scatter of fields and a diffusion of subjects and although critics have scoffed at studies that they have said only prove the obvious, the product of 25 years of behavioralism has been voluminous and often stimulating in quality. Common in this output is the assumption that politics is process, the ceaseless interaction of individuals and groups on each other in a flow of activities in and around public governance. The most commanding concept devised to fit this flow with form has been the concept of the system. In this view the focus of political science is not the individual in solitude nor the multitude nor even the group. The American political scientist David Easton's *The Political System* (1953), a seminal work in empirical political theory, conceived of the political system as part of the total social system. For Easton the political system comprised all those activities having to do with the formulation and execution of social policy; that is, "policy-making process." The identifying criterion of political behaviour for Easton was the "authoritative allocation of values" for the society, by which he distinguished his sense of the subject matter of political science from that of Lasswell, who had argued that political science was concerned with the distribution and contet of patterns of value throughout the society. In this conception of system, which is not unrelated to the conception of system in physics and biology, linkages are found between the system and its environment. Inputs (demands) flow into the system and are converted into the outputs (decisions and actions) that constitute the authoritative allocation of values for the society; that is, the distribution of rewards in wealth, power, and status that the system may provide.

Variants of systems theory

There are various versions of systems theory, although all suppose the existence of the system and more or less consciously pattern it after systems models in other disciplines. There has thus been developed a view of systems, based upon engineering models and ideas, that rejects the view of Easton that there can be a general theory of political systems. Another theorist, drawing on cybernetics, has viewed the political system as a communications net. Although expositors of systems concepts generally suppose they are working with scientific theory, critics have said that they are merely creating new taxonomies. What is certain is that systems theory has introduced new and—to more traditional scholars in the field—unusual terms into the vocabulary of politics. Thus, instead of such terms as the state and sovereignty, the new idiom speaks of systems, inputs and outputs, feedbacks, circular loops, netwoks, legitimacy symbols, information storage and retrieval, political socialization, interest articulation and aggregation, cluster blocs, zero-sum games, macropolitics and micropolitics, and much else drawn from the concepts and languages of other sciences and from statistics, with hopeful assertions about the possibility of making politics predictable.

Although decision making as a research field in political science may be analytically regarded as part of the process by which inputs are converted into outputs in systems theory, it also has an origin independent of systems theory. Interest in game theory provided a stimulant to studies of decision making, which already had an established place in the lore and lexicon of politics and administration. In decision theory the paradigm is the rational actor, and the research problem is the ascertainment of the most efficient means to effect given goals, which may or may not be rational. Investigators into decision making are generally less concerned with the system as a whole than with the activity that they regard as the most important part of it, namely, the way in which decisions are or should be made according to rational calculation. A body of writing on decision making by judges has earned a sub-classification as "jurimetrics."

Interest groups, elites, and political parties. Studies of interest groups, elites, and political parties also have their own independent origins, although they, too, have been brought within the framework of systems analysis. Interest groups and political parties, for example, have been described as agencies for the articulation and aggregation of interests, which in turn provide inputs (demands) for the political system to convert into outputs (decisions and actions). But interest-group analysis antedates the advent of behavioralism. The modern concern for the subject starts perhaps with studies of prohibition and other pressure groups during the 1920s. More generalized and theoretical treatments of interest groups and political parties resulted in part from a revived interest in the work of Bentley in the 1930s and 1950s. The study of elites was begun at least as early as 1936 (Lasswell's "influentials"), and it came to the forefront in the 1950s in community studies made by political scientists in Atlanta, Chicago, New York, New Haven, and elsewhere. Such studies had long been familiar in sociology, but they acquired special significance in political science, because democratic values seemed threatened by the possible existence of elites. Despite declarations of their concern to establish a value-free science, American political scientists have thus tended to cleave to the traditional democratic ethic, in which elites are presumed to have no place.

Analysis of political attitudes and voting behaviour. What some have called the behavioral revolution had its greatest successes in the analysis of public opinion, political attitudes, and electoral behaviour. Especially in the period after World War II, the refinement of statistical techniques in public-opinion polling, the analysis of voter behaviour, and the development of new research concepts have brought the study of opinions and attitudes closest to the goal of the scientific outlook and some considerable distance from the mark made by Merriam in the 1920s. The Survey Research Center at the University of Michigan has become an important national centre for the collection of data on elections and voter behaviour.

CURRENT TRENDS

Reactions against "scientism"

The effort to establish a value-free, objective political science in the United States in the two decades after World War II has won what is doubtless a permanent place for the scientific study of politics, but it has also bred a critical reaction. In the late 1960s opponents of "scientism" rejected what they felt was the increasing subjection of spontaneity and human values to determinisms in every aspect of life and argued that political science was an example of the pervasiveness of technology and of a search for rationality in a social complex that might be irrational and out of control. Although political science had developed skillful and sensitive techniques for the quantification of data termed political, there is no orthodoxy on the scope of political science nor on the delimitation of separate areas of research. Quite different outcomes emerge from the basic assumptions as to the focus of the discipline, whether it be power, government, system, process, decision making, or policy formulation. The failure of the discipline to settle the question of identity has made for a creative and inventive exploration of many avenues of research, but the achievement of a unified general theory of political behaviour on which there is common consent still lies ahead.

Even outside the United States, non-normative political science is not extensive, although it is not unknown. England has made many creative contributions to political theory and law throughout its history, as have other European countries, but it has also produced substantial works that can be classified as positive (*i.e.,* non-normative) political science. Among these are the studies of R.T. McKenzie and D.E. Butler in the field of political parties, voter behaviour, and pressure groups, S.E. Finer on interest groups, and U.W. Kitzinger on German elections. Notable work has been done in France in the field of political parties by Maurice Duverger and François Goguel. Work has also been done on the political socialization of children in French schools by Charles Roig and François Billon-Grand. In Denmark, beginnings were made in the systematic study of political science in 1959 with the founding of the Institute of Governmental Studies at the University of Århus. In Finland the first extensive studies on voting behaviour appeared in 1956.

In Japan there was a flowering of social sciences after World War II, but political science at first did not grow at the same pace as other social sciences because of lack of agreement about both subject matter and method, a difficulty felt from the beginning by the man recognized to be the founder of Japanese academic political science, Onozuka Kiheiji, who published *Principles of Political Science* in 1903. With the opening of the behaviour of public officials and institutions to scientific scrutiny, however, there has been much political inquiry that would qualify as political science, including, for example, systems analysis and works on political culture, political development, and process and behaviour.

Political science in Communist countries

Although in the past the objective study of political subjects by researchers in Communist regimes has been difficult, if not impossible, a somewhat more permissive policy in some countries has led to what may be the beginnings of scholarly political science. The most advanced political science is to be found in Poland and Yugoslavia, and Romania and Czechoslovakia have come to recognize political science as a discipline. The Soviet Union does not yet sanction political science, but scholars do conduct empirical research, which was endorsed in 1962 by the U.S.S.R. Academy of Sciences under the term "concrete sociological investigations." The greatest movement has been in the conduct of public-opinion polls using advanced Western techniques. Interest in the development of political science was evidenced in the publication in 1969 of *Politicheskaya nauka v SShA: Kritika burzhuaznykh Kontseptsy vlasti* ("Political Science in the U.S.A.: A Critique of Bourgeois Conceptions of Power"), by V.G. Kalensky. Although there is no officially sanctioned political science, there is a Soviet Association of Political Sciences, which has sent delegates to the meeting of the International Political Science Association. Despite its title, the Soviet Association of Political Sciences is most heavily oriented toward the state and the law, and its members have been critical of what they regard as the anti-Marxist bias of bourgeois political science. One of its members, however, F.M. Burlatsky, published a major article in *Pravda* in 1965 calling for the establishment of a genuine political science in which the findings would emerge from the data.

In Yugoslavia a political-science association was established in 1951, and in 1962 a faculty of the political sciences was established at the University of Zagreb. In Czechoslovakia a political-science association was formed in 1965 and became a member of the International Political Science Association; and in 1968 a political-science association was formed in Romania. Political science in Yugoslavia has tended to centre on traditional divisions of the discipline, such as political theory, comparative government, and international relations. In Poland, political science has centred on the study of political behaviour, on community power structures, on voter behaviour, and on public opinion. Techniques of considerable sophistication have been employed.

Political science is one of the means by which people seek to understand the human condition and man's fate. Or, as Aristotle believed, politics is the most important of human activities, and the sovereign science. For 24 centuries, at least, the greatest intellects and scholars have striven to state the universal elements of just order in human affairs. None has succeeded in this, so far, hopeless ambition, although most have contributed some special insight and added to the common wisdom. The effort to achieve magisterial comprehension will doubtless continue; and the search will change direction as experience requires, with the aid that new perceptions, concepts, and methods will provide. Progress toward establishing general laws, however, may never be as steady or as swift in political science as it has been in laboratory sciences like physics; for, as Albert Einstein once said, politics is more difficult than physics. (Ea.L.)

Study of international relations

HISTORICAL DEVELOPMENT OF THE STUDY OF INTERNATIONAL RELATIONS

Three contemporary forces go far to account for the impressive growth of scholarly studies of international relations and foreign policy in the 20th century. An autonomous academic discipline has emerged, related to geography, history, law, sociology, psychology, general political science, philosophy, and other fields, yet belonging to no one of these. The first impelling force was noted above: the growing demand to find better, less dangerous, more effective means of guiding relations among peoples, societies, governments, and economies. A second force is the result of the monumental upwelling of intellectual activity in modern times based on the belief that systematic observation and inquiry will dispel ignorance and serve the betterment of mankind. One sentence by Diderot characterizes this force as well today as when it was written in the 18th century: "Everything must be examined, everything must be shaken up, without exception and without circumspection." The third force is the consequence of the popularization of political affairs, including one of its most important sectors, foreign affairs. Only late in the 19th century did the traditional view that foreign and military matters should remain an exclusive preserve of rulers and special elites yield to the opposite belief that such matters constitute an important concern and responsibility of all the people. This popularization of international relations made logical the idea that education should include instruction in foreign affairs and that knowledge in the field should be advanced in the interests of public control over international political and military matters. The experience of World War I, the war to make the world safe for democracy, strengthened the conviction that not enough was known about international relations and that the universities should reduce ignorance in the field through more research and teaching.

The popularization of foreign and military affairs

Between the two world wars. A strong impulse toward the development of international studies in universities came in the 1920s. New centres, institutes, and schools devoted to teaching and research in international relations were founded. Courses were organized and general textbooks on the subject began to appear. Private organizations were formed, and large grants of philanthropic funds were channelled to the support of scholarly journals, to the advancement of citizenship in world affairs through special training institutes, conferences, and seminars, and to the stimulation of university research.

Initially, three subject areas commanded the most attention. All three had roots in the period of World War I. In the revolutionary upheavals at the end of the war, great portions of the government archives of imperial Russia and imperial Germany were opened and made public in a series of documentary publications. Very exciting scholarly work began to appear that pieced together the hitherto-unknown history of prewar alliances, secret diplomacy, and military planning. These materials were integrated to provide explanations of the origins of World War I. The two decades between the two great wars were the heyday of diplomatic history, and the most famous of the students of international affairs were historians. With great ingenuity and industry, they presented the world with superb examples of the art and science of diplomatic history.

The second subject that captured attention was bound up with the hope and expectation of a new world order in the making through the League of Nations. Some of the schools of international relations that were founded in the

1920s had the explicit purpose of preparing civil servants for what was expected to be the dawning age of international government. Thus the genesis and organization of the League, the history of earlier plans for international federations, and the analysis of the problems and procedures of international organization and international law were investigated with enthusiasm.

The third study of consequence during the early part of the interwar period was an offshoot of the peace movement and was concerned with scholarly investigations of international warfare: its cause, its costs, and its sociological and psychological aspects. In addition to the data and the interpretations dredged up in the study of war, the interest in the question "why war?" brought a host of new social scientists—economists, sociologists, and psychologists—into active participation in international studies for the first time. They were pioneers in what later came to be known as the "behavioral approach" to international relations.

Realpolitik in the 1930s

The breakdown of the League, the rise of the aggressive dictatorships, and the coming of World War II in the 1930s caused a reaction against the international government and peace-inspired themes in the study of international relations. Idealism and moralism were criticized, and "realism" became the new thought in the field. The image was built at that time that the first stage of academic development of international studies was the handiwork of starry-eyed idealists and peace visionaries who ignored the hard facts of international politics. This characterization is untrue, the fact being that the scholarship on world affairs of the '20s and the '30s was extensive and sound in the organization of data and in the development of some fundamental concepts.

In the European tradition since early modern times, the knowledge of international relations had been loosely ordered in two branches of learning. The first is diplomatic history, which has been considered to reflect the variety of political experience, the particularity of events, and the contingencies in the actual practices of diplomacy and war. The second is international law, which has been viewed as registering the "residue of history"—the fundamental principles of conduct, the uniformities in international phenomena, and the permanent aspects of practice. The effect of the new field of international relations was to broaden the traditional organization almost beyond recognition.

New areas of inquiry during the interwar period

Some of the topics that today are considered novel and of recent origin were being explored vigorously in the two interwar decades; by the time of World War II, they already had acquired large bibliographies. It is instructive to recall a few of those topics in order to correct the stereotype that moralist teachings were then entirely dominant: the relationship of problems of racial and ethnic minorities to international affairs, the effects of the population explosion on foreign policies, the linkage between raw materials and other of the "life-support systems" of the planet with the actions of nations, the effects of imperialism and colonialism, the strategic aspects of international relations including the effects of geographical location and space on military power and the influence on governments of what has come to be called the "military-industrial complex," the economic inequalities of nations, and the role of public opinion, national differences, and cultural orientations in world affairs. If these studies tended to be short on theory and long on description, nevertheless the topics investigated remain relevant.

Certain individual scholarly contributions of the 1930s deserve particular notice because they were forerunners of what was to be developed after World War II. Harold D. Lasswell was making explorations of the relationships between world politics and the psychological realm of symbols, perceptions, and images. Abram Kardiner and his associates were laying the groundwork for a psychoanthropological approach to the analysis of national behaviour and culture, which later became a popular but short-lived theory of international relations. Frederick L. Schuman was producing foreign policy analyses that synthesized analytic comment with accounts of current international events. Schuman thus set the style that is still followed by government interpreters of foreign policy developments and by the news analysts of world affairs.

Quincy Wright was leading one of the first team research projects in the field and was investigating numerous aspects of international behaviour in a very broad approach to the study of war. Carl J. Friedrich, Frederick L. Schuman, Harold Sprout, Nicholas Spykman, E.H. Carr, Brooks Emeny, and others were developing the main lines of analysis of what became the power-politics explanation of international relations.

Some 30 years later, one begins to appreciate that the definition of the study of international relations and the widening of its scope were the fundamental contributions of the scholars of the interwar period. Many of the innovators of the 1930s found their services enlisted by governments during World War II for work in intelligence, propaganda, and political analysis. In this respect, the war stimulated systematic social-science investigations of international phenomena. On the other hand, World War II became a divide for academic international relations. The war made a drastic change in the agenda of world politics. The postwar intellectual climate shifted away from many of the earlier interests, emphases, and problems. There was a readiness in the early postwar years for an analysis that would cut through the details of studies of myriads of international topics and that would provide a focussed view of the fundamental nature of international politics. An intellectual hunger for theory existed.

The postwar ascendancy of political realism. Hans Morgenthau's *Politics Among Nations,* first published in 1948, met this need for theory. Writing in 1959, Stanley Hoffmann expressed what was, in all likelihood, the opinion of most students of international relations:

Hans Morgenthau and the "realist" theory of power politics

> The theory which has occupied the center of the scene in this country [the United States] during the last ten years is Professor Morgenthau's "realist" theory of power politics.

At the time of this writing, the influence of the Morgenthau text continues to be strong in most countries outside the Communist world. The realist theory still requires the attention of new students of international relations. A reader is best advised to explore the theory at its source. *Politics Among Nations* remains an impressive study; it is clearly conceived and well argued.

At the heart of the realist theory is the concept of interests. Politics is defined as the struggle for power, whether in domestic or international settings. The struggle for power is part of human nature and takes form in society according either to the competition or to the alignment of interests. Collaboration occurs when parties find their interests are coinciding. Rivalry, competition, and conflict result from the clash of interests. Accommodations are possible through the application of political skill.

In an international system composed of sovereign nation-states, the survival of both the states and the system depends on the intelligent pursuit of national interests and on the realistic calculations of national power. As long as the state system persists, the only truly constructive way to participate in international politics is through skilled diplomacy. Religious and ideological crusades threaten the ruin of both the individual states and the system, and disasters follow from attempts to reform nations toward the ideal of universal trust and cooperation.

Thus the realist theory of power politics was brought forward in the late 1940s to stand guard against idealists: those who would think and act in the visionary ways of moralism and legalism in world affairs. No impressive new formulation of political idealism appeared to carry the challenge to the realist position, and the "great debate" of realism versus idealism gradually faded from the scene.

Many scholars of international relations neither opposed nor accepted the power-politics theory. Some simply were engrossed in other aspects of international-relations teaching and research. Large sums of money were made available in the 1950s for the development of foreign-area studies, and general theoretical concerns played little part in the growth of area specialization. Other scholars agreed with Morgenthau's statement that theory and research should have a "concern with human nature as it actually is, and with the historic processes as they actually take place," but they did not believe that the realist conceptualization provided a sufficient explanation of observed international behaviour.

The behavioral decade: mid-1950s to mid-1960s. An important new influence was the arrival in the field of a number of fresh ideas, conceptualizations, models, and paradigms that were loosely identified in ensemble as behavioral theory. The new movement distracted attention from the realist–idealist question. The unanticipated appearance in the mid-1950s of a large number of possible alternative ways to organize international data and to orient inquiries in international relations soon appeared to threaten the very foundations of scholarly communication. Simply to list a number of the conceptual innovations suggests the reason for the anxiety that the discipline might lapse into complete incoherence. In addition to power theory, there appeared a welter of theories, each with its distinctive label: decision making, system, conflict, deterrence, capabilities, field, communication, integration, development, environmental, cognitive, and, finally, game theory. Much of the intellectual effort of the "behavioral decade" went into the task of attempting to understand, compare, interpret, and integrate all these ideas. The scholarly goal of the period became to build an integrated framework of theories—to carry out conceptual mapping.

The behavioral aim to integrate the mass of ideas

To describe the efforts at theoretical integration that were made and the problems that were encountered would require book-length treatment. Suffice it to say that the comparing and integrating of the elements of theory turned out to be difficult. The more the matter was investigated, the more the specialists in theory questioned the necessity of arriving at one comprehensive structure of theory. The international realities that theory is supposed to reference and explain are varied and diverse, so why should one not expect that a number of separate theories would be needed to account for different parts and aspects of international relations?

Increasingly, explanations that trace the forces of international relations to any single source have been seen to be unsatisfactory. The struggle for power among nations, for example, can be accepted as a fact in past and current international politics, but to theorize that all other factors are subordinate to or dependent upon this one is to exclude too much that is important and interesting in international phenomena. Similarly, the formulation that asserts that the character of nations and, hence, the character of their participation in international relations are determined by child-rearing practices is simple, appealing; at the same time, however, it is unacceptable because it theorizes a single cause where multiple causation prevails. The Communist theory that international relations are the historical expression of the class struggle also falls into the single-source classification of theory.

The attitude nurtured in the behavioral decade emphasized the necessity of recognizing social multidimensionality and, therefore, multiple causation and the multiple forms of explanation and theory. Under this perspective, one might well conclude that the concept of the struggle for power among nations and the idea that international relations is really a manifestation of a global struggle between social classes both relate to human conflict and that a theory of conflict should encompass these as well as other conflict interpretations. Indeed, the anticipated integrating effect on theory and research in international relations was a main motive in advancing general conflict theory in the '60s, but it is important to add that conflict theory came to coexist with integration theory and game theory, both of which approach some conflict phenomena from different conceptual angles.

Development of the theory of conflict

By the end of the behavioral decade, the multiple-theory perspective had come into the ascendancy in North America and western Europe. International-relations scholars in the Soviet Union who were following closely the Western literature in the field reported with some satisfaction that Western theory had become eclectic and was in disarray in contrast to their own, which they declared to be based on the unified theory of the science of Marxism. At the same time, they gave indications that some of the Western trends in international-relations thinking were being introduced in socialist countries and that a muted contest between traditional and behavioral approaches was underway.

By the 1970s only the realist theory of power politics survived as a relatively simple and comprehensive explanation of international politics in a conceptual environment that otherwise had become pluralistic and complex to the extent that "theory" could no longer be outlined quickly or conveniently in a classroom or before an audience. A situation had begun to develop that put the question of a single theory of international relations in about the same class with that of a single theory of biology.

Contemporary perspectives in international relations. *Foreign-policy and international-systems perspectives.* One important clarification developed from the effort of the behavioral decade to bring the various theories together in a unified structure. It was the consensus that the academic organization of the discipline has two principal parts. It is convenient to call each part a perspective on the subject. The parts are called the foreign-policy and the international-system-analysis perspectives.

The foreign-policy perspective covers many theory and research interests. Fully generalized, it embraces all the inquiries that look into the domestic sources of external or international phenomena. Thus, a study of any set of traits, structures, or processes arising within a national society or polity that can be demonstrated to determine or influence importantly how that society or polity participates in international relations belongs to the foreign-policy perspective. The decision-making approach to international politics meets the requirement, for example. The analysis of the information that decision makers use, of their perceptions and motives, of the influences on them exerted by public opinion, and of the organizational settings in which they operate is a manifestation of the foreign-policy perspective. Studies that seek to relate the facts of wealth and power of a nation to its international status and role provide other illustrations.

Comparative foreign-policy analysis is an area of theory and research effort that first appeared in the mid-1960s. Its objectives are, first, to examine the data of domestic sources of external conduct country by country using standard criteria of data selection and analysis, and, then, to compare across countries for generalized findings on foreign-policy performances. When the details of the domestic sources and the external performances have been compared, theories about the domestic–external linkages and about the groupings of countries according to the types of linkage are expected to develop. The comparative foreign-policy approach, so described, develops theory through inductive research procedures.

The second perspective is that of international system analysis. Whereas foreign-policy analysis concentrates on the actors, international system analysis is preoccupied by the interaction. The term "interaction" suggests challenge and response, give and take, or move and countermove. Diplomatic histories feature the narratives of action and response in international situations and interpret the meanings in the exchanges. The theory of the balance of power is an example of an international-system conception. Explanations and descriptions of bargaining in international negotiations fit the perspective and so also do studies of arms races and other escalating processes. A model of the international trading system would be an example of a structural approach to international system studies, while an examination of how and why a coalition of states disintegrates would represent a process approach to international system analysis.

The theorist of international systems may gain a general outlook on the phenomena he studies through a system perspective, but he must bring in a certain amount of empirical detail when he identifies the components, relationships, and environments of the system that is the object of his theoretical inquiry. System theory does not turn out to be any single formulation; it is more in the nature of a conceptual anchoring point for a variety of specific formulations. Thus one system theorist may define the components of his system as geographical nation-states related according to rules and political structures, whereas another theorist may define the components of his system as nonterritorial, nonstate transactional units only partly related by the influences wielded by national governments.

The general system perspective. Although the theoretical development of the system idea may lead to very diverse outcomes, another more general concept—that of open, adaptive systems—may provide the most promising approach to a comprehensive understanding of the dynamics of relations among nations. Without imposing any single school of thought or any single interpretation of world affairs, it has a loosely unifying effect on the outlook of students of the field. The so-called general system perspective on international relations may be compared to the map of a little-explored continent. Outlines, broad features, and a continental delineation are not in question, but everything else remains in doubt, is subject to controversy, and awaits exploration. One commentator has remarked that general system theory is not really a theory but instead is "a program or a direction in the contemporary philosophy of science." (From Anatol Rapoport, "Systems Analysis," in *International Encyclopedia of the Social Sciences,* vol. 15, p. 452, 1968.)

General concept of open, adaptive systems

As noted above, the quest for theoretical unification during the behavioral decade resulted in the widespread acceptance of two perspectives—a foreign-policy approach and an international-system approach. The general concept of open, adaptive systems provides a conceptual bridge connecting the two perspectives and creates a loose bond joining many of the diverse theoretical formulations prevailing in the field. An examination of the line of general thinking that builds the bridge and provides the bond is, therefore, well worth attention.

One begins to think about almost any open, adaptive system that involves human beings as a living system. If the system is living, its pervading characteristic is activity. Acting units, however they are recognized and defined, are doing things, participating in events, carrying forward processes, and creating effects. The effects created by activity include progressive influences on the actors. That is to say, the acting unit is immune neither to the effects of its own participation nor to the participatory influences generated by other acting units. It is this situation that establishes the condition of the openness of a system. Streams of influencing activity contain two kinds of processing: the first kind is regular, which is to say, governed by rules and repetitive in form, and the second kind is unexpected, irregular, and variant.

It is the second type that stimulates change and that initiates the special processing called adaptation. A living system is open and is able to adapt, however, only if it in some way has gained access to information on the state of affairs that joins it with its environment and, further, in some way has achieved means to direct and to change its stream of activity. Thus in addition to the fundamental concept terms of openness and adaptation, the general system perspective incorporates the ideas of communication and corrective action or, more generally, of communication and control.

Another view adds the observation that there is a systematic deception afflicting human understanding when the unit of action—the actor or the initiator—is seen as an entity. For social phenomena, the system perspective advises us that we make a grievous error when we identify individual persons, groups, organizations, nations, and so on as separate, uniquely named things. Recognized correctly, all these are nothing other than organized and interlocking activity flows, proceeding across historical time in somewhat regularized fashion. John Doe is to be described, literally, as an organized packet of active processes in exchange with an environment and utilizing communication and control to survive and adapt. The same description fits Japan: an incredibly complex network of related action much more than a place, a people, or a name.

Not only does the general system perspective urge that the "actor" be considered as a configuration of activity but it also prompts a recognition that, most often, the configuration itself has a hierarchical organization. In fact, the use of the term "system" is ordinarily an effort to convey the meaning that smaller organized activity flows serve larger activity flows and that the functional linking of subordinate parts to the operating whole is the process that defines what an actor really is. Thus the conventional explanation is that any recognizable living system is made up of related "components" and that each component, when examined at its own level, is found to be a functioning system in its own right. It also may be called a "subsystem." Enough of the system conceptualization has been suggested here to show next how these fundamentals have been translated for the purposes of international-relations theory.

In world history, the Earth has been populated by hundreds of separate and relatively isolated social systems. Each system, as a somewhat ordered stream of interrelated activities, has had exchanges with its particular environment in its own time and in its own way and has employed communication and control capabilities either to succeed at adaptation or to fail at it and, therefore, perish. What today is called international relations is that sector of exchange of a social system with its particular environment that has to do with interlinked action flows to and from other separate social systems. The knowledge problem of international relations is to understand, describe, and explain such flows to and from social systems from their source to their termination. Social systems thus consist of complexes of "internal" interacting subsystems or components, only some of which connect their processes with the action flows between that social system and another.

Internal processes and interactive processes of nations

The next logical thought then is that there can be only two sources of international conduct. One originates in the activity complexes within each participating social system and the other arises from the effects of the interlinked action flows to and from the participating social systems that together make up the membership of an international system. Hence, two basic perspectives of theory and research on international relations are distinguished by the primary attention given either to the origins of conduct arising from internal processes or to origins of conduct arising from the effects of the processes of exchange between social systems—or, to put it more succinctly, from the effects of interaction.

Obviously, the two perspectives are related because each casts its beam on a sector of a whole phenomenon of action and interaction. If he expects to be understood, the theorist must specify not only the basic perspective he wishes to emphasize but also particular identifications of the "units" of action, the kinds of action flows, and the linkages of processes and effects in this particular conception of the system that concerns him. In a multidimensional social world, the theorist can exercise his choice of a focus of inquiry in many different ways. Multiple theories are the outcome that may be expected from the introduction of the concept of general systems into the study of international relations.

The rise of quantitative research and computers. The emphasis on theory building and theory integrating after World War II was perhaps a reaction against the emphasis on methodology inherent in the historical and institutional studies of the 1930s. In turn, the renewed interest in empirical research after the behavioral decade probably was a reaction against the excessively intensive theorizing. Only part of the renewal of interest in data gathering and data analysis can be accounted for in this way, however. The marked increase in quantitative data studies after the mid-1960s has resulted from the direct influence of the computer. Computerization came late to the field of international relations; how to use the machines to good advantage was not apparent at first because most observations on the conditions of international relations are recorded in narrative or literary form.

Uses of computers in international relations

The first big discovery about computers in the study of international relations was that they could be made to serve the function of marvelously efficient librarians. By placing data collections under computer management, researchers could improve their investigations by incorporating vastly larger collections of facts in their work. Facts that had a narrative, nonquantitative form could be included in storage and retrieval systems almost as readily as numerical data. The second revelation was that by systematic coding and counting, many kinds of nonquantitative data could be transformed into quantitative indicator information and then be condensed and evaluated by mathematical and statistical procedures. Vistas on new research possibilities were opened and exploratory studies were made in a number of directions. Through quantitative, computer-assisted research detailed, painstaking examination of historical records has again become important. The new research also gives promise of bringing the interests of academic theory and research into closer accord with the interests of government analysts and the practitioners of diplomacy.

RELATIONS BETWEEN SCHOLARSHIP AND ACTION IN FOREIGN AFFAIRS

The differences between the interests of scholars and practitioners of international affairs have appeared to be more prominent than the similarities. A steady concern of scholars has been to avoid both the fact and the reputation of serving as apologists for official foreign policies. One of the first teachings of the idealism of the discipline's founding period at the time of World War I was to maintain an attitude of future-oriented reform of the international order. Existing systems had lesser importance according to the early objectives of the field. The principle of scientific detachment in social-science research also has contributed to the scholarly effort to evaluate international events and developments from a global standpoint rather than from the perspective of any one country's foreign-policy position. On the other side, practitioners have been inclined more to indifference than to hostility in their attitudes toward academics in international relations. They frequently have professed that for their day-to-day work they have found little of value in the theory and research contributions of the field. One looks in vain, therefore, for many signs of direct influence in either direction. An indirect and subtle exchange has occurred, nevertheless, and it has had importance in the direction of foreign affairs.

Mutual influences of government and academia

New international programs or new foreign-policy directions undertaken by governments often have attracted so much interest in the universities that research programs have been initiated and special subfields of international studies have resulted. Examples of such subfields include: (1) "national development" stimulated by the foreign-assistance programs to aid underdeveloped countries; (2) the "area studies" that emerged from World War II because of the problems that Western governments had in finding knowledgeable personnel who understood the languages, histories, cultures, geography, and politics of Asian, African, and Latin-American regions; and (3) "national security," resulting from the heavy influence of military factors on foreign policy and especially from the nuclear-control problem in the post-World War II period.

The indirect influence flowing in the other direction from university studies into governmental thinking may be traced in a number of noteworthy examples during the past two decades: first, the realist formulation of the power-politics theory has filtered into the foreign-policy thinking of the U.S. government to such an extent that most foreign-policy decisions have been defended by arguments of national interest and power calculation and opposing views have been discounted for reflecting insufficient "hard-nosed" realism. Another interesting example was the preoccupation of Pres. John F. Kennedy's administration with the details of the process of foreign-policy decision making in crisis periods. This orientation can be traced to the popularity of decision-making theory advanced in the academic field in the preceding decade. One other example has been identified in the Soviet Union's revision of foreign-affairs doctrine to accommodate the deterrence theory of nuclear defense. Concepts drawn from the literature of American "defense intellectuals" can be readily seen in Soviet works on military doctrine.

Although the past relations of scholars and practitioners of international relations have been relatively mild and indirect, the future may change this customary relationship. The implications of system theory can be expected to percolate slowly but deeply into the understanding of both officials and the public. As this occurs, there will be greater appreciation of how truly complex and far-reaching the flows of international events are. The problems of survival and adaptation will be seen in a new light and the realization of how interdependent modern nation-states are—how much they are bound to a common fate—should lead to more careful, better researched, and more deliberate formulations of foreign policy. The search for more accurate appraisals of the state of the world in its international aspect, which is the vocation of theory and research in academic international relations, should bring officials and researchers together.

Demand for data

On the other side of the relationship, the surge of growth in quantitative, computer-assisted studies in the universities is creating an increasing appetite for more frequent and more exact reports on the state of the world and in more and more of its aspects. The academic community in whatever country one might name lacks the resources and trained manpower to satisfy this growing appetite for data. The demand could well lead to increased responsiveness of the data-gathering services of national governments and to large reductions both in secrecy and in governmental interference with free public communication among nations. The current practice of withholding large amounts of information about developments in the flows of international events is an atavism left over from the era of aristocratic diplomacy. Today it has become a threat to the survival of mankind.

If the data on the conditions and relationships of the world's social systems (now made more manageable and more available for immediate use through computer systems) are brought into full and untrammeled circulation, the academic field of international relations and the governmental analysis and planning agencies will have much more in common than they have had. The end result should be the development of entirely new, more competent, and more realistic approaches to the formulation and execution of foreign policies. (C.A.McC./Ed.)

Comparative law

HISTORICAL DEVELOPMENT OF COMPARATIVE LAW

The expression comparative law is a modern one, first used in the 19th century when it became clear that the comparison of legal institutions deserved a systematic approach, in order to increase understanding of foreign cultures and to further legal progress. From early times, however, certain scholars and researchers have made use of the comparative technique, conscious of the advantages to be gained.

Ancient roots of law. In the 6th century BC according to legend, the Greek lawgiver Solon, faced with the task of compiling the laws of Athens, gathered together the laws of various city-states. Similarly, in the 5th century BC, a Roman commission was reported to have consulted the statutes of the Greek communities in Sicily before giving Rome the famous *Laws of the Twelve Tables.* Aristotle, in the 4th century, is said to have collated the constitutions of no less than 158 city-states in his effort to devise a model constitution. Thus, from ancient times it would seem that those wishing to set up a just system have sought inspiration and example from abroad. The true expansion of comparative law, however, was hindered by a number of obstacles—such as the parochialism of social groups, contempt for foreigners, or "barbarians," and belief in the sacredness or everlasting inviolability of inherited legal rules.

Although certain practices and institutions that crept into Roman law undoubtedly originated in the imperial provinces, Roman legal science took no cognizance of comparative law. Nor can the medieval universities in Europe be said to have displayed great concern for comparative law. Over the centuries, their interest was limited to Roman law, supplemented in certain areas or modified to some extent by canon law. While members of the first school of thought (called glossators) confined themselves to the task of elucidating the meaning of the Roman codes of law, their successors (the postglossators) undertook the systematic arrangement and adaptation of that law to prevailing social conditions. At no time was there an effort to compare laws. The customary laws that one found here and there could hardly hold any interest for scholars labouring to give society a model of ideal justice and to discover or elucidate a higher law above man's making. Indeed, in their opinion, local laws were no more than rubbish and evidently doomed to decay. To compare these local practices would have been a waste of time; to compare them with Roman laws would have been almost indecent.

The ancient search for "higher" or ideal law

Role of judges. Such contempt was not characteristic of the attitude of the judges and lawyers whose duty it was to administer justice, mainly by applying the customary law. Their material contained areas of uncertainty and required adaptation to social needs. In the work of ascertaining the content of a custom, and in the task of filling the gaps of customs, judge or lawyer had to consider which customs to allow to prevail. In so doing, he had to decide whether one custom was more just than another and how far he should go in introducing concepts of ideal justice (based on Roman law) that were being promoted by the universities. Two processes were thus at work: the elimination of conflicting local customs and the acceptance and rejection of elements of Roman law. With regard to the first process, the comparative aspects of the work took place behind the scenes, and consequently the results of melding the different local or municipal laws are known, but the reasoning leading to the result is not. With regard to the second process, by contrast, certain publications place the act of comparison in full view. This was particularly noticeable in England, where some writers—such as Sir John Fortescue in the 15th century and Saint-Germain in the 16th—took upon themselves the comparison of common law and Roman law, and in 1623 Sir Francis Bacon suggested to James I that a work be drafted comparing English and Scots law, as a preliminary step toward the unification of the two systems.

19th-century beginnings. Despite the occasional use of the comparative technique, nevertheless, comparative law itself was not recognized as a separate branch or as a fundamental technique of legal science until the 19th century. In particular, it played no part in legal education. It was quite unthinkable that the pursuit of justice should be taught by reference to a host of customary rules that were incomplete, sometimes archaic, and generally regarded as barbaric. A foundation of ethical and political principles rather than sociological considerations, an appeal to reason rather than a study of human behaviour or judicial precedent—these were deemed the true criteria of progress.

End of dualism between ideal law and everyday practice

With the coming of the 19th century, codification of the law put an end to the dualism existing in many countries between an ideal system, as taught in the universities, and the laws that were applied in everyday practice. Codification of those everyday laws gave them the status of a national law, thoroughly purged of anachronisms and arranged in a systematic manner. That codified law became the cornerstone of legal education. This promotion of local customs, regarded henceforth as being fully consonant with natural justice, may be considered as the underlying cause of the appearance and rise of comparative law.

In short, the attitude toward comparative law tends to change when a country makes its national law the object of legal study and law students begin contrasting it with foreign counterparts. In Europe this dawning change was evident early in the 19th century. Legal periodicals were founded in Germany in 1829 and in France in 1834 to further a systematic study of foreign law. In France, the civil and mercantile laws of modern states were translated with "concordances" referring to the corresponding provisions of the French codes; and in England in 1850–52, Leone Levi published a work entitled *Commercial Law, Its Principles and Administration; The Mercantile Law of Great Britain Compared with Roman Law and the Codes or Laws of 59 Other Countries.*

A chair of comparative legislation was set up in 1831 in the Collège de France; and this was followed, in 1846, by a chair of comparative criminal law in the University of Paris. In 1869 the Société de Législation Comparée was founded in France, followed in 1873 by the Institut de Droit International and the International Law Association. In England, the Society of Comparative Legislation was founded in 1895, and the Quain Professorship of Comparative Law was created at London University in 1894. Similarly, chairs in comparative law were founded and projects in foreign law undertaken all over the continent of Europe, but with particular vigour in France.

First International Congress of Comparative Law

International efforts. The 19th century drew to a close with an important event—the meeting of the First International Congress of Comparative Law in Paris in 1900. Experts from every part of Europe delivered papers and discussed the nature, aims, and general interest of comparative law. Particular emphasis was laid on its role in the preparation of a "common law for the civilized world," the contents of which would be laid down by international legislation. The stress, however, was on comparative legislation and codification because (with the exception of one English jurist) the congress had attracted only jurists from continental European nations, all of which had coded law, in contrast to English customary, or common, law. Consequently, the idea of an enacted world law was the natural outcome of its proceedings.

The upheavals resulting from World War I (1914–18) prompted a change in direction. From then on, European interest began to extend beyond the continental systems themselves, first, to those of the common-law countries (chiefly England and the United States), then still further afield to the Socialist systems, and finally, after 1945, to the laws of the newly independent states of Asia and Africa. The new territory for legal study that has thus been opened up has resulted in references to comparative law, rather than to comparative legislation.

METHODOLOGICAL CONSIDERATIONS IN CONTEMPORARY COMPARATIVE LAW

The world contains a vast number of national legal systems. The United Nations brings together representatives of some 127 states, but these states are far outnumbered by legal networks, since not all states—notably federal ones—have accomplished unification within their own frontiers. It is thus an enormous task to try to compare the laws of all the different jurisdictions. This problem, however, should not be overly magnified. Differences between the diverse systems are not always of the same order; some are sharp; others are so closely similar that a specialist in one branch of a legal "family" often may easily extend his studies to another branch of that family. For this reason, one can distinguish two types of research in comparative law. The exponent of "microcomparison" analyzes the laws belonging to the same legal family. By observing their differences, he will decide whether they are justified and whether an innovation made in one country would have value if introduced elsewhere. The researcher pledged to "macrocomparison," on the other hand, investigates those systems differing most widely from each other in order to gain insight into institutions and thought processes that are foreign to him. For the "pure jurist," concerned mainly with legal technicalities, microcomparison holds the greater attraction; whereas macrocomparison is the realm of the political scientist or legal philosopher, who sees law as a social science and is interested in its role in government and the organization of the community.

Two types of research in comparative law

Microcomparison. Microcomparison demands no particular preparation. The specialist in one national system is usually qualified to study those of various other countries of the same general family. His chief need is access to bibliographical material. In the United States, each state has its own statutes and, to some purposes, its own common law. Thus, the American lawyer must be a microcomparatist as he takes the 50 state systems and the federal law into daily account in his practice of the law. The same is true, to a large extent, of the Australian, or Indian, or Kenyan lawyer, who must take into account not only his own national system but also the laws of England and of other common-law jurisdictions in the Commonwealth. Whatever can be said of the common-law systems holds largely true for the Roman-law and Socialist families. French comparative law students encounter little difficulty in contrasting the laws of certain countries, so long as they confine their study to French, German, Italian, and Dutch law, which are related in tradition and structure and serve a similar type of society.

Macrocomparison. The situation differs greatly in consideration of macrocomparison. Here no comparison is possible without previously identifying and thoroughly mastering the fundamentals of the law systems as they differ from place to place. The jurist must, as it were, forget his training and begin to reason according to new criteria. If he is French, English, or American, he must recognize that in some folk societies of the Far East, the upright citizen never crosses the threshold of a courtroom and acknowledges no subjective rights; instead, the citizen's behaviour is governed by rites handed down from his ancestors, ensuring him the approval of the community. Likewise, if the Western jurist is to understand the law of the Muslim or Hindu, he must realize that the law is contained in rules of conduct laid down by a religion for its followers, and for its followers only. These rules, creating obligations and not rights, rank above all worldly matters and, in particular, are not to be confused with the regulations that a national government may, at a given time, enact and ratify. Further, in comparing his system of law with that of a Communist nation, the Westerner must remember that on no account does the citizen of a Marxist-Leninist state regard the rule of law as an ideal for society. Far from it, for his dream is to see law—which to him is synonymous with injustice and coercion—wither away in an affluent society founded on human solidarity and fellowship. A considerable shifting of legal gears is necessary before a French or German jurist can grasp the vital importance that the English or American lawyer traditionally attaches to the concept of due process and the rules of evidence; in continental eyes, procedural rules take second place to substantive law.

Mastering fundamentals of a foreign system

The specialist of macrocomparison also picks out the structural differences existing between certain systems. Accordingly, the Anglo-American lawyer must be aware of the importance of the distinction between public and private law—between law involving the state and law involving only individuals. The jurist in a Roman-law country must, conversely, appreciate the significance of the concepts of common law (unwritten customary law of various kinds) and equity (the use of injunctions and other equitable remedies), neither of which have counterparts in his own system. The lawyer from a centralized country must familiarize himself with the distinction between federal law and the laws of secondary jurisdictions (states, provinces, cantons, and so forth)—a distinction that is of fundamental importance in many countries. If he is from a nation like England or France that acknowledges the sovereignty of the national parliament, he must give due weight to the prominence of constitutional law in countries that permit courts to review the constitutional validity of legislative acts—especially in countries such as the United States and the Federal Republic of Germany. The jurist in a "bourgeois" country must appreciate the policy of collective ownership of means of production in Socialist states.

Parliamentary versus constitutional law

Classification of families of law. The terms microcomparison and macrocomparison, reflecting the language of economics, are in keeping with the idea that legal systems can be grouped into families, such as common-law, Roman, and Socialist. But it must be acknowledged that the number of identifiable families and the appropriate classification of a given system are questions always open to argument. The legal system of a given country, for instance, may exhibit some features that relate it to a particular family and others that may escape that classification. Such blurring of distinctions is particularly true of law in countries of Africa and the Middle East, where certain sectors of the law have been transformed by Western ideas (as in criminal and mercantile law and procedure) leaving other sectors (such as personal status, family law, and land law) faithful to traditional principles of the region. The phenomenon is not peculiar to those countries, however.

Wide differences also may be detected between legal systems that are commonly regarded as belonging to the same family. American law, for instance, without hesitation is ranked as a member of the common-law family; yet countless differences set it apart from English law, in large part because the United States has a federal and England a unitary system of government.

PURPOSES OF COMPARATIVE LAW

Historical and cultural comparisons. First of all, there has been a tendency to view comparative law from the standpoint of its value to the historical study of legal decision making—a consideration that was responsible for establishing the first chairs of comparative law in 19th-century Europe. Ideas regarding the place of law in society and the nature of the law itself—whether divine or secular, whether dealing with substantive or procedural rules—obviously become appreciably clearer when comparative law is joined to historical research. Indeed, to some extent historical background may aid in forecasting the future of certain national systems

A closely related consideration prompts many Western jurists, political scientists, and sociologists to acquaint themselves with non-Western methods of reasoning. Comparative studies reveal that the citizen of some countries of Asia and Africa looks upon the concept of a just social order with thoughts and feelings far removed from those of Western man. The notions of a rule of law and of rights of the individual—fundamental to Western civilization—are not wholly recognized by those societies that, faithful to the principle of conciliation and concerned primarily with harmony within the group, do not favour excessive Western-style individualism or the modern Western ideal of legal supremacy. Thus comparative law may enable statesmen, diplomats, and jurists to understand foreign points of view, and it may frequently help to create better international understanding.

Understanding another culture's reasoning

Commercial uses. Comparative law may be used for essentially practical ends. The businessman, for instance, needs to know what benefits he may expect, what risks he may run, and generally how he should act if he intends to invest capital or make contracts abroad. It was with this purpose in mind that the first French institute of comparative law was set up in Lyon in 1920; its mission was to instruct French legal advisers on foreign trade. It was this practical aspect that also encouraged the growth of comparative law in the United States, where the essential aim of the law school has been usually to turn out practitioners; and one need hardly mention the strong link in Germany between big industry and the various institutes of comparative law. Sometimes it is said that studies with such a focus should not be considered a part of comparative law, but practical considerations certainly have helped to finance and promote the development of comparative legal studies in general.

Aid to national law. The improvement of national legislation was the prime consideration during the 19th century in countries that were codifying or recodifying their legal systems. Numerous later additions to the Code Napoléon, drawn up in 1804, for instance, were of foreign origin. Many other nations, of course, followed France's lead and introduced into their own systems elements of the French Napoleonic codes and institutions of French public law. It is well worth noticing that a book on French administrative laws was published in German by Otto Mayer before Mayer felt himself able to write a textbook on German administrative law.

The foreign inspiration of a number of legal rules or institutions is a well-known phenomenon, sometimes so all-embracing that one speaks of "reception"—reception, for instance, of the English common law in the United States, Canada, Australia, India, and Nigeria; reception of French law in French-speaking Africa, Madagascar, Egypt, and Indochina; reception of Swiss law in Turkey; and reception of both German and French law in Japan, along with even some reception of American common law. The study of comparative law has found a special place in countries where such a reception has occurred.

Use in international law. In modern times the spirit of nationalism has often tended to frustrate the development of an international law that would overcome individual national differences. One task facing statesmen and jurists is to inject new life into this effort, adapting it to the exigencies of the modern world. Those engaging in international trade, for instance, do not know with certainty which national law will regulate their agreements, since the answer depends to a large extent on a generally undecided factor—namely, which national court will be called upon to decide the questions of competence. Thus, the sole lasting remedy would seem to be the development of an international law capable of governing all legal questions outside the jurisdiction of a single state. Such a project can succeed only through the medium of comparative law.

(R.Da./Ed.)

BIBLIOGRAPHY. The most comprehensive and detailed bibliographical background for the history of the social sciences is to be found in the *Encyclopaedia of the Social Sciences,* 15 vol. (1930–35), still valuable, especially in its historical sections, despite its age; and in the *International Encyclopedia of the Social Sciences,* 17 vol. (1968). The best individual work, by now a classic in the history of ideas, on the period leading up to the emergence of the individual social sciences, is PRESERVED SMITH, *A History of Modern Culture,* 2 vol. (1930–34, reprinted 1962). JAMES WESTFALL THOMPSON, *A History of Historical Writing,* 2 vol. (1942), is useful in this respect also. On the age immediately preceding the rise of the social sciences, the best study by far is LESTER G. CROCKER, *Nature and Culture: Ethical Thought in the French Enlightenment* (1963), and *An Age of Crisis: Man and World in Eighteenth Century French Thought* (1959), recommended to be read or consulted in that order. The best general work on the history of the social sciences and the history of social philosophy in the West is HARRY ELMER BARNES and HOWARD BECKER, *Social Thought from Lore to Science,* 2nd ed., 2 vol. (1952). The first volume deals with the issues and problems of all the social sciences, with major emphasis on the period following the breakup of the Middle Ages; the second volume is primarily concerned with sociology in the 20th century.

ERIC J. HOBSBAWM, *The Age of Revolution: 1789–1848* (1962), is an important and fascinating treatment of the social, cultural, and intellectual aspects of the age in which the individual social sciences emerged in western Europe. ROBERT A. NISBET, *The Sociological Tradition* (1966), although concerned primarily with sociology, deals with the specific ways in which the ideologies and themes of the democratic and industrial revolutions became translated into social theory. The same author's *Social Change and History* (1969) deals in detail with the incorporation of the theory of social evolution into the social sciences of the 19th century. For the rise and development of the individual social sciences in the 19th and 20th centuries, the following works are recommended. (*Anthropology*): ROBERT H. LOWIE, *The History of Ethnological Theory* (1937); and MARVIN HARRIS, *The Rise of Anthropological Theory: A History of Theories of Culture* (1968). (*Economics*): ERICH ROLL, *A History of Economic Thought,* 3rd ed. rev. (1954); and the extremely readable ROBERT HEILBRONER, *The Worldly Philosophers: The Lives, Times, and Ideas of the Great Economic Thinkers,* 3rd ed. (1967). (*Political science*): GEORGE SABINE, *A History of Political Theory,* 3rd ed. (1959), best on the three centuries preceding the 20th; FRANCIS W. COKER, *Recent Political Thought* (1934), excellent for the early 20th century; and LEO STRAUSS and JOSEPH CROPSEY (eds.), *History of Political Philosophy* (1963). (*Sociology*): Barnes and Becker, referred to above, for detailed information on the history of sociology in the 19th and early 20th centuries; Nisbet, also referred to above, dealing with the relation between political ideologies and the currents of sociological thought in the late 19th century; and LEWIS A. COSER, *Masters of Sociological Thought* (1971), a very good general history of sociology in 19th- and 20th-century Europe and America. (*Social psychology*): FAY BERGER KARPF, *American Social Psychology: Its Origins, Development and European Background* (1932), the best account of social psychology in the 19th and early 20th centuries, and THOMAS C. WIEGELE, *Biology and the Social Sciences* (1982), on the effect biological research is having on the social science disciplines.

For authoritative and easily available accounts of very recent developments in the social sciences, the reader is advised to turn to the articles in this encyclopaedia on the individual social sciences both for content and for bibliographical sources. The *International Encyclopedia of the Social Sciences,* already referred to above, is indispensable in this respect.

Anthropology: Histories of anthropological science include T.K. PENNIMAN, *A Hundred Years of Anthropology,* 3rd ed. rev. (1965), which covers all of anthropology; and P. MERCIER, *Histoire de l'anthropologie* (1966), which covers only cultural anthropology. The principal textbooks are M.J. HERSKOVITS, *Man and His Works* (1948); F.M. KEESING, *Cultural Anthropology: The Science of Custom* (1958); and J. POIRIER (ed.), *Ethnologie générale* (1968). In cultural anthropology—aside from two works by the "fathers" of the discipline, L.H. MORGAN, *Ancient Society* (1877); and E.B. TYLOR, *Anthropology: An Introduction to the Study of Man and Civilisation* (1881)—some of the classic general works are FRANZ BOAS, *The Mind of Primitive Man* (1911), and *Race, Language and Culture* (1940); BRONISLAW MALINOWSKI, *A Scientific Theory of Culture, and Other Essays* (1944); A.R. RADCLIFFE-BROWN, *Structure and Function in Primitive Society* (1952); A.L. KROEBER (ed.), *Anthropology Today* (1953); CLAUDE LEVI-STRAUSS, *Anthropologie structurale* (1958; Eng. trans. 1963); G. BALANDIER, *Anthropologie politique* (1967; Eng. trans. 1971); and M. MAUSS, *Oeuvres* (1968). Studies of individual peoples that have become classics include W.H.R. RIVERS, *The Todas* (1906); M. GRANET, *Fêtes et chansons anciennes de la Chine* (1919; Eng. trans., *Festivals and Songs of Ancient China,* 1932); BRONISLAW MALINOWSKI, *The Argonauts of the Western Pacific* (1922), and *Coral Gardens and Their Magic* (1935); A.R. RADCLIFFE-BROWN, *The Andaman Islanders* (1922); MARGARET MEAD, *Coming of Age in Samoa* (1928); P. SCHEBESTA, *Bambuti, die Zwerge vom Kongo* (1932); FRANZ BOAS, *The Religion of the Kwakiutl Indians* (1930); R.F. FORTUNE, *Sorcerers of Dobu: The Social Anthropology of the Dobu Islanders of the Western Pacific* (1932); R.W. FIRTH, *We, the Tikopia: A Sociological Study of Kinship in Primitive Polynesia* (1936); M. GRIAULE, *Masques Dogons* (1938); M.J. HERSKOVITS, *Dahomey: An Ancient West African Kingdom* (1938); M. LEENHARDT, *Gens de la grande terre* (1937); E.E. EVANS-PRITCHARD, *The Nuer* (1940); and E.R. LEACH, *Political Systems of Highland Burma: A Study of Kachin Social Structure* (1954).

Sociology: Among older titles of major significance, generally considered classics, are W.G. SUMNER, *The Folkways* (1907, reissued 1940); C.H. COOLEY, *Human Nature and the Social Order* (1902, reissued 1967); G.H. MEAD, *Mind, Self, and Society,* ed. by C.W. MORRIS (1934); E. FARIS, *The Nature of Human Nature* (1937); R.E. PARK and E.W. BURGESS, *Introduction to the Science of Sociology* (1921, reissued 1969); and W.F. OGBURN, *Social Change with Respect to Culture and Original Nature,* new ed. (1950; suppl. ch., 1964). A comprehensive summary of early sociological theory is available in H.E. BARNES (ed.), *An Introduction to the History of Sociology* (1948). The following titles provide excellent coverage of the main directions and subfields of recent and contemporary sociology: R.E.L. FARIS (ed.), *Handbook of Modern Sociology* (1964); R.K. MERTON, L. BROOM, and L.S. COTTRELL (eds.), *Sociology Today,* 2 vol. (1959, reissued 1965); G.D. GURVITCH and W.E. MOORE (eds.), *Twentieth Century Sociology* (1945); J.G. MARCH (ed.), *Handbook of Organizations* (1965); and N.J. SMELSER and J.A. DAVIS (eds.), *Sociology* (1969). Two influential general texts are L. BROOM and P. SELZNICK, *Sociology,* 4th ed. (1968); and G.A. LUNDBERG *et al., Sociology,* 4th ed. (1968). Leading books treating communities and societies as wholes are TALCOTT PARSONS, *Societies: Evolutionary and Comparative Perspectives* (1966); I.T. SANDERS, *The Community,* 2nd ed. (1966); R.K. MERTON, *Social Theory and Social Structure,* rev. ed. (1957); G.C. HOMANS, *Social Behaviour: Its Elementary Forms* (1961); G.E. LENSKI, *Human Societies: A Macrolevel Introduction to Sociology* (1970); and A.L. STINCHCOMBE, *Constructing Social Theories* (1968). The technical and statistical methods used in sociology are presented in E.F. BORGATTA (ed.), *Sociological Methodology* (1968); and J.H. MUELLER, K.F. SCHUESSLER, and H.L. COSTNER, *Statistical Reasoning in Sociology,* 2nd ed. (1970). ERNEST GELLNER, *Soviet and Western Anthropology* (1980), illustrates differing approaches to anthropological study.

Social psychology: Excellent general textbooks include EDWIN P. HOLLANDER, *Principles and Methods of Social Psychology,* 2nd ed. (1971); and ROGER W. BROWN, *Social Psychology* (1965). A comprehensive account of research is G. LINDZEY and E. ARONSON (eds.), *Handbook of Social Psychology,* 2nd ed., 5 vol. (1968–69). A useful account of theories is MARVIN E. SHAW and PHILIP R. CONSTANZA, *Theories of Social Psychology* (1970). Social psychology approached through detailed analysis of social interaction is described in MICHAEL ARGYLE, *Social Interaction* (1969). A. PAUL HARE, EDGAR F. BORGATTA, and ROBERT F. BALES (eds.), *Small Groups* (1965), is a most useful collection of papers. Research on social psychology in industry is described in BERNARD M. BASS, *Organizational Psychology* (1965). Social behaviour in relation to personality is dealt with in EDGAR F. BORGATTA and WILLIAM W. LAMBERT (eds.), *Handbook of Personality Theory and Research* (1968). A so-called symbolic interactionist approach is represented by GREGORY P. STONE and HARVEY A. FARBERMAN, *Social Psychology Through Symbolic Interaction* (1970); and by ERVING GOFFMAN, *Relations in Public* (1971).

Criminology: Important titles include H. MANNHEIM, *Comparative Criminology,* 2 vol. (1965); M.E. WOLFGANG and F. FERRACUTI, *Il comportamento violento* (1966; Eng. trans., *The Subculture of Violence,* 1967); H. MANNHEIM (ed.), *Pioneers in Criminology,* 2nd ed. (1972); L. RADZINOWICZ, *Ideology and*

Crime (1966), a survey containing much historical material; J.T. SELLIN and M.E. WOLFGANG, *The Measurement of Delinquency* (1964), an attempt to find an operational definition of serious offenses and to construct a "crime index"; H. MANNHEIM, *Social Aspects of Crime in England Between the Wars* (1940); F.H. MCCLINTOCK and N.H. AVISON, *Crime in England and Wales* (1968), partly an updating of the preceding work; L.T. WILKINS, *Social Deviance* (1964), an attempt to bridge the gap between social research and social action from the viewpoint of the statistician; H. MANNHEIM and L.T. WILKINS, *Prediction Methods in Relation to Borstal Training* (1955); S. and E. GLUECK, *Predicting Delinquency and Crime* (1959), and *Ventures in Criminology* (1964); and A.K. BOTTOMLEY, *Criminology in Focus: Past Trends and Future Prospects* (1979).

Economics: The best introduction to methodological issues in economics is still L. ROBBINS, *The Nature and Significance of Economic Science,* 2nd ed. (1935). A more modern position on empirical testing is well conveyed by M. FRIEDMAN in *Essays in Positive Economics* (1953). The best way to learn about economics is to browse through an introductory textbook such as P.A. SAMUELSON, *Economics: An Introductory Analysis,* 8th ed. (1970); R.G. LIPSEY and P.O. STEINER, *Economics,* 2nd ed. (1969); or A.A. ALCHIAN and W.R. ALLEN, *University Economics* (1964). Good histories of economic thought include H.L. HEILBRONER, *The Worldly Philosophers,* rev. ed. (1953), a pleasure to read; O.H. TAYLOR, *A History of Economic Thought* (1960), arranged chronologically; and E. WHITTAKER, *A History of Economic Ideas* (1940), arranged topically. J.A. SCHUMPETER, *History of Economic Analysis* (1954); and M. BLAUG, *Economic Theory in Retrospect,* 2nd ed. (1968), are advanced references. Excellent articles on the great economists, as well as superb but sometimes quite difficult essays on the leading branches of modern economics may be found in the *International Encyclopedia of the Social Sciences,* 16 vol. (1968). Further readings for advanced students are those published by the American Economic Association, *Survey of Contemporary Economics, I,* ed. by H.S. ELLIS (1948); *A Survey of Contemporary Economics, II,* ed. by B.F. HALEY (1952); and *Surveys of Economic Theory,* 3 vol. (1965). Each branch of economics has its own text. (*Microeconomics*): C.E. FERGUSON, *Microeconomic Theory,* 2nd ed. (1969); K.J. COHEN and R.M. CYERT, *Theory of the Firm: Resource Allocation in a Market Economy* (1965). (*Macroeconomics*): G. ACKLEY, *Macroeconomic Theory* (1961); T.F. DERNBURG and D.M. MCDOUGALL, *Macroeconomics,* 3rd ed. (1968). (*Development economics*): C.P. KINDLEBERGER, *Economic Development,* 2nd ed. (1965); B.H. HIGGINS, *Economic Development,* 2nd ed. (1968). (*Public finance*): R.A. MUSGRAVE, *The Theory of Public Finance* (1959). (*Monetary economics*): A.G. HART and P.B. KENEN, *Money, Debt and Economic Activity,* 3rd ed. (1961); L.V. CHANDLER, *The Economics of Money and Banking,* 5th ed. (1969). (*International economics*): C.P. KINDLEBERGER, *International Economics,* 4th ed. (1968). (*Labour economics*): A.M. CARTTER and F.R. MARSHALL, *Labour Economics: Wages, Employment, and Trade Unionism* (1966); L.C. HUNTER and D.J. ROBERTSON, *Economics of Wages and Labour* (1969). (*Industrial organization*): E.H. CHAMBERLAIN (ed.), *Monopoly and Competition and Their Regulation* (1954); J.S. BAIN, *Industrial Organization* (1959). (*Agricultural economics*): T.W. SCHULTZ. *The Economic Organization of Agriculture* (1953), and *Transforming Traditional Agriculture* (1964). (*Growth economics*): R.G.D. ALLEN, *Macro-Economic Theory: A Mathematical Treatment* (1967), a fairly difficult book; most texts on macroeconomics and economic development devote a chapter or two to growth theory. (*Mathematical economics*): A.C. CHIANG, *Fundamental Methods of Mathematical Economics* (1967). (*Econometrics*): A.A. WALTERS, *An Introduction to Econometrics* (1968); C. CHRIST, *Econometric Models and Methods* (1966). T.W. HUTCHISON, *The Politics and Philosophy of Economics: Marxians, Keynesians and Austrians* (1981), a history of economic thought; and DANIEL BELL and IRVING KRISTOL (eds.), *The Crisis in Economic Theory* (1981).

Political science: Although works of classical political philosophy are both venerable and extensive, few of them qualify as modern political science because they are neither quantitative nor, in most respects, even empirical in tone and temper. ARISTOTLE'S *Politics* and MACHIAVELLI'S *The Prince* come closest to meeting empirical standards. AUGUSTE COMTE, *Cours de philosophie positive,* 6 vol. (1830–42; Eng. trans., *The Positive Philosophy of Auguste Comte,* 2 vol., 1853), and *Système de politique positive,* 4 vol. (1851–54; Eng. trans., *System of Positive Polity,* 4 vol., 1875–77), are seminal statements in the 19th century on a science of society. LUDWIG GUMPLOWICZ, *Grundriss der Sociologie* (1885; Eng. trans., *The Outlines of Sociology,* 1899; 2nd ed., 1963); and GUSTAV RATZENHOFER in *Wesen und Zweck der Politik,* 3 vol. (1893), argue the case for the primacy of groups in studies of the state. A useful summary statement of the sociologies of the 19th century is NICHOLAS S. TIMASHEFF, *Sociological Theory: Its Nature and Growth,* 3rd ed. (1967). A good general work on the efforts of German jurists in the 19th century to cope with the facts of federalism is RUPERT EMERSON, *State and Sovereignty in Modern Germany* (1928). Among Americans who either studied in Germany in the 19th century or showed the influence of new European thought in their work were THEODORE D. WOOLSEY, *Political Science; or, The State Theoretically and Practically Considered,* 2 vol. (1877); JOHN W. BURGESS, *Political Science and Comparative Constitutional Law* (1891); WOODROW WILSON, *The State: Elements of Historical and Practical Politics* (1889), and *Congressional Government* (1885, reprinted 1961); and FRANK J. GOODNOW, *Politics and Administration: A Study in Government* (1900, reprinted 1967). An Englishman who studied American institutions was JAMES BRYCE, *The American Commonwealth,* 3 vol. (1888; abridged ed., 1959).

An excellent statement of the content of political-science teaching in America is ANNA HADDOW, *Political Science in American Colleges and Universities, 1636–1900* (1939, reprinted 1969). The most notable precursor of the behavioral approach in the 20th century was ARTHUR F. BENTLEY, *The Process of Government: A Study of Social Pressures* (1908, reprinted 1949). Others were GRAHAM WALLAS, *Human Nature in Politics,* 4th ed. (1962); and WALTER LIPPMANN, *Public Opinion* (1922; paperback ed., 1965). Besides works of the Chicago School mentioned in the article, the following may be noted: CHARLES E. MERRIAM, *Chicago: A More Intimate View of Urban Politics* (1929, reprinted 1970); LEONARD D. WHITE, *The Prestige Value of Public Employment in Chicago* (1929); and HAROLD D. LASSWELL and DANIEL LERNER (eds.), *The Policy Sciences: Recent Developments in Scope and Method* (1951), an effort to bring scientific method to the study of choices in public policy. Support for the establishment of a value-free science of politics was also provided by STUART A. RICE, *Quantitative Methods in Politics* (1928, reprinted 1969), who wrote the first general work on the application of statistical methods to the study of politics; GEORGE E.G. CATLIN, *The Science and Method of Politics* (1927); and WILLIAM BENNETT MUNRO, *Invisible Government* (1928). A useful summary survey of political science around the world after the end of World War II is *Contemporary Political Science,* published in 1950 by the UNITED NATIONS EDUCATIONAL, SCIENTIFIC AND CULTURAL ORGANIZATION.

International relations: ADDA B. BOZEMAN, *Politics and Culture in International History* (1960), is the best guide to histories and concepts of past international systems. Other useful references that survey premodern international relations are COLEMAN PHILLIPSON, *The International Law and Custom of Ancient Greece and Rome,* 2 vol. (1911); RICHARD L. WALKER, *The Multi-State System of Ancient China* (1953); GEOFFREY F. HUDSON, *Europe and China: A Survey of Their Relations from the Earliest Times to 1800* (1931); FRANK M. RUSSELL, *Theories of International Relations* (1936); and SHMUEL N. EISENSTADT, *Political Systems of Empires* (1963). Accounts of the earlier development of the academic study of international relations may be found in EDITH E. WARE, *A Study of International Relations in the United States: A Survey for 1937* (1938); SIR ALFRED ZIMMERN, *The Study of International Relations* (1931); and GRAYSON L. KIRK, *The Study of International Relations in American Colleges and Universities* (1947). Representative of the work of the 1930s that widened the scope of the field are HAROLD D. LASSWELL, *World Politics and Personal Insecurity* (1935; paperback ed., 1965); FREDERICK L. SCHUMAN, *International Politics: An Introduction to the Western State System and the World Community,* 6th ed. (1958); C.K. LEITH, *World Minerals and World Politics* (1931, reprinted 1970); W.S. THOMPSON, *Danger Spots in World Population* (1929); PAUL RADIN, *The Racial Myth* (1934); NICHOLAS SPYKMAN, *America's Strategy in World Politics* (1942, reprinted 1970); CARL J. FRIEDRICH, *Foreign Policy in the Making* (1938); BROOKS EMENY, *The Strategy of Raw Materials* (1935); ABRAM KARDINER, *The Psychological Frontiers of Society* (1945); QUINCY WRIGHT, *A Study of War,* 2nd ed., 2 vol. (1965); and E.H. CARR, *The Twenty Years' Crisis, 1919–1939,* 2nd ed. (1946, reprinted 1964). The theory of political realism is expressed by HANS J. MORGANTHAU in *Politics Among Nations* (1948). The idealist attempt to answer the challenge of political realism may be traced in THOMAS I. COOK and MALCOLM MOOS, *Power Through Purpose: The Realism of Idealism as a Basis for Foreign Policy* (1954); and JOHN H. HERZ, *Political Realism and Political Idealism* (1951). Four books on the psychological and cultural aspects of the behavioral study of international relations providing an excellent introduction are J. DAVIS SINGER, *Human Behavior and International Politics* (1965); OTTO KLINEBERG, *The Human Dimension in International Relations* (1964); JOSEPH H. DE RIVERA, *The Psychological Dimension of Foreign Policy* (1968); and HERBERT C. KELMAN, *International Behavior* (1965). The most famous book of its time on a psychocultural interpretation of national behaviour was RUTH BENEDICT, *The Chrysanthemum and the Sword* (1946, reprinted

1967). RICHARD C. SNYDER, H.W. BRUCK, and BURTON SAPIN, *Decision-Making as an Approach to the Study of International Politics* (1954), is still basic reading on the decision-making approach, but see also the data in GLENN D. PAIGE, *The Korean Decision* (1968). The most convenient compilation of varied examples of the theory and research of the behavioral decade is JAMES N. ROSENAU, *International Politics and Foreign Policy,* rev. ed. (1969). For conflict theory and international applications of game theory, see KENNETH E. BOULDING, *Conflict and Defense: A General Theory* (1962); THOMAS C. SCHELLING, *The Strategy of Conflict* (1960); and ANATOL RAPOPORT, *Fights, Games and Debates* (1960). EDWARD MCWHINNEY, *Conflict and Compromise: International Law and World Order in a Revolutionary Age* (1981), studies the effectiveness of law in resolving disputes among nations.

An introductionto general system theory in the social sciences is WALTER BUCKLEY, *Sociology and Modern Systems Theory* (1967). MORTON A. KAPLAN, *System and Process in International Politics* (1957); and CHARLES A. MCCLELLAND, *Theory and the International System* (1966), take different approaches to the applications of general systems ideas in the study of international relations. Important material on computer applications is included in DAVIS B. BOBROW and JUDAH L. SCHWARTZ, *Computers and the Policy-Making Community; Applications to International Relations* (1968).

Comparative law: A good historical survey is H.C. GUTTERIDGE, *Comparative Law,* 2nd ed. (1949). J.H. WIGMORE, *A Panorama of the World's Legal Systems,* 3 vol. (1928; 1-vol. ed., 1936), gives a general account of 16 legal systems. The ASSOCIATION OF AMERICAN LAW SCHOOLS, *A General Survey of Events, Ideas, Persons and Movements in Continental Legal History* (1912), is a pioneer book, in which the legal traditions of various European countries are discussed by specialists from each country. R. DAVID and J.E.C. BRIERLEY, *Les Grands systèmes de droit contemporains* (1964; Eng. trans., *Major Legal Systems in the World To-day,* 1968), a more up-to-date book originally devised for students, it describes the problems and value of comparative law and provides information on civil, Socialist, common, and religious and traditional law, ending with a valuable bibliography. *The International Encyclopaedia of Comparative Law,* proposed 16 vol. (1971–), will, when completed, constitute a major source of information.

Social Structure and Change

Social structure and social change are general concepts used by social scientists, particularly in the fields of sociology and social and cultural anthropology. They are often conceived of as polarized twin concepts, social structure referring to permanence, social change to the opposite. The relationship between the two concepts is, however, more complicated. "Structure," for instance, does not necessarily indicate lack of change. Those features of a society, or any other social group, that are regarded as parts of its structure are always generated by dynamic processes. For example, the kinship structure of a given society (the typical composition of household units and the rules governing marriage and line of descent) is maintained by continuous changes in families, as marriages are concluded; children are born, grow up, and become adults; and people die. Second, although many social processes show a cyclical pattern—the formation, dissolution, and reformation of families being one example—social life never repeats itself completely. The kinship relations in one generation are never an exact replica of those in the previous one. The same processes that serve to maintain the social structure may also lead to social change and modification of the structure over a long period.

The concepts of social structure and social change pertain not only to basic characteristics of human social life but also to certain ideals and preferences. The structure, or order, of the society, generally regarded as harmonious and conducive to the general well-being, has also been seen as conflict-ridden and repressive. Similarly, social change has been conceived of both as progress and as decay, as emancipation on the one hand and as deviance from good tradition on the other. Such widely varying evaluations have influenced different theories concerning the nature of social structure and social change, and they continue to be reflected, to some extent, in present-day social thought.

This article is divided into the following sections:

SOCIAL STRUCTURE

The term structure has been used with reference to human societies since the 19th century. Before that time, it had been already applied to other fields, particularly construction and biology. Its biological connotations are evident in the work of several social theorists of the 19th and early 20th centuries, such as Herbert Spencer in England. He and others conceived of society as an organism, the parts of which are interdependent and thereby form a structure that is similar to the anatomy of a living body.

Structure compared to anatomy

The metaphor of construction is clear in the work of Karl Marx, where he speaks of "the economic structure [*Struktur*] of society, the real basis on which is erected a legal and political superstructure [*Überbau*] and to which definite forms of social consciousness correspond." This phrase expresses the Marxian view that the basic structure of society is economic, or material, and determines, at least to a large extent, the rest of social life, which is defined as spiritual or ideological.

Although social scientists since Spencer and Marx have disagreed on the concept of social structure, their definitions have certain elements in common. In the most general way, social structure may be defined as those features of a social entity (a society or group within a society) that have a certain permanence over time, are interrelated, and determine or condition to a large extent both the functioning of the entity as a whole and the activities of its individual members.

As may be inferred from this definition, several ideas are implicit in the notion of social structure. The concept expresses the idea that human beings form social relations that are not arbitrary and coincidental, but exhibit some regularity and persistence. The concept also refers to the observation that social life is not amorphous but is differentiated into groups, positions, and institutions that are interdependent, or functionally interrelated. These differentiated and interrelated characteristics of human groupings, although constituted by the social activities of individuals, are not a direct corollary of the wishes and intentions of these individuals; instead, individual choices are shaped and circumscribed by the social environment. The notion of social structure implies, in other words, that human beings are not completely free and autonomous in choosing their activities, but rather they are constrained

Ideas implicit in structure

by the social world they live in and the social relations they form with one another.

The social structure is sometimes simply defined as patterned social relations—those regular and repetitive aspects of the interactions between the members of a given social entity. Even on this descriptive level, the concept is highly abstract: it selects only certain elements from ongoing social activities. The larger the social entity considered, the more abstract the concept tends to be. What is considered as the social structure of a small group is generally much nearer to the daily activities of its individual members than that which is regarded as the social structure of a larger society. In the latter case the problem of selection is acute: what to include or not include as components of the social structure. The solution to the problem varies with the different theoretical views according to which characteristics of the society are regarded as particularly important.

Aspects of structure

Apart from these different theoretical views, some preliminary remarks on general aspects of the social structure of any society may be made. Most generally, social life is structured along the dimensions of time and space. Specific social activities take place at specific times, and time is divided into periods that are connected with the rhythms of social life—the routines of the day, the month, and the year. Specific social activities are also organized at specific places; particular places, for instance, are designated for such activities as working, worshiping, eating, or sleeping. Territorial boundaries delineate these places. These boundaries are defined by rules of property, which in any society structure the use and possession of scarce goods. In any society, moreover, there is a more or less regular division of labour. Yet another universal structural characteristic of human societies is the regulation of violence. The use of violence is everywhere a potentially disruptive force; at the same time, it is a means of coercion and coordination of activities. Human beings have formed political units, such as nations, within which the use of violence is strictly regulated and which, at the same time, are organized for the use of violence against outside groups.

In any society, furthermore, there are arrangements within the structure for sexual reproduction and the care and education of the young. These arrangements partly take the form of kinship and marriage relations. Finally, systems of symbolic communication, particularly language, everywhere structure the interactions between the members of a society.

Within the broad framework of these and other general features of human society, there is an enormous variety of social forms between and even within societies. Several theories have been developed to account for both the similarities and the varieties. In these theories certain aspects of social life are regarded as basic and, therefore, central components of the social structure.

Murdock's taxonomic scheme

Some social scientists use the concept of social structure as a device for creating an order for the various aspects of social life. Thus, the U.S. anthropologist George P. Murdock, in his *Social Structure* (1949), a comparative study of kinship systems, used the concept as a taxonomic scheme for classifying, comparing, and correlating aspects of kinship systems of different societies. In other studies, the concept is of greater theoretical importance; it is regarded as an explanatory concept, a key to the understanding of human social life. Some of the more prominent of these theories are reviewed here.

Structural functionalism. A.R. Radcliffe-Brown, a British social anthropologist, gave the concept of social structure a central place in his approach and connected it to the concept of function. In his view, the components of the social structure have indispensable functions for each other—the continued existence of the one component is dependent on that of the others—and for the society as a whole, which is seen as an integrated, organic entity. Radcliffe-Brown defined the social structure empirically as patterned, or "normal," social relations (those aspects of social activities that conform to accepted social rules or norms). These rules bind society's members to socially useful activities.

Structural functionalism was elaborated further by Talcott Parsons, a U.S. sociologist, who, like Radcliffe-Brown, was strongly influenced by the French social scientist Émile Durkheim. While Radcliffe-Brown focused on so-called primitive societies, Parsons attempted to formulate a theory that was valid for large and complex societies as well. For Parsons, the social structure is essentially normative; it consists of "institutionalized patterns of normative culture." Social behaviour is structured insofar as it conforms to norms, ranging from general ideas of right and wrong (values) to specific rules of behaviour in specific situations. These rules vary according to the positions of the individual actors: they define different roles, such as various occupational roles, or the roles of husband–father and wife–mother. Norms also vary according to the type of activities or sphere of life: they form clusters called social institutions, such as the institution of property or the institution of marriage. Norms, roles, and institutions are components of the social structure on different levels of complexity.

Relation of structure and norms

Theories of class and power. Parsons' work has been criticized for several reasons. One has been the comparatively meagre attention he paid to inequalities of power, wealth, and other social rewards. Other social theorists, including functionalists like the U.S. sociologist Robert K. Merton, have given these distributional properties a more central place in their concepts of social structure. For Merton and others, the social structure consists not only of normative patterns but also of the inequalities of power, status, and material privileges, which give the members of a society widely different opportunities and alternatives.

Social structure by class

In complex societies these inequalities define different strata, or classes, which form the stratification system, or class structure, of the society. Both aspects of the social structure, the normative and the distributive aspect, are strongly interconnected, as may be inferred from the observation that members of different classes often have different and even conflicting norms and values.

This leads to a consideration contrary to structural functionalism: certain norms in a society may be established, not because of any general consensus about their moral value, but because they are forced upon the population by those who have both the interest and the power to do so. To take one example, the "norms" of apartheid in South Africa reflect the interests and values of only one section of the population, which has the power to enforce them upon the majority. In theories of class and power this argument has been generalized: norms, values, and ideas are explained as the result of the power inequalities between groups with conflicting interests.

The Marxian view of structure

The most influential theory of this type has been Marxism, or historical materialism. The Marxian view is succinctly summarized in Marx's phrase that "the ideas of the ruling class are, in every age, the ruling ideas." These ideas are regarded as reflections of class interests and are connected to the power structure, which is identified with the class structure. This Marxian model, which was claimed to be particularly valid for capitalist societies, has met with several criticisms. One basic problem is its distinction between economic structure and spiritual superstructure, which are identified with social being and consciousness, respectively. This suggests that economic activities and relations are in themselves somehow not conscious, as if they were conceivable without knowing and thinking human beings.

Nevertheless, the Marxian model has become influential even among non-Marxist social scientists. The distinction between material structure and nonmaterial superstructure continues to be reflected in sociological textbooks as the distinction between social structure and culture. Social structure here refers to the ways people are interrelated or interdependent; culture refers to the ideas, knowledge, norms, customs, and capacities that they have learned and share as members of a society.

Structuralism. The concept of structure in the study called structuralism, as in structural functionalism and the class and power theories, is theoretical and explanatory. Unlike those other studies, however, it is not descriptive. The concept here refers to the underlying, unconscious regularities of human expressions, which are not observ-

Lévi-Strauss and underlying structure

able but explain what is observed. Claude Lévi-Strauss, a French anthropologist, derived this concept from structural linguistics as developed by the Swiss linguist Ferdinand de Saussure. Any language is structured in the sense that its elements are interrelated in nonarbitrary, regular, rule-bound ways; a competent speaker of the language largely follows these rules without being aware of doing so. The task of the theorist is to detect this underlying structure, including the rules of transformation that connect the structure to the various observed expressions.

According to Lévi-Strauss, this same method can be applied to social and cultural life in general. He constructed theories concerning the underlying structure of kinship systems, myths, and customs of cooking and eating. The structural method, in short, purports to detect the common structure of widely different social and cultural forms. The structure does not determine the concrete expressions; the variety of expressions it generates is potentially unlimited. The structures that generate the varieties of social and cultural forms ultimately reflect, according to Lévi-Strauss, basic characteristics of the human mind.

Structuralism became an intellectual fashion in the 1960s in France, where such different writers as Roland Barthes, Michel Foucault, and Louis Althusser were also regarded as representatives of the new theoretical current. Structuralism in this wide sense, however, is not one coherent theoretical perspective. The Marxist structuralism of Althusser, for example, is far removed from Lévi-Strauss's anthropological structuralism. The structural method, when applied by different scholars, appears to lead to different results.

The criticisms launched against structural functionalism, class theories, and structuralism indicate that the concept of social structure is problematic. Yet the notion of social structure is not so easy to dispense with, because it expresses ideas of continuity, regularity, and interrelatedness in social life. Other terms are often used that have similar, but not identical, meanings, such as social network, social figuration, or social system. The British sociologist Anthony Giddens has suggested the term "structuration" in order to express the view that social life is, to a certain extent, both dynamic and ordered.

SOCIAL CHANGE

Social change in the broadest sense is any change in social relations. In this sense, social change is an ever-present phenomenon in any society. In order to give the concept a more restricted meaning, it has been defined as change of the social structure. A distinction is made then between processes within the social structure, which serve, at least partially, to maintain the structure (social dynamics), and processes that modify the structure (social change). Because the concept of social structure does not have one generally accepted and unambiguous meaning, however, this distinction does not clearly determine which social processes belong to the field of social change.

The specific meaning of social change depends first of all on the social entity considered. Changes in a small group may be important on the level of that group itself, but negligible on the level of the larger society. Similarly, the observation of social change depends on the time span taken; most short-term changes are negligible if a social development is studied in the long run. Even if one abstracts from small-scale and short-term changes, social change is a general characteristic of human societies: customs and norms change, inventions are made and applied, environmental changes lead to new adaptations, conflicts result in redistributions of power.

Biological roots of change

This universal human potential for social change has a biological basis. It is rooted in the flexibility and adaptability of the human species—the near absence of biologically fixed action patterns on the one hand and the enormous capacity for learning, symbolizing, and creating on the other hand. The human biological constitution makes changes possible that are not biologically (genetically) determined. Social change, in other words, is only possible by virtue of biological characteristics of the human species, but the nature of the actual changes cannot be reduced to these species traits.

Historical background. Several ideas of social change have been developed in various cultures and historical periods. Three of them may be distinguished as the most basic: (1) the idea of decline or degeneration, or, in religious terms, the fall from an original state of grace; (2) the idea of cyclical change, a pattern of subsequent and recurring phases of growth and decline; and (3) the idea of continuous progress. These three ideas were already prominent in Greek and Roman antiquity and have characterized Western social thought from that time. The concept of progress, however, became the most influential idea, especially since the 18th-century Enlightenment. Social thinkers like Anne-Robert-Jacques Turgot and the Marquis de Condorcet in France and Adam Smith and John Millar in Scotland advanced theories on the progress of human knowledge and technology.

Social evolution as progress

Progress was the key idea in 19th-century theories of social evolution, and evolutionism was the common core shared by the most influential social theories of the century. Evolutionism implied that mankind as a whole progresses along one line of development; that this development is predetermined and inevitable, since it corresponds to definite laws; that some societies are more advanced in this development than other ones; and that Western society is the most advanced and therefore indicates the future of the rest of mankind. Auguste Comte, a French philosopher and sociologist, advanced a "law of three stages," according to which mankind progresses from a theological stage, which is dominated by religion, through a metaphysical stage, in which abstract speculative thinking is most prominent, and onward toward a positivist stage, in which scientific theories based on empirical research come to dominate.

The most encompassing theory of social evolution was developed by Herbert Spencer, who, unlike Comte, linked social evolution to biological evolution. According to Spencer, biological organisms and human societies follow the same universal, natural evolutionary law: "a change from a state of relatively indefinite, incoherent, homogeneity to a state of relatively definite, coherent, heterogeneity." In other words, as societies grow in size, they become more complex; their parts differentiate, specialize into different functions, and become, consequently, more interdependent.

Morgan's evolutionary rankings

Evolutionary thought also dominated the new field of social and cultural anthropology in the second half of the 19th century. Anthropologists such as Sir Edward Burnett Tylor and Lewis Henry Morgan classified contemporary societies on an evolutionary scale. Morgan ranked them from "savage" through "barbarian" to "civilized." Tylor postulated an evolution of religious ideas from animism through polytheism to monotheism. Morgan classified societies on the basis of the level of technology, or sources of subsistence, which he connected with the kinship system. He assumed that monogamy was preceded by polygamy, and patrilineal descent by matrilineal descent.

Marx and Friedrich Engels too were highly influenced by evolutionary ideas. The Marxian distinctions between primitive communism, the Asiatic mode of production, ancient slavery, feudalism, capitalism, and future socialism may be interpreted as a list of stages in one evolutionary development, although the Asiatic mode did not fit well in this scheme. Marx and Engels were impressed by Morgan's anthropological theory of evolution, which became evident in Engels' book *Der Ursprung der Familie, des Privateigentums und des Staats* (1884; *The Origin of the Family, Private Property and the State*).

The originality of the Marxian theory of social development lay in its combination of dialectics and gradualism. In Marx's view social development was a dialectical process: the transition from one stage to another took place through a revolutionary transformation, which was preceded by increasing deterioration of society and intensifying class struggles. Underlying this discontinuous development was the more gradual development of the forces of production (technology and organization of labour).

Marx was influenced by the countercurrent of Romanticism, which was opposed to the idea of progress. This influence was evident in his notion of "alienation," which meant that in the course of social development people had increasingly lost control over the social forces that they had produced by their own activities. Romantic counterprogressivism was, however, much stronger in the work of other social theorists of the century, such as Ferdinand Tönnies, a German sociologist. He distinguished between the community (*Gemeinschaft*), in which people were bound together by common traditions and ties of affection and solidarity, and the society (*Gesellschaft*), in which social relations had become contractual, rational, and nonemotional.

Gemeinschaft and *Gesellschaft*

Durkheim and Max Weber, sociologists who began their careers at the end of the 19th century, showed ambivalence toward the ideas of progress. Durkheim regarded the increasing division of labour as a basic process, which was at the roots of modern individualism, but could also lead to "anomie," or lack of moral norms. Weber rejected evolutionism by arguing that the development of Western society was quite different from that of other civilizations and therefore historically unique. It was characterized, according to Weber, by a peculiar type of rationalization, which had brought modern capitalism, modern science, and rational law, but also, on the negative side, a "disenchantment of the world" and increasing bureaucratization.

The work of Durkheim, Weber, and other social theorists around the turn of the century marked a transition from evolutionism toward more static theories. Evolutionary theories were criticized on empirical grounds—they could be refuted by a growing mass of research findings—and because of their determinism and Western-centred optimism. Theories of cyclical change that denied long-term progress gained popularity in the first half of this century; these included the theory of the Italian economist and sociologist Vilfredo Pareto on the "circulation of elites" and those of Oswald Spengler and Arnold Toynbee on the life cycle of civilizations. Although the interest in long-term social change never disappeared, it faded to the background, especially when, from the 1920s until the 1950s, functionalism, emphasizing an interdependent social system, became the dominant paradigm both in anthropology and in sociology. "Social evolution" was substituted for the more general and neutral concept of "social change."

Decline of evolutionism

From the 1950s and increasingly through the 1960s and 1970s there was a revival of interest in long-term social change. Neo-evolutionist theories were proclaimed by several anthropologists, including Ralph Linton, Leslie A. White, Julian H. Steward, Marshall D. Sahlins, and Elman Rogers Service. These authors hold to the idea of social evolution as a long-term development of mankind, which is patterned and cumulative in some respects. Neo-evolutionism differs from 19th-century evolutionism in that it does not assume that all societies go through the same stages of development; much attention is paid to variations between societies as well as to relations of influence among them. The latter concept has come to be known by the term acculturation. Moreover, social evolution is not regarded as predetermined or inevitable but is rather conceived in terms of probabilities. Finally, evolutionary development is not equated with progress.

The revival of interest in long-term social change has been partly induced by the problems of the so-called underdeveloped countries. In order to explain the gaps between rich and poor countries, Western sociologists and economists in the 1950s and 1960s elaborated modernization theories. These theories implied a covertly Western-centred evolutionism insofar as they assumed that poor countries had stagnated on a relatively low level of development and could and should develop, or modernize, in the direction of a Western-type society. Modernization theories have been criticized for their lack of attention to international power relations, in which the richer countries dominate the poorer ones. These relations have been brought into the centre of attention by more recent theories of international dependency or, in Immanuel Wallerstein's terms, the "world capitalist system."

Modernization theories

Since about 1965 there has been some convergence between sociology and anthropology on the one hand and history on the other. Historians have become interested in theories of long-term social change, while many sociologists and anthropologists increasingly turn toward history for the empirical testing and refinements of their theoretical viewpoints.

Patterns of social change. The common assumption of theories of social change, old and new, is that the course of such change is not arbitrary but, to a certain degree, regular or patterned. The three traditional ideas of social change—those of decline, cyclical change, and progress—have influenced modern theories. However, insofar as these theories are nonnormative, or scientifically determined, they do not distinguish explicitly between decline and progress. Such values cannot be derived from empirical observations alone but depend on normative evaluations, or value judgments. In nonnormative terms, then, two basic patterns of social change emerge: the cyclical and the one-directional. Often the time span of the change determines which pattern is observed.

Cyclical change. A regular alternation of stages characterizes cyclical change. Much of ordinary social life is organized in cyclical changes: those of the day, the week, and the year. These short-term cyclical changes may be regarded as conditions necessary to structural stability. Other changes that have a more or less cyclical pattern are less regular. For example, business cycles, recurrent phenomena of capitalist and industrial societies, are patterned to some extent yet hard to predict in concrete cases. A well-known theory of the business cycle is that of the Soviet economist Nikolay D. Kondratyev, who tried to show the recurrence of long waves of economic boom and recession on an international scale. He charted the waves from the end of the 18th century, with each complete wave comprising a period of about 50 years. Subsequent research has shown, however, that the patterns in different countries have been far from identical.

Kondratyev's waves

Long-term cyclical changes are rendered by theories on the birth, growth, flourishing, decline, and death of civilizations. Toynbee conceived world history in this way in the first volumes of *A Study of History* (1934–61), as did Spengler in his *Untergang des Abendlandes* (1918–22; *Decline of the West*). These theories have been criticized for their conception of civilizations as natural entities with sharp boundaries because this tends to neglect the interrelations between civilizations.

One-directional change. A continuation in terms of more or less characterizes one-directional change. Such change is usually cumulative; it implies growth or increase, such as that of population density, the size of organizations, or the level of production. However, the direction of the change may also be one of decrease, or a combination of growth and decrease. An example of this last process is what the American cultural anthropologist Clifford Geertz has called "involution," found in some agrarian societies: population growth coupled with decreasing per capita wealth. Or the change may be a shift from one to the other pole of a continuum—from religious to scientific ways of thinking, for example. Such a change may be defined as either growth (of scientific knowledge) or decline (of religion).

The concept of involution

The simplest type of one-directional change is linear: the extent of social change is constant over time. Another type of regular social change is exponential growth, in which the growth percentage is constant over time and the change accelerates correspondingly. Population growth and production growth often approximate this pattern during some periods of time.

A pattern of long-term growth may also conform to a three-stage S-curve: at the beginning of the period under consideration the change is almost imperceptibly slow, then accelerates, then slackens, until it approaches a supposed upper limit. The model of the demographic transition in industrializing countries exhibits this pattern. In the first stage, premodern or preindustrial, both the birthrate and the mortality rate are high, and, consequently, the population grows very slowly; then mortality decreases, and the population grows much faster; in the third stage both the birthrate and the mortality rate have become low, and the population growth approaches zero. The same model has been suggested, more hypothetically, for the rate of technological and scientific change.

Combined patterns of change. Cyclical and one-directional changes may be combined in one way or another. Very often short-term changes are cyclical while long-term development is in one direction. Figures on the production rates of industrializing countries conform, more or less, to this pattern, short-term business cycles occurring within long-term economic growth.

All of these pattern models cannot be applied simply and easily to social reality. They are at best approximations of parts of social reality. Comparing the model with the reality is not always possible because of a lack of reliable data. Moreover, and more importantly, many social processes do not lend themselves to precise quantitative measurement. Processes like bureaucratization or secularization, for example, can be defined as changes in a certain direction, but it is hard to measure the extent to which the change in a given period has taken place. It is doubtful that models like those of linear or exponential growth can be used in such cases.

Determination of long-term change

It remains to be seen whether or not long-term social change in a certain direction may be ascertained. Many investigations have sought answers to this question for Western society since the Middle Ages. The transformation of medieval society into the Western nations of the 20th century may be conceived in terms of several interconnected, long-term, one-directional changes; some of the more important of these include commercialization, increasing division of labour, growth of production, formation of national states, bureaucratization, growth of technology and science, secularization, urbanization, spread of literacy, increasing social and geographical mobility, and growth of organizations. Many of these changes have also occurred in non-Western societies. Most changes have not originated in the West, but some important changes did originate there—particularly such complex transformations as the rise of capitalism and the Industrial Revolution. These subsequently had a strong impact on non-Western societies. Groups of people outside western Europe have been incorporated in a global division of labour, in which the Western nation-states dominated both politically and economically.

General trends of social evolution

The extent to which these changes are part of a global, long-term social development is the central question of social evolution, which is conceived as a very long-term one-directional change for mankind as a whole. Although knowledge concerning this question is far from complete, some very broad and general trends may be hypothesized on firm ground. First, technological innovations and growing empirical knowledge led to an increasing control of natural forces for the satisfaction of human needs. Parts of this development were the use and control of fire, the cultivation of plants and the domestication of animals (dating from about 8000 BC), the use of metals, and the process of industrialization. This technological development, combined with long-term capital accumulation, led to rising production levels and, therefore, made possible population growth and increasing population density. Energy production and consumption grew, if not per capita then at least per square mile.

Interconnected with technological development and growth of production were the process of division of labour and social differentiation. On the one hand, it was only by division of labour and corresponding specialization of knowledge and abilities that the technical control of natural forces could increase beyond certain limits. On the other hand, the growth of production as a result of technological innovations contributed to further social differentiation; more people, in other words, could specialize in activities that were not immediately necessary for survival. Growing size and density of populations and social differentiation led to increased interdependence between growing numbers of people over longer distances. In hunting and gathering societies people were strongly interdependent within their small bands, depending on very little outside their groups. In modern times, most of the world's people are enmeshed in worldwide networks of interdependence.

These processes were not inevitable in the sense that they corresponded to any "law" of social change. They had the tendency, however, to spread whenever they occurred. For example, once the set of transformations known as the agrarian revolution had taken place anywhere in the world, their extension over the rest of the world was predictable. Societies that adopted these innovations grew in size and became more powerful. As a consequence, other societies had only three options: to be conquered and incorporated by a more powerful agrarian society; to adopt the innovations; or to be driven away to marginal places of the globe. Something similar might be said of the Industrial Revolution and other power-enhancing innovations, such as bureaucratization and the introduction of more destructive weapons. This last example illustrates that these processes should not be equated with progress in general.

Causal connections between social processes

Explanations of social change. One way of explaining social change is to show causal connections between two or more processes. This may take the form of a kind of determinism or reductionism, which explains all social change by reducing it to one supposed autonomous and all-determining causal process. A more cautious assumption is that one process has relative causal priority, without implying that this process is completely autonomous and all-determining. Following are some of the processes conceived as having caused social change.

Natural environment. Changes in the natural environment may vary from climatic ones to those caused by the spread of diseases. For example, both the worsening of climatic conditions and the epidemics of the Black Death have been submitted as factors that explain the crisis of feudalism in 14th-century Europe. Changes in the natural environment may be either independent of human social activities or the result of these activities. Deforestation and erosion, air pollution, and the exhaustion of natural resources belong to the last category, and they in turn may have far-reaching social consequences.

Demographic processes. Population growth and increasing population density represent, in particular, demographic forms of social change. Population growth may lead to geographical expansion of a society, military conflicts, and the intermingling of cultures. Increasing population density may also stimulate technological innovations, which may increase division of labour and social differentiation, commercialization, and urbanization. This has been observed as affecting western Europe from the 11th to the 13th century and England in the 18th century, where population growth was a factor in the Industrial Revolution. On the other hand, population growth may contribute to economic stagnation and increasing poverty, as may be witnessed in several Third World countries today.

Technological innovations. According to several theories of social evolution, technological innovations are regarded as the most important determinants of societal change. The social significance of such technological breakthroughs as the invention of the smelting of iron, the introduction of the plow in agriculture, the invention of the steam engine, and the development of the computer is indeed evident. Of course, it is possible to dispute the relative importance of such innovations when compared to other determinants of social change.

Economic processes. Technological changes are often considered in conjunction with economic processes, including the formation and extension of markets, modifications of property relations (such as the change from feudal lord–peasant relations to contractual proprietor–tenant relations), and changes in the organization of labour (such as the change from independent craftsmen to factories). Historical materialism, as developed by Marx and Engels, is the most influential theory that gives priority to economic processes, but it is not the only one. Materialist theories have been developed even in opposition to Marxism, one being the "logic of industrialization" thesis by the U.S. scholar Clark Kerr, which states that industrialization everywhere has similar consequences, whether the property relations are called capitalist or communist.

Ideas. Other theories have stressed the significance of ideas in the causation of social change. Comte's law of three stages is such a theory. Weber regarded religious ideas as important in contributing to economic development or stagnation; according to his controversial thesis, the individualistic ethic of Christianity, and in particular

Protestantism, partially explains the rise of the capitalist spirit, which brought economic dynamism in the West.

Social movement. A change of collective ideas is not merely an intellectual process; it is often connected to the formation of a new social movement. This in itself might be regarded as a potential cause of social change. Weber called attention to this factor in conjunction with his concept of "charismatic leadership." The charismatic leader, by virtue of the extraordinary personal qualities attributed to him, is able to create a group of followers who are willing to break established rules.

Political processes. Changes in the regulation of violence, in the nature of the state organization, and in international relations may also determine social change. For example, the German sociologist Norbert Elias has analyzed the formation of states in western Europe as a relatively autonomous process that led to an increasing control of violence and, consequently, to rising standards of self-control. According to recent theories of political revolution, the functioning of the state apparatus itself and the nature of interstate relations are of decisive importance in the outbreak of a revolution: it is only when the state is not able to fulfill its basic functions of maintaining law and order and territorial integrity that revolutionary groups have any chance of success.

Criticism of determinism

Each of these processes is a possible determinant of other ones; none of them is the only determinant. One reason why deterministic, or reductionist, theories run into difficulties is that the process they use to explain the process as a whole is actually not autonomous but has to be itself explained. Moreover, social processes are often intertwined to such a degree that it would be misleading to consider them separately. For example, there are no sharp and fixed borderlines between economic and political processes, nor between economic and technological processes. Technological change may in itself be regarded as a specific type of cultural or conceptual change. The causal connections between distinguishable social processes are a matter of degree and vary over time.

Mechanisms of social change. The scope of any causal explanation of social change in which initial conditions or basic processes are specified is limited. A more general and theoretical way of explaining is to construct a model of recurring mechanisms of social change. Such mechanisms, incorporated in different theoretical models, include the following.

Mechanisms of one-directional change: accumulation, selection, and differentiation. Some evolutionary theories stress the essentially cumulative nature of human knowledge. Because human beings are innovative, they add to existing knowledge, replacing less adequate ideas and practices with more adequate ones. As they learn from mistakes, they select new ideas and practices in a trial-and-error process (sometimes compared to the process of natural selection). The expansion of collective knowledge and capabilities beyond a certain limit is only possible by specialization and differentiation. Growth of technical knowledge stimulates capital accumulation, which leads to rising production levels. Population growth may also be incorporated in this model of cumulative evolution: it is by the accumulation of collective technical knowledge and means of production that human beings can multiply their numbers; this growth then leads to new problems that stimulate further innovations.

Mechanisms of curvilinear and cyclical change: saturation and exhaustion. Models of one-directional change assume that change in a certain direction induces further change in the same direction; models of curvilinear or cyclical change, on the other hand, assume that change in a certain direction creates the conditions for change in another (perhaps even the opposite) direction. More specifically, it is often assumed that growth has its limits and that in approaching these limits the change curve will inevitably be bent. Ecological conditions like the availability of natural resources, in particular, set limits to population growth and economic growth.

Shorter term cyclical changes are explained by comparable mechanisms. Some theories of the business cycle, for example, assume that the economy is saturated periodically with capital goods: investments become less necessary and less profitable, the rate of investments diminishes, and a negative spiral resulting in a recession sets in. After a period of time, however, essential capital goods will have to be replaced: investments are pushed up again and a phase of economic expansion begins.

Group conflict as a cause of change

Conflict, competition, and cooperation. Group conflict has often been viewed as a basic mechanism of social change, especially of those radical and sudden social transformations identified as revolutions. Marxists in particular tend to depict social life in capitalist society as a struggle between a ruling class, which wishes to maintain the system, and a dominated class striving for radical change; social change then is the result of that struggle. These ideas are basic to what Dahrendorf has called a conflict model of society.

The notion of conflict becomes more relevant for the explanation of social change if it is broadened to include competition between rival groups as well. Nations, firms, universities, sports associations, and artistic schools are groups between which such rivalry occurs. Competition stimulates the introduction and diffusion of innovations, especially when they are potentially power-enhancing. Thus, the leaders of non-Western states feel the necessity of adopting Western science and technology, even though their ideology may be anti-Western, because it is only by these means that they can maintain or enhance national autonomy and power.

Additionally, competition may lead to the growing size and complexity of the entities involved. The classic example of this process, analyzed by Marx, is the tendency in capitalism for monopolies to form as small firms are driven out of competition by larger ones. Marx's analysis has been applied to another area by Norbert Elias, who explained the formation of national states in western Europe as the result of competitive struggles between feudal lords.

Competition is also put forward in individualistic theories, which conceive social change as the result of the actions of individuals pursuing their self-interest. With the help of game theory and other mathematical devices it has been shown that individuals acting on the basis of self-interest will cooperate, given certain conditions, in widening social networks.

Cultural lag

Tension and adaptation. In structural functionalism, social change is regarded as the adaptive response to some tension within the social system. When some part of an integrated social system changes, a tension between this and other parts of the system is created, which will be resolved by the adaptive change of the other parts. An example is what the U.S. sociologist William Fielding Ogburn has called cultural lag, which refers in particular to a gap that develops between fast-changing technology and other slower paced sociocultural traits.

Young urban professionals

Diffusion of innovations. Some social changes are to be regarded as the result of the diffusion of innovations, such as technological inventions, new scientific knowledge, new beliefs, or a new fashion in the sphere of leisure. Diffusion is not automatic but selective; an innovation is only adopted by people if they are motivated to do so and if it is compatible with important aspects of their culture. One reason for the adoption of innovations by larger groups is the example of higher status groups, which are reference groups for other people. Successful innovations, which affect the majority of the people of a society, tend to follow a pattern of diffusion from higher to lower status groups. More specifically, most early adopters of innovations in modern Western societies, according to several studies, are young, urban, and highly educated, with a high occupational status. Often they are motivated by the wish to distinguish themselves from the mass of the population. After diffusion has taken place, however, the innovation is no longer a symbol of distinction, which motivates the same group to look for something new again. This mechanism may explain the succession of trends in several fields.

Planning and institutionalization of change. Social change may be, to a certain extent, the result of goal-directed, large-scale social planning. The possibilities of planning by government bureaucracies and other large organizations have increased in modern societies. Most social planning is short-term, however; the goals of planning are often not attained, and, even if the planning is successful in terms of the stated goals, it often has unforeseen consequences. The wider the scope and the longer the time span of planning, the more difficult it is to attain the goals and to avoid unforeseen and undesired consequences. This has become especially clear in Communist societies, where the most serious efforts have been taken to put the ideal of integral and long-term planning into practice. Large-scale and long-term social developments in any society are still largely unplanned.

Planning implies institutionalization of change, but institutionalization does not imply planning. Many unplanned social changes in modern societies are institutionalized; they originate in organizations permanently oriented to innovation, such as universities and the research departments of governments and private firms, but their social repercussions are not controlled. It is in the fields of science and technology especially that change is institutionalized, producing social change that is partly intended and partly unintended.

These mechanisms of social change are not mutually exclusive. On the contrary, some of them are clearly interconnected. For example, innovation by specialized organizations is stimulated by competition. Several mechanisms may be combined in one explanatory model of social change.

CONCLUSION

The interconnection of structure and change

Social structure and social change are central theoretical concepts of the social sciences that refer to basic and complementary characteristics of social life in general—permanence, continuity, and repetitiveness on the one hand, dynamics and changeability on the other. Both concepts are interconnected: the social structure cannot be conceptualized adequately without some notion of actual or potential change, and social change as a more or less regular process is inconceivable without the notion of continuity. To the degree that change processes are regular and interconnected, social change itself is structured. Any separation of the two concepts, as though they refer to divergent fields, is therefore misleading. This is not to deny that the relative stress on either structural continuity or dynamic change varies in social scientific theories and empirical studies. Since about 1965 there has been a shift from "structure" to "change" in social theory. Change on different levels—social dynamics in everyday life, short-term transformations and long-term developments in society at large—has become the focus of attention.

BIBLIOGRAPHY. A general reader on social structure is PETER M. BLAU (ed.), *Approaches to the Study of Social Structure* (1975). The most important theoretical works in structural functionalism are A.R. RADCLIFFE-BROWN, *Structure and Function in Primitive Society* (1952, reissued 1965, reprinted 1968); and TALCOTT PARSONS, *The Social System* (1951, reprinted 1964). For coverage of the debate on structural functionalism, see N.J. DEMERATH and RICHARD A. PETERSON (eds.), *System, Change, and Conflict* (1967, reprinted 1968). A more empirical type of functionalism is represented by ROBERT K. MERTON, *Social Theory and Social Structure: Toward the Codification of Theory and Research,* new ed. (1968), in which due consideration is given the distributive aspects of the social structure. These are stressed even more by PETER M. BLAU, *Inequality and Heterogeneity: A Primitive Theory of Social Structure* (1977). RALF DAHRENDORF, *Class and Class Conflict in Industrial Society* (1959; originally published in German, 1957), advances a power-and-conflict model of society. Other, more sophisticated power models are contained in PETER M. BLAU, *Exchange and Power in Social Life* (1964); STEVEN LUKES, *Power: A Radical View* (1974); and NORBERT ELIAS, *What Is Sociology?* (1978; originally published in German, 3rd ed., 1978). An introduction to structuralism is DAVID ROBEY (ed.), *Structuralism* (1973). CLAUDE LÉVI-STRAUSS, *Structural Anthropology,* 2 vol. (1963–76; originally published in French, 1958–73), contains several articles on the structural method and its applications. Examples of different empirical applications of the concept of social structure are GEORGE PETER MURDOCK, *Social Structure* (1949, reissued 1965); PETER M. BLAU and OTIS DUDLEY DUNCAN, *The American Occupational Structure* (1967, reprinted 1978); and PETER V. MARSDEN and NAN LIN (eds.), *Social Structure and Network Analysis* (1982). A synthesis of different views is offered by ANTHONY GIDDENS, *Central Problems in Social Theory: Action, Structure and Contradiction in Social Analysis* (1979, reprinted 1983).

On the history of ideas concerning social change, see ROBERT A. NISBET, *Social Change and History: Aspects of the Western Theory of Development* (1969). An introduction to 18th- and 19th-century evolutionism is LOUIS SCHNEIDER, *Classical Theories of Social Change* (1967). Original texts in social evolutionism are HERBERT SPENCER, *The Principles of Sociology,* 3 vol. in 4 (1876–96, reprinted in 3 vol., 1975), and *Herbert Spencer: Structure, Function, and Evolution,* ed. by STANISLAV ANDRESKI (1971, reissued 1972); LEWIS HENRY MORGAN, *Ancient Society* (1877, reissued 1985); and EDWARD B. TYLOR, *Primitive Culture,* 2 vol. (1871, reissued 1970). Good selections of Marxian texts are KARL MARX, *Selected Writings in Sociology and Social Philosophy,* ed. by T.B. BOTTOMORE and MAXIMILIAN RUBEL (1956, reprinted 1964); and KARL MARX and FREDERICK ENGELS, *Selected Works,* 2 vol. (1935, reissued in 1 vol., 1968). The most influential study in Marxist evolutionism is FREDERICK ENGELS, *The Origin of the Family, Private Property and the State* (1902, reissued 1978; originally published in German, 1884). A criticism of Spencer's evolutionism is contained in ÉMILE DURKHEIM, *Émile Durkheim on the Division of Labor in Society* (1933, reissued 1984 as *The Division of Labor in Society;* originally published in French, 1893); while MAX WEBER, *The Protestant Ethic and the Spirit of Capitalism* (1930, reprinted 1985; originally published in German, 1920, in vol. 1 of his *Gesammelte Aufsätze*), contains a criticism of historical materialism.

Anthropological neo-evolutionism is represented by: LESLIE A. WHITE, *The Evolution of Culture: The Development of Civilization to the Fall of Rome* (1959); JULIAN H. STEWARD, *Theory of Culture Change: The Methodology of Multilinear Evolution* (1955, reprinted 1973); MARSHALL D. SAHLINS and ELMAN R. SERVICE (eds.), *Evolution and Culture* (1960, reprinted 1982); W.F. WERTHEIM, *Evolutie en revolutie: De golfslag der emancipatie* (1971), from which an abridged English trans., *Evolution and Revolution: The Rising Waves of Emancipation* (1974), was made; and ELMAN R. SERVICE, *Cultural Evolutionism: Theory in Practice* (1971). A sociological textbook with an evolutionary approach is GERHARD LENSKI and JEAN LENSKI, *Human Societies: An Introduction to Macrosociology,* 4th ed. (1982). S.N. EISENSTADT, *Tradition, Change, and Modernity* (1973, reprinted 1983), represents a sophisticated version of the modernization theory. Good examples of historical sociology are NORBERT ELIAS, *The Civilizing Process,* 2 vol. (1978; originally published in German, 1939); BARRINGTON MOORE, JR., *Social Origins of Dictatorship and Democracy: Lord and Peasant in the Making of the Modern World* (1966, reissued 1984); and IMMANUEL WALLERSTEIN, *The Modern World-System,* vol. 1, *Capitalist Agriculture and the Origins of the European World-Economy in the Sixteenth Century* (1974). Akin to these books are comprehensive historical studies on long-term developments, such as WILLIAM H. MCNEILL, *The Rise of the West: A History of the Human Community* (1963, reissued 1965); and FERNAND BRAUDEL, *Capitalism and Material Life, 1400–1800* (1974; originally published in French, 1967). An overview of this field is given by THEDA SKOCPOL (ed.), *Vision and Method in Historical Sociology* (1984).

General theoretical books on social change are: WILBERT E. MOORE, *Social Change,* 2nd ed. (1974), and *Order and Change: Essays in Comparative Sociology* (1967), two treatises in the functionalist tradition; EVA ETZIONI-HALEVY and AMITAI ETZIONI (eds.), *Social Change: Sources, Patterns, and Consequences,* 2nd ed. (1973), a reader representing various approaches; WILLIAM FIELDING OGBURN, *Social Change with Respect to Culture and Original Nature,* new ed. (1950, reprinted 1965); AMITAI ETZIONI, *The Active Society: A Theory of Societal and Political Processes* (1968, reprinted 1971), which explores the possibilities of planned change; ROBERT L. HAMBLIN, R. BROOKE JACOBSEN, and JERRY L.L. MILLER, *A Mathematical Theory of Social Change* (1973); HENRY TEUNE and ZDRAVKO MLINAR, *The Developmental Logic of Social Systems* (1978); and KENNETH E. BOULDING, *Ecodynamics: A New Theory of Societal Evolution* (1978, reprinted 1981).

Controversial theories on the cyclical development of civilizations have been advanced by OSWALD SPENGLER, *The Decline of the West,* 2 vol. (1922, reissued 1981–83; originally published in German, 1918–22); and ARNOLD J. TOYNBEE, *A Study of History,* 12 vol. (1934–61, reprinted 1948–61). A theory of the circulation of elites can be found in VILFREDO PARETO, *The Mind and Society: Treatise on General Sociology,* 4 vol. (1935, reprinted 1983; originally published in Italian, 2nd ed., 3 vol., 1923), and *Sociological Writings,* ed. by S.E. FINER (1966, reprinted 1976). An empirical test of theories of economic growth and the business cycle is given by ANGUS MADDISON, *Phases of Capitalist Development* (1982). The concept of involution is explained by CLIFFORD GEERTZ, *Agricultural Involution: The Process of Ecological Change in Indonesia* (1963, reprinted 1968).

The significance of demographic processes is analyzed by CARLO M. CIPOLLA, *The Economic History of World Population,* 7th ed. (1978), and an analysis of one stage of development is presented in MARK NATHAN COHEN, *The Food Crisis in Prehistory: Overpopulation and the Origins of Agriculture* (1977, reprinted 1979). A classic account of the influence of technological change is V. GORDON CHILDE, *Man Makes Himself,* rev. ed. (1951, reissued 1983). CLARK KERR *et al., Industrialism and Industrial Man: The Problems of Labor and Management in Economic Growth,* 2nd ed. (1964, reissued 1973), represents a non-Marxist materialist view. Theories of political revolution are developed by CRANE BRINTON, *The Anatomy of Revolution,* rev. and expanded ed. (1965); and THEDA SKOCPOL, *States and Social Revolutions: A Comparative Analysis of France, Russia, and China* (1979). EVERETT M. ROGERS, *Diffusion of Innovations,* 3rd ed. (1983), deals with the social aspects of technological innovations. The individualistic approach to social change processes is exemplified by DOUGLASS C. NORTH, *Structure and Change in Economic History* (1981); MANCUR OLSON, *The Rise and Decline of Nations: Economic Growth, Stagflation, and Social Rigidities* (1982, reprinted 1984); and ROBERT AXELROD, *The Evolution of Cooperation* (1984).

An influential criticism of deterministic theories of social development is KARL R. POPPER, *The Poverty of Historicism,* 2nd ed. (1960). Examples of pessimistic social forecasting with much attention to ecological conditions are DONELLA H. MEADOWS, *et al., The Limits to Growth,* 2nd ed. (1974, reprinted 1982); and MIHAJLO MESAROVIC and EDUARD PESTEL, *Mankind at the Turning Point* (1974, reissued 1976). Much more optimistic examples are HERMAN KAHN and ANTHONY J. WIENER, *The Year 2000* (1967); DANIEL BELL, *The Coming of Post-Industrial Society: A Venture in Social Forecasting* (1973, reprinted 1976); and CLARK KERR, *The Future of Industrial Societies: Convergence or Continuing Diversity?* (1983). For a history of ideas about future social developments, see KRISHAN KUMAR, *Prophecy and Progress: The Sociology of Industrial and Post-Industrial Society* (1978).

(N.Wi.)

Social Welfare

The basic concerns of social welfare—poverty, disability and disease, the dependent young and elderly—are as old as society itself. The laws of survival once severely limited the means by which these concerns could be addressed; to share another's burden meant to weaken one's own standing in the fierce struggle of daily existence. As societies developed, however, with their patterns of dependence between members, there arose more systematic responses to the factors that rendered individuals, and thus society at large, vulnerable.

Religion and philosophy have tended to provide frameworks for the conduct of social welfare. The edicts of the Buddhist emperor Aśoka in India, the sociopolitical doctrines of ancient Greece and Rome, and the simple rules of the early Christian communities are a few examples of systems that addressed social needs. The Elizabethan Poor Laws in England, which sought relief of paupers through care services and workhouses administered at the parish level, provided precedents for many modern legislative responses to poverty. In Victorian times a more stringent legal view of poverty as a moral failing was met with the rise of humanitarianism and a proliferation of social reformers. The social charities and philanthropic societies founded by these pioneers formed the basis for many of today's welfare services.

Because perceived needs and the ability to address them determine each society's range of welfare services, there exists no universal vocabulary of social welfare. In some countries a distinction is drawn between "social services," denoting programs, such as health care and education, that serve the general population, and "welfare services," denoting aid directed to vulnerable groups, such as the poor, the disabled, or the delinquent. According to another classification, remedial services address the basic needs of individuals in acute or chronic distress; preventive services seek to reduce the pressures and obstacles that cause such distress; and supportive services attempt, through educational, health, employment, and other programs, to maintain and improve the functioning of individuals in society. Social welfare services originated as emergency measures to be applied when all else failed; they are now generally regarded as a necessary function in any society, a means not only of rescuing the endangered but also of fostering a society's ongoing, corporate well-being.

This article treats first the personal social services—those provided on an individual basis to persons in need; collectively these services constitute the professional field of social work. A discussion of government-sponsored social security programs, such as social assistance and social insurance, follows.

The article is divided into the following sections:

Social work: the personal social services

The majority of personal social services are rendered on an individual basis to people who are unable, whether temporarily or permanently, to cope with the problems of everyday living. Recipients include families faced with loss of income, desertion, or illness; children and youths whose physical or moral welfare is at risk; the sick; the disabled; the frail elderly; and the unemployed. When possible, services are also directed toward preventing threats to personal or family independence.

The social services generally place a high value on keeping families together in their local communities, organizing informal support from friends or neighbours when kinship ties are weak. Where necessary, the services provide substitute forms of home life or residential care, and they play a key role in the care and control of juvenile delinquents and other socially deviant groups, such as drug and alcohol abusers.

MODERN EVOLUTION

In the advanced industrial societies the personal social services have always constituted a "mixed economy of welfare," involving the statutory, voluntary, and private sectors of welfare provision. Although their role is crucial, they account for only a small proportion of total welfare expenditures. The most substantial increases in expenditures have occurred in social security systems, which provide assistance to specific categories of claimants on the basis of both universal and selective criteria. The development of modern social security systems from the 1880s reflects not only a gradual but fundamental change in the aims and scope of social policy but also a dramatic shift in expert and popular opinion with regard to the relative significance of the social and personal causes of need (see below *Social security: government welfare programs*).

Poor relief

In the belief that personal shortcomings were the chief cause of poverty and of people's inability to cope with it, the major 19th-century systems of poor relief in western Europe and North America tended to withhold relief from all but the truly destitute, to whom it was given as a last resort. This policy was intended as a general deterrent to idleness. The poor-law relieving officer was the precursor of both the public assistance officials and the social workers of today in his command of statutory financial aid. The voluntary charitable agencies of the time differed on the relative merits of deterrent poor-law services on the one hand, implying resistance to the growth of statutory welfare, and on the provision of alternative assistance to the needy, coupled with the extension of statutory services, on the other hand. From the 1870s the Charity Organization Society and similar bodies in the United States, Britain, and elsewhere held strongly to the former option, and their influence was widespread until the outbreak of World War II.

Settlement movement

The settlement movement in Britain and the United States drew voluntary workers into direct contact with the serious material disadvantages suffered by the poor. The pioneer of this movement was the vicar Samuel A. Barnett, who in 1884 with his wife and a number of university students "settled" in a deprived area of London, calling their neighbourhood house Toynbee Hall. Two visitors to this settlement soon introduced the movement into the United States—Stanton Coit, who founded Neighborhood Guild (later University Settlement) on the Lower East Side of New York City in 1886, and Jane Addams, who with Ellen Gates Starr founded Hull House on the Near West Side of Chicago in 1889. From these prototypes the movement spread to other U.S. cities and abroad through Europe and Asia.

The origins of modern social casework can be traced to the appointment of the first medical almoners in Britain in the 1880s, a practice quickly adopted in North American and most western European countries. The almoners originally performed three main functions: ascertaining the financial eligibility and resources of patients faced with the rising costs of medical care, providing counseling services to support patients and their families during periods of ill health and bereavement, and procuring adequate practical aids and other forms of home care for discharged patients. Elsewhere secular and religious charitable associations providing financial help, educational welfare, and housing for the poor began to employ social workers.

By the turn of the century there were various schemes for organizing charitable work on "scientific" principles according to nationally agreed standards of procedure and services. In Britain, the United States, Germany, and, later, Japan, leading charities worked in conjunction with poor-law and public assistance authorities, an approach endorsed in 1909 in the majority report of the British Royal Commission on the Poor Law. The first schools of social work, usually run by the voluntary charitable agencies, appeared in the 1890s and early 1900s in London, New York City, and Amsterdam, and by the 1920s there were similar ventures in other parts of western Europe and North America and in South America. The training programs combined casework methods and other practical forms of intervention and support, with particular emphasis on working in cooperation with individuals and families to restore a level of independence.

Early social work training

From the 1900s onward the surveys conducted by Charles Booth in London and Seebohm Rowntree in York and by other researchers began to transform conventional views of the role of the state in social welfare and the relief of poverty, and the social causes of poverty came under scrutiny. At the same time, the scope of social work was growing, with the spread of settlement houses, to include group work and community action.

In most countries social welfare services, or personal social services, rather than being separately organized and administered, are often attached to other major social services, such as social security, health care, education, and housing. This is explained by the course of their historical development. The means open to policy-making and administration in the personal social services are often incompatible. For example, the demands of the general integration and coordination of care programs can conflict with the provision of services that take due account of the needs of specific client groups. Also to be reconciled are the provision of individual services and the provision for family and neighbourhood needs.

Statutory and voluntary social services have evolved in response to needs that could not be fully met by individuals either alone or in association with others. Among the factors determining the present nature of such services are, first, that the growth in the scale and complexity of industrial societies has added to the obligations of central and local governments. Second, the increasing wealth and productivity of industrial societies has heightened public expectations regarding standards of living and standards of justice, at the same time augmenting the material capacity to meet those expectations. Third, the processes of social and economic change have grown to such proportions that individuals are increasingly ill-equipped to anticipate and cope with the adverse effects of such change. Fourth, it is difficult and sometimes impossible to recognize and provide for the idiosyncratic needs arising from the interaction of social and personal life.

Any family can experience crises that it is powerless to control. The hardships of ill health and unemployment can be compounded by loss of income; divorce and separation can impede the welfare and development of young children; and long-term responsibility for dependent relatives can impair the physical and emotional well-being of those who provide the care.

A very small number of families experience such intractable problems that they require almost continuous help from personal social services. Some of these families present problems of deviant behaviour, including family violence and child abuse, irregular attendance or nonenrollment in school, alcohol and drug abuse, and crime and delinquency. Not all poor families, however, make heavy demands on social welfare services; indeed considerable hardship could be alleviated through more efficient use of existing services.

Over time, social workers have acquired a special responsibility for people whose particular needs fall outside the aegis of other professions and agencies. Apart from the requirements of individuals and families with serious long-term social and emotional problems, personal social services meet a wide spectrum of needs arising from the more routine contingencies of living. Inevitably personal social services are primarily concerned with reacting to a crisis as it occurs, but today much effort is being invested in preventive work and in the enhancement of welfare in the wider community. In this respect comparison can be made with the traditional aim of social security—the reduction of poverty—and the more ambitious objective of income maintenance (see below *Social security: government welfare programs*).

The organization of personal social services in different societies is extremely variable. Ethnicity and urban deprivation have added new dimensions to need that cut across the traditional client categories of families, children, youth, the sick and handicapped, the unemployed, the aged, and the delinquent. Nevertheless, there are continuities and consistencies in the pattern of needs that characterize these major client groups.

MAJOR AREAS OF CONCERN

Family welfare. Social philosophers and caseworkers generally regard family life as the ideal context for the promotion of social welfare. Family welfare programs seek to preserve and strengthen the family unit through both economic assistance, where available, and personal assistance with a variety of services. Personal assistance services include marriage counseling in most developed countries and in urban centres of developing countries; maternal, prenatal, and infant care programs; family planning services; family-life education, which promotes both the enrichment of family relationships and the improvement of home economics; "home-help" or "homemaker" services providing household assistance to families burdened with chronic illness, handicaps, or other dependencies; and care of the aged through such programs as in-home meal services, transportation, regular visitation, and reduced-cost medicines.

Family services

Child welfare. A paramount concern in all family welfare programs is the welfare of children. Whenever possible, children's services are rendered within the setting of home life. Income assistance to parents may help ensure the basic security of the family structure. Maternal, prenatal, and child health-care programs are important in all societies but especially so in those affected by widespread disease and malnutrition; infant and maternal mortality rates are in fact the most basic indexes of child welfare. The increasing number of working mothers worldwide has given rise to day-care services ranging from simple custodial supervision to educational and health-care programs. In some countries, industries are required to provide such facilities for their employees, in recognition of the changing economic pressures on family life.

Day-care services

While the family unit is imbued with great value by most child-welfare programs, these programs must also address the special needs of unwed mothers and their children, broken families, and children whose families, although intact, are sources of abuse and neglect rather than love and nurture. Attitudes vary greatly among the world's societies toward pregnancy out of wedlock. Historically, social and even physical persecution have been common in some communities, but most modern societies recognize a responsibility toward the welfare of unmarried mothers and their children. In industrial countries, and in some developing countries through private charity, services typically include medical care and delivery and counseling regarding the decision to keep the baby or to give it up for adoption. In many countries institutional homes provide for the care both of unwed expectant mothers and of mothers and babies after delivery, in a setting sheltered from the often rigid strictures of family and community. Procedures of adoption vary considerably worldwide, but arrangements are frequently carried out by social service agencies in cooperation with legal authorities.

Whereas orphans once made up the majority of children living in institutional homes, the number of children who lose both parents through death has been greatly reduced by medical advances. Institutional and foster care are now provided predominantly to children whose home lives have been disrupted, permanently or temporarily, by illegitimacy, marital discord, financial hardship, parental irresponsibility, neglect, or abuse. While foster care might be considered preferable because it offers the intimate atmosphere of family living, some children, such as those severely affected by parental abuse or emotional disturbance, may adjust more comfortably to the more impersonal environment of an institution. Although it cannot be determined conclusively whether the increasing incidence of reported child abuse is attributable to falling standards of parental care or to improvements in detection and reporting, much effort has been invested in supervision, social education, and cooperation between personal social services and health care, education, police, and housing authorities.

Institutional and foster care

Youth welfare. The underlying aim of most social welfare services for young people, apart from those services that address immediate basic needs, is to prepare them for the assumption of responsible roles in the adult world. The majority of programs provide adult-supervised leisure-time group activities, which may range from cultural and social events to athletics to hiking and camping. Participation in such programs is high in most European countries. The Soviet youth organizations, called Pioneers and Komsomol, are the largest in the world. Some programs, such as Boy Scouts, Girl Scouts or Girl Guides, Young Men's Christian Associations, and Young Women's Christian Associations, have spread nearly worldwide, stimulating the formation of similar groups tailored to local needs. In addition to group activity, youth welfare programs also provide counseling and guidance services on a more individual basis to help meet the personal, social, educational, and vocational needs of young people.

While the above services are intended to provide constructive outlets for the energies of young people, there remain many destructive influences in society. Social services have directed increasing attention to the problem of delinquency in an effort to provide alternatives to the traditional juvenile court/institutional methods of control. In some urban areas so-called street workers approach the problem at its source. Recognition of the importance of peer groups in youth behaviour has led to the use of group therapy in many correctional institutions and in communities as a preventive service or as an adjunct to parole.

Delinquency

Welfare of the elderly. The elderly now constitute the largest single client group using personal social services worldwide. In all advanced industrial societies the proportion of infirm elderly is on the increase, and, although they constitute only a small minority of the retired population, their claim on social services is disproportionately heavy. Because social care for the elderly is often labour-intensive, most countries give full support to the promotion of family care and the expansion and rationalization of informal care on a voluntary or quasi-voluntary basis. Services include transportation, friendly visiting, home delivery of hot meals, nurse visitation, and reduced-cost medical supplies. Senior centres sponsor group activities such as crafts, entertainment, outings, and meals on a regular basis. Nursing homes, variously funded, provide medical and custodial care for those who are unable to live independently. Paradoxically, the majority of retired people lead independent lives, seldom utilizing personal social services. Indeed, fit elderly people are increasingly in demand as a source of voluntary service.

Group welfare. The settlement movement arose in response to the collective needs of deprived urban communities. Settlement houses today, and similar community centres and other organizations, seek to promote the common welfare of local groups that may differ in language, national origin, race, or religion. Whereas, in the United States, attempts were formerly made to Americanize such groups by supplanting foreign traits of language and custom with American ones, the emphasis of educational and training programs has changed; language and other assimilating skills are taught, but the preservation of cultural diversity is also promoted. In addition to educational and cultural programs, settlements may offer legal advocacy, recreational activities, and health clinics.

Throughout the 20th century the resettlement of massive numbers of refugees forced from their homes has placed great demand on social welfare services. In Europe and North America various church denominations have taken an active role in relief and other welfare work for such groups as well as for migrant and transient elements within the general population.

Refugees and migrants

Welfare of the sick and disabled. Serious illness and disability account for many of the problems addressed by social services. In addition to the need for adequate primary care, the ill and disabled also frequently face disruption or loss of income, inability to meet family responsibilities, the long-term process of recovery or adjustment to handicaps, and ongoing care in the form of medication, therapy, and the observance of dietary or other precautions.

In some countries, medical social workers are local-authority social workers who have been attached to hospitals, local general-practice health centres, and child guidance agencies. They provide the counseling and other supportive services required by the physically ill and the disabled and their families. Especially in countries where free medical care is not available to the poor, the responsibility for means-testing gives the workers an additional, advisory role with respect to their clients' financial problems. Personal social services make arrangements for domiciliary care in the form of regular visits from home-helpers and occupational therapists; special appliances and home adaptations are supplied either by personal social services or by health services. In the case of severely disabled people personal social services run day-care centres to provide relief for family care providers and small residential homes for the most dependent disabled when they no longer require hospital care.

Welfare of the mentally ill. The social aspects and consequences of mental illness were recognized early in the history of social work. The speciality of psychiatric social work developed initially as an adjunct to hospital care in urban areas. Such services have also been provided under military auspices, particularly in wartime. In developed countries today the psychiatric social worker serves at all levels of patient care; social casework may contribute to diagnosis and the course of treatment; educational and counseling services help other family members cope with the problems of hospitalization, treatment, and aftercare; close work with housing authorities and employers can facilitate the readjustment of patients into community life by means of foster care, halfway houses, sheltered workshops, and regular employment.

Psychiatric social work

Personal social services have been a major contributor to the development of community care for the mentally ill and the mentally handicapped. In the industrialized world generally, though less so in the Soviet Union, policy calls for a reduction in the number of patients hospitalized on a long-term basis; this goal can be achieved only by returning patients to their families or accommodating them in neighbourhood hostels providing adequate support and supervision. The bulk of this responsibility has fallen on local authorities and voluntary agencies, which provide the professional staff and volunteers. Treatment programs are also increasingly designed to prevent hospital admissions and to avoid compulsory admission in all but exceptional cases.

THE WORK OF THE PERSONAL SOCIAL SERVICES

Social work training. In practice the demand for personal social services does not fall into clearly defined categories. Welfare needs often overlap, and the needs of individuals often affect their families or associates. The range of skills required for effective service provision is equally complex. Inevitably, therefore, opinions differ on the training and deployment of social workers.

In the United States, the United Kingdom, Canada, Australia, New Zealand, Japan, and India the bulk of training is provided in the higher-education system, whereas in France, West Germany, Norway, and Sweden it is conducted mainly in separate institutions. Most social workers are employed in either statutory or voluntary agencies; outside the United States very few are engaged in private practice. There is much diversity in their training and deployment, but the role of social workers has broadened, making them individually responsible for a wide range of methods and client groups. In some cases specialized social workers are deployed in teams. Opinions differ on the relative effectiveness of the alternative methods of intervention—direct casework, or counseling, on the one hand and indirect social-care planning on the other. Voluntary and private agencies tend to perform more specialized roles, centred on particular client groups and age groups requiring special methods of care and service delivery.

Administration of services. *Basic organization.* There are marked national variations in the organization and funding of personal social services. To begin with, there are differences in the relative importance of the statutory, voluntary, and private sectors. Second, even if governments are the major contributors, the proportional allocation of funds for the statutory and nonstatutory sectors varies from country to country. Third, there are variations in the relative importance of central, regional, and local governments with respect to statutory funding, policy-making, and service delivery. Fourth, there are also variations in the degree of administrative autonomy granted to the personal social services.

The social service staff

The paid staff of statutory personal social services includes social workers, community workers, social care assistants, home-helps (homemakers), workers who supply mobile meals, occupational therapists, and psychologists working in a variety of field, day-care, and residential settings. Although social workers account for a small proportion of the social service workforce, they constitute the majority of its professional staff. Their job is to provide casework, or counseling, services in cooperation with individuals and families and to engage in tasks of social-care planning, such as seeing to the delivery of direct services in kind and fostering the involvement and support of informal care providers and volunteers. In most industrial societies social workers have more or less exclusive responsibility for mandatory duties related to fostering, adoption, and other work affecting parental rights as well as for the management of substitute home care or residential care for the main client groups. Probation officers act as social workers with a special attachment to the courts, the administration of probation usually being separate from that of other statutory personal social services.

The increasing orientation toward community care calls for social policies that strengthen the association between formal personal social services and informal networks of social care without losing sight of their differences. The formal public, or statutory, sector and the voluntary and private sectors all have paid career staffs whose objectives and management are bound by explicit rules. The primary tasks of the public sector are laid down by statute; most voluntary and private organizations are registered, respectively, as charities and companies. In countries such as the United States, the United Kingdom, West Germany, The Netherlands, and Japan formal voluntary and private agencies receive direct or indirect grants from the statutory sector in return for agreed amounts of contracted work. In the developing countries many welfare agencies are internationally organized and jointly financed by charitable donations and government grants.

Informal care

Informal care is spontaneously provided in the context of families, neighbourhoods, and other loosely structured community-based associations. Without these supporting networks the personal social services would be overwhelmed by demand. Consequently they often make small grants to informal self-help groups and supplement the unpaid services provided to dependents by their relatives and friends. Professional social workers and community workers are increasingly deployed in the recruitment, training, and general assistance of informal care providers. Payment for fostering is a long-established practice in many countries, and this policy has spread to the care of other groups such as the handicapped and the infirm elderly.

Personal social services are prime movers in the humanitarian trend toward caring for dependent people in their own communities, to which the high cost of residential care adds an economic incentive. It is evident that there is no clear boundary between the formal and informal sectors of social welfare. Nevertheless, informal care cannot take the place of formal services, the two sectors being mutually supportive rather than alternative sources of social welfare. Formal social services are a matter of legal obligation; their providers and users are normally strangers to each other, whereas informal care is given and received on the basis of personal relationships. Formal services have a wide membership and are delivered on a continuous basis, without regard to personal considerations. Informal care is highly localized and, although it may reflect intense loyalty and devotion, is often less reliable than formal care in the long run because family and neighbourhood networks are vulnerable to personal crisis and social change. Such care also does not usually extend to those without living relatives or other close associates. There are, of course, changes in priority within formal social services in response to trends such as the increasing incidence of reported child abuse, especially in the United States and the United Kingdom, the growing proportions of unemployed and infirm elderly, and the heightened awareness of racial inequality and injustice.

The United States. In the United States the main social assistance and personal social service programs are county- and state-administered, with substantial federal government support. Many programs are delegated to local governments, and voluntary organizations are heavily subsidized by public bodies via contracts for provision of services. The Department of Health and Human Services is the chief federal agency, and each state has a counterpart of this agency. In addition there is a small but popular and growing private market for fee-charging social services that overlaps the voluntary sector.

Federal policies for the personal social services have changed significantly since the 1960s. The Social Security amendments of 1962 put a premium on the role of rehabilitative casework, although states could also include homemaking and foster care. Between 1967 and 1977, however, income maintenance services (Aid to Families with Dependent Children excepted) were regrouped under the Social Security Administration, and primary responsibility for personal social services was transferred to the Office of Human Development. The 1974 amendments to the Social Security Act (Title XX) considerably extended the scope of eligibility for social services, giving priority to preventive work and positive efforts to improve the quality of life rather than to the traditional focus on poverty abatement. Casework, or counseling, however, lost ground to community-oriented service programs such as day-care provision, mental health centres, and nutrition programs. Problems of child abuse and alcohol and drug dependence have steadily assumed greater importance.

There has been significant growth in employer-sponsored welfare programs in the private sector and service-purchase schemes linking public, voluntary, and private agencies, accompanied by increasing use of paid volunteers. The promotion of for-profit entrepreneurial services and decentralization of funding and policy management from federal to state agencies is intended to diversify still further the mixed economy of welfare that typifies the personal social services of the United States.

Treatment of child abuse

In both the United States and Canada special treatment programs have been developed for the prevention and treatment of child abuse, but lower priority has been given to preschool and family support programs designed to encourage better parenting and child development. The U.S. Child Abuse Coordinating Program set up in 1972 is based on an interservice approach involving municipal and quasi-public bodies, one of which provides the agency officers. American child protection law is extremely complex because of its dual federal and state components, and, although the best interests of children are generally paramount, it is thought difficult to consider them in isolation from those of the parents.

The mental health care legislation of 1970 and 1972 stepped up the funding of community mental health centres in poor areas, but it was not until the Mental Health Systems Act of 1980 that priority federal funding began to reach those with the worst economic or ethnic disadvantages among the chronically ill, the retarded, and the elderly. There is a growing problem of homelessness among the more mobile patients discharged from mental hospitals, who need higher incomes and more social support if they are to resume independent lives.

Social services for elderly American citizens constitute a typical mixed economy of welfare. Amendments to the Older Americans Act of 1965 have led to the establishment of a network of more than 600 Area Agencies on Aging, which are area-wide planning and coordinating agencies. Locally sponsored senior citizen centres provide group meals and counseling, homemaker, information, referral, transportation, educational, legal, and recreational services. There are also a strong volunteer sector and a rapidly expanding private market. Provisions for the frail elderly under Medicaid and Medicare do not include long-term social care, and the poorest groups are dependent on social insurance and social assistance for the requisite finance. Many not-for-profit and for-profit agencies have developed nursing-home and other special housing schemes that are linked to various reverse-equity mortgage options. Nearly three-quarters of all the states have tax policies designed to reduce the cost of independent living for elderly homeowners with low incomes.

The United Kingdom. In the United Kingdom, as a result of the Seebohm reforms of 1970–71, the funding and organization of personal social services are highly centralized at the local authority level. In each local authority a single social services department serves all categories of client and welfare need. In Scotland, however, the probation service is separate. Personal social services are provided from catchment area offices, although some local authorities delegate this responsibility to small "patch" teams serving neighbourhoods. Roughly half of local authority funding comes from the central government; nevertheless, within strict cash limits, the local authorities exercise wide discretionary powers over the organization and deployment of personal social services. Social work training is centrally regulated, and there is only one (general) qualification in professional social work.

Status of social work

Although income maintenance was transferred to the central government in 1948, local-authority social workers continue to provide small cash grants to families with children when shortage of money is deemed likely to cause a family breakdown. In Britain the separation of income maintenance and social work services was part of an overall policy designed to end the historically stigmatizing association between public assistance and social work in particular and the more general association between poor relief and social welfare. It was also hoped that social work and the other personal social services would shed their low status and become more acceptable in all sectors of society. This philosophy was adopted by the Seebohm Report of 1968 and reflected in the Local Government and Social Services Act (1970), but the resources for a truly universal network of services oriented toward preventing problems were not forthcoming.

British child care law developed in piecemeal fashion over a long period. Nevertheless, it places a clear obligation on the local authorities to protect children at risk and to receive them into care when their welfare is at stake because their parents are deemed unable to provide satisfactory care. Under certain circumstances local authorities can assume full parental rights until a child reaches the age of 18. Separate provisions are made for compulsory admission into care through juvenile court proceedings, when children are "in need of care and control" on various defined grounds, or through matrimonial, divorce, separation, wardship, or criminal proceedings. Care orders may also be issued under the Children and Young Persons Act of 1969, as amended by the Criminal Justice Act of 1982, when children or young persons are found guilty of an offense that, if committed by an adult, would be punishable by imprisonment. Observation and assessment centres and secure community homes with educational facilities on the premises are run by the Department of Health and Social Security.

There are strict regulations on boarding out children in care with foster parents, including thorough investigation of prospective homes, frequent inspections, and the keeping of case records. In English law, adoption is an almost complete and irrevocable transfer of a child from one family to another. Adoption orders are made in the Magistrates', County, or High courts, and adoption proceedings can be initiated only by registered, not-for-profit adoption agencies (including local authorities).

Although English law makes extensive provision for the protection of children, personal social services have a well-established tradition of working with children and families on the basis of a cooperative partnership whenever possible. This tradition includes avoidance of recourse to legal intervention or residential care unless it is in the best interests of the children concerned.

British mental health legislation

With regard to the mentally ill and mentally handicapped, the British Mental Health Act of 1959 anticipated the general trend toward voluntary treatment and voluntary hospital admission, and legislation in 1982 introduced even stricter criteria for the protection of patients' rights. Since 1983 certain procedures in the admission and discharge of mentally ill patients have been reserved to a new category of specially trained approved social workers. In cases of compulsory detention, patients have a strengthened right of appeal to the Mental Health Review Tribunals, and there are special provisions for the guardianship of certain types of discharged patients. Although there are still serious deficiencies in community care for the mentally ill or handicapped as well as the elderly and the physically handicapped, various joint government and local-authority funding schemes have helped to reduce the numbers in institutional care.

Services for the elderly and the physically handicapped account for roughly half of all British local-authority personal social service expenditure, mainly because of the steady increase in the numbers of the frail elderly and the high cost of care for the minority who live in residential homes. Extensive efforts have been made to improve the quality of domiciliary support, but relatives carry the main burden of home care. There are special housing schemes for the elderly sponsored by statutory, voluntary, and private agencies, and a growing number of local authorities employ paid volunteers to visit elderly people and help them with a range of daily tasks. Perhaps the best guarantee of independence in old age, however, is an adequate income from social security (see below *Social security: government welfare programs*).

The formal voluntary sector makes its own important contribution to the care of all the major need groups, although it is heavily dependent on direct and indirect financial support from both central and local governments. Within the voluntary sector the churches have always played a major part in the provision of both community and residential care. Nevertheless as statutory funding has lagged well behind demand, the private market, especially with respect to services for the elderly, has begun to expand.

Australia. In Australia the state governments and the local authorities, with some federal funding, have the main responsibility for personal social services. Each state has a welfare department, usually an amalgamation of the former children's and public relief departments, providing a general range of casework and community services. Some of the municipal authorities also provide welfare services in conjunction with their public health, educational, housing, and legal aid services. In addition there is a well-established tradition of volunteer work that is subsidized by statutory bodies, sometimes provided on a dollar-for-dollar matching basis. Some of the religiously based charities, such as the Brotherhood of Lawrence, the Society of St. Vincent de Paul, and the Salvation Army, are pioneers in work with severely deprived groups.

France. In France personal social services are not administratively autonomous. A variety of social workers and social care workers are employed by other major public services, such as social security, hospitals, community health care, education, housing, and the courts. There are several types of social worker, including the family social worker (*assistante sociale*) and other specialists in child protection, medical social work, and court work; the homemaker (*travailleuse familiale*); child development workers specializing in the care of handicapped children; social allowance guardians with special responsibilities for families in serious financial difficulties; and the community worker (*animateur socioculturel*), who serves neighbourhood groups. Apart from the statutory services there is an extensive network of semipublic agencies (*caisses*) based on trade unions, family associations, and religious denominations, as well as a variety of independent, not-for-profit organizations financed by state grants.

Social work specialists

The French system of child care is explicitly family-oriented. It is based on services financed by the Ministry of Health and the Ministry of Justice, in cooperation with other family income support services. The judicial services are called upon only if parents refuse to cooperate. Social workers are employed in maternal and child health centres and in municipal and family allowance agencies. Special child-protection officers work closely with pediatric nurses in cases of actual or suspected child abuse, and the procedures for removing children from the home and for providing substitute care are in principle similar to those in Britain. Child care services are unified at *département* level, and there is close liaison between the courts and specialized medical services in child protection work.

The reforms of the 1960s and 1970s improved the quality of French social services not only for children but also for the mentally and physically handicapped and the elderly. Since the late 1950s domiciliary care and sheltered housing provisions have been strengthened and diversified, objectives that were upheld in the Laroque Report of 1960 and in the provisions of the Sixth (1971–75) and Seventh (1976–80) Plans. The plans specifically referred to the growing need for more trained staff and for more sheltered housing, residential homes, and nursing homes in addition to increased community care and more generous income support within a better-coordinated framework of health and welfare programs at neighbourhood, local, and regional levels. Social care services for the mentally ill are mainly controlled by the health and employment authorities, but the social workers attached to the regional and local *caisses* play a major part in the provision and coordination of community care.

West Germany. In the Federal Republic of Germany there is a long tradition of cooperation between the statutory and voluntary sectors and between these formal agencies and the informal networks of family and neighbourhood care. These arrangements exemplify the principle of subsidiarity (the belief that informal care should, whenever possible, take precedence over state intervention) in European Roman Catholic welfare philosophy, although in West Germany all of the major religious denominations play an important part in social welfare service. The health care provisions of the income maintenance services do not extend to the longer term welfare needs of the elderly mentally ill or handicapped or those of the physically disabled. These are met largely from public aid. About half of the total expenditure on welfare services comes from the Aid for Care program, which channels much of its funding through the larger not-for-profit charitable organizations.

Roman Catholic welfare philosophy

Sweden. The modern Swedish welfare state emerged from poor-law and charitable traditions in which the churches were prominent. Since the years between the two world wars, the scope and funding of statutory agencies have steadily increased. Local authorities, assisted by central government grants, provide most personal social services and a social assistance scheme, in which investigation of needs and means is undertaken by social workers. There has been a trend toward the unification of specialist agencies into local joint welfare boards, but the municipal communes still exercise considerable local discretion in the organization of their services. Although the extensive role of the state in Swedish welfare has elicited much comment, the scale of voluntary effort is equally noteworthy, as it is in Norway and Denmark.

Israel. Israel has a complex system of welfare services distributed by central ministries, with subdivisions for all the major need groups, including services for wounded soldiers and surviving dependents, a Jewish agency with special responsibilities for immigrants, and a universal labour union (Histadruth) with extensive roles in insurance and welfare and a long tradition of mutual aid based on local collectives (kibbutzim) and cooperative villages (moshavim). This has been supplemented by a network of community centres funded by the central and local governments and by membership fees and overseas donations.

Japan. Japanese social welfare provision is uniquely

Work-based social services

reliant on employer- and work-based social services, although there is also an extensive but relatively underfunded system of statutory local-authority personal social services for the major need groups. Social workers in these municipal agencies are responsible for both discretionary income support and protective social care. In major cities they cooperate with a growing number of voluntary agencies, of which the Minsei-iin is the oldest and largest. As in the case of income support, health care, and housing, access to welfare services for most Japanese workers largely depends on the size and financial stature of the organizations employing them. Although traditional familial ties are still pervasive, they are weaker in the large cities, as a result of social and geographical mobility. At the same time the number and proportion of the dependent elderly show a marked increase. Accordingly, Japanese policy has turned toward the expansion of statutory services, and much has been done to foster neighbourhood networks of mutual aid that go beyond the traditional notions of kinship and obligation.

Socialist countries. It is as difficult to make generalizations about social welfare in socialist countries as it is in the case of the democratic societies referred to above. Nevertheless, in the foremost socialist societies the state provides the formal social services, and the workplace and the trade unions play a large part in service management and delivery. In these planned economies, where work is both a civic right and a formal obligation, social assistance for the unemployed is minimal. In the absence of firm data on this area of provision it must be presumed that families shoulder the main financial responsibility for many of the exceptional needs covered by discretionary provision in the West.

There are no professional social workers in either the Soviet Union or China; social service workers perform similar functions, especially with regard to child protection and delinquency. The Soviet Union has a long tradition of nurtured interdependence between the formal social services and a complex network of mutual aid, lay counseling, and supportive services. The latter are distributed by street, block, and house committees in the towns and cities, by agricultural collectives in the countryside, and by the parallel agencies of the trade unions and the Communist Party.

Self-reliance and mutual aid

The Chinese system of social welfare is also strongly based on the industrial or agricultural workplace. Many essential social services, such as health care, are funded from the profits of collective work and administered by neighbourhood committees. Throughout the People's Republic the guiding welfare principles are self-reliance and mutual aid. Although, in exceptional cases, families receive grants-in-aid to help with care for dependent relatives, Article 13 of the 1950 Marriage Law states that children and parents are jointly responsible for mutual support in hardship and old age. At the same time extensive and sustained support is given to schemes of mutual support that extend to neighbourhoods and workplaces, and priority is given to the needs of dependent persons without families of their own.

Decentralization of services

The trend in Yugoslavia has been toward the decentralization of personal social services and the promotion of neighbourhood voluntary work. State-sponsored organizations such as the Alliance of Friends of the Young and the Pensioners' Associations act in conjunction with a growing network of professionally staffed social work centres financed by the 600 communities that are the basic units of local government. Similar developments can be seen in the other socialist countries of eastern Europe, where, as in the Soviet Union and China, there is a strong commitment to the expansion of informal provision for family dependents and neighbours.

Developing countries. In former colonies, such as Ghana, Sri Lanka, Jamaica, India, the Philippines, and Francophone Africa, the basic welfare services grew out of modified versions of the European poor laws, charitable and missionary activities, and the introduction of Western juvenile justice procedures. The oldest school of social work in Latin America was founded in Santiago, Chile, in 1925, and the Ratan Tata Foundation established the first Indian school in Bombay in 1936. New training institutions have since proliferated throughout the so-called Third World, many of them sponsored by the United States Agency for International Development.

Disaster relief

In developing countries, where formal social services are generally under-resourced, traditional networks of informal care are the main source of assistance in adversity and old age. High rates of migration and unplanned urban growth, however, have weakened these networks in impoverished rural areas and overwhelmed the limited public services in new cities and towns. Indigenous overcrowding and poor housing, unemployment and low wages, and inadequate sanitation and endemic disease are not responsive to Western methods of personal social service intervention. Priority, often within severe economic restraints, must go to major programs of preventive health care, family planning, basic education, income support, and slum clearance. Nevertheless, community development work is also important in these processes of social development. In the poorest rural areas, where the majority of people live at or well below subsistence level, disaster relief is heavily supplemented by international aid agencies such as the United Nations and its associated agencies, including the World Health Organization and the International Labour Organization (ILO), charities such as Oxfam and the Save the Children Fund, and the governments of richer nations. In the longer term the enhancement of living standards depends on horticultural improvements, reforestation, water conservation, and those irrigation schemes that can be managed within small communities.

CONCLUSION

It is clear that the processes of economic and social change create new prospects and new hazards for every generation. This requires constant adjustment on the part of the social services. Political considerations and levels of resources largely determine how social services are organized and how responsibility is apportioned between the statutory, voluntary, and private sectors. Even in prosperous societies the scale and diversity of needs is such that the formal social services are obliged to utilize and support informal systems of social care and mutual aid. The idea of the welfare state as a universal provider for largely passive populations has never had any reality in fact, nor much serious support in political theory. There is widespread evidence of a general trend toward the development of closer links between the formal and informal systems of social care, although this might lead to further variation in social welfare services as societies become more sensitive to their indigenous cultural diversity and develop their own responses to change. (R.A.P.)

Social security: government welfare programs

In international usage the term social security has come to mean all collective measures established by legislation to maintain individual or family income or to provide income when some or all sources of income are disrupted or terminated or when exceptionally heavy expenditures have to be incurred (*e.g.,* in bringing up children or paying for health care). Thus social security may provide cash benefits to persons faced with sickness and disability, unemployment, crop failure, loss of the marital partner, maternity, responsibility for the care of young children, or retirement from work. Benefits may be provided in cash or kind for medical need, rehabilitation, domestic help during illness at home, legal aid, or funeral expenses. Social security may be provided by court order (*e.g.,* to compensate accident victims), by employers (sometimes using insurance companies), by central or local government departments, or by semipublic or autonomous agencies.

The ILO uses three criteria to define a social security system. First, the objective of the system must be to grant curative or preventive medical care, to maintain income in case of involuntary loss of earnings or of an important part of earnings, or to grant a supplementary income to persons having family responsibilities. Second, the system must have been set up by legislation that attributes specified individual rights to, or that imposes specified obligations on, a public, semipublic, or autonomous body. And third, the system should be administered by a public, semipublic, or autonomous body.

In its statistics the ILO includes provisions according to which the liability for the compensation of employment injuries is imposed directly on the employer, although such schemes do not strictly meet the third criterion above. For this reason employer liability is included here.

An alternative but wider term for social security in the countries that are members of the European Economic Community is social protection, which includes voluntary schemes not set up under legislation. In some countries the term social security is used in a narrower sense. For example, in the United Kingdom only statutory benefits in cash are regarded as social security. The term social services is used to cover social security; health, education, and housing services; and provisions for social work and social welfare. In the United States the term social security is restricted to the federal social insurance system (OASDI) as distinct from state benefits and "welfare," which in Europe would be called social assistance. In some countries (*e.g.,* Denmark and the United Kingdom) the reduction of poverty historically has been a central aim of social security policy, and the concept of maintaining income has been grafted on at a later stage. In other countries, such as France, measures to deal with poverty have been seen as quite separate from the income maintenance aims of social security.

A report published in 1984, prepared by 10 international experts appointed by the director of the ILO, set out the ultimate aims of social security.

> Its fundamental purpose is to give individuals and families the confidence that their level of living and quality of life will not, insofar as is possible, be greatly eroded by any social or economic eventuality. This involves not just meeting needs as they arise but also preventing risks from arising in the first place, and helping individuals and families to make the best possible adjustment when faced with disabilities and disadvantages which have not been or could not be prevented. . . . It is the guarantee of security that matters most of all, rather than the particular mechanisms such as contributory or tax financing, the insurance or service model of delivery, or the ownership of facilities (public/private, profit/non-profit) by which that guarantee is given. . . . The means should not be confused with the ends.

Approximately 140 countries have some type of social security scheme. Nearly all of these countries have schemes covering work-related injury and old-age and survivors' pensions. Well over half have provisions for sickness, and nearly half have provisions for family allowances. The least commonly provided schemes are for unemployment, though at least 40 countries have them.

THE RATIONALE FOR SOCIAL SECURITY

Industrial societies

Because general social security schemes based on compulsory insurance did not develop until the last 20 years of the 19th century, it has often been argued that social security in its modern form has been a response to industrialization, which caused large numbers of people to become dependent for their security solely on earnings from employment. Indeed many families became dependent on one male earner and thus on his capacity to find work, to undertake it, and to remain in it. Moreover, industrialization led to the migration of people toward centres of work, thus separating them from the support given by the wider family. In addition, the development of compulsory education prolonged the period during which children were dependent on their parents; later the system of enforced retirement created dependency at the other end of life. This situation is contrasted with the idealized image of the extended rural family with access to land, on which both husband and wife worked, children started work early, and old people continued to work until they became too frail or disabled to do so. On the basis of this oversimplification, some theorists have proposed that social security developed out of a need peculiar to industrial societies and that there is less need or no need for such programs in the rural areas of developing countries today.

It is true that support from the extended family, often enforced by local custom and religious beliefs, contributes to the survival of peasant societies. But by no means all the rural populations of developing countries have access to land, and many people work for wages in agricultural estates and mines. Moreover, peasant farmers are subject to formidable risks of crop failure, quite apart from the risks associated with the shorter average life span that characterizes developing countries. Although there is a need for social security in rural societies, the importance of specific risks may vary from region to region. Moreover, the irregular incomes in cash and kind emanating from agriculture do not lend themselves to the payment of regular social insurance contributions. Thus, what may be lacking is the economic and administrative base for providing such security. Furthermore, provision for sickness and old age is not generally seen as the highest priority by peasant farmers overwhelmed by problems of weather and debt.

Agricultural societies

While industrialization has undoubtedly added to the need for social security by breaking up the extended family and leading to urban poverty, it is by no means the sole reason why the system evolved. Two of the first three countries to make provision for old-age pensions were primarily agricultural societies—Denmark in 1891 and New Zealand in 1898. The Danish scheme was clearly an attempt to alleviate rural rather than urban poverty. And it is notable that the first province in Canada to develop compulsory health insurance (1962) was Saskatchewan, which was overwhelmingly agricultural. Thus, statutory social security may evolve for a variety of reasons. Moreover, it depends to a considerable degree on the economic level attained by the groups that might be covered and the administrative capacity of the country to operate such a scheme. It is certainly the case that, as countries become wealthier, there is greater willingness to defer consumption by paying insurance contributions or taxes.

HISTORICAL EVOLUTION

In many societies charity has been the traditional way in which provision was made for the poor. Charitable giving has been encouraged by many different religions, and in many parts of the world religious agencies have long collected charitable donations and distributed help to those in need.

Obligations on communities to pay taxes to provide for the poor can be traced back for hundreds of years in a number of different societies. For example, part of the function of the Christian tithe or the Islāmic zakat was to provide for the poor. Town poor laws were passed in Germany from 1520 onward, and a law of 1530 clearly placed on towns and communities the obligation of sustaining the poor. In 1794 the Prussian states assumed the responsibility of providing food and lodgings for those citizens unable to support and fend for themselves. From the 16th century it became recognized in England that there were people who could not find work, and legislation was passed to provide work for the poor and houses of correction for rogues and idlers. From 1598 a clear obligation was placed on parishes to levy local taxes and appoint overseers of the poor to give relief to those who could not work and to provide work for those who could. This was the essence of the Elizabethan Poor Laws, an early provision of social assistance.

Poor Laws

The Elizabethan Poor Laws were poorly enforced in the 17th century but widely used and liberalized by the end of the 18th century. A new Poor Law enacted in 1834, and reflecting a harsh moral view of poverty, required the poor persons to be admitted to the workhouse so as to receive relief only in kind, with occasional exceptions, but this again was by no means uniformly enforced, though it added greatly to the unpopularity of the Poor Laws. Some U.S. states copied the Elizabethan Poor Laws but exempted recent immigrants. The English Poor Laws were also introduced in Jamaica in 1682 for destitute European immigrants and much later in Mauritius (1902) and Trinidad (1931). In Latin America the Spanish colonists, instead of establishing a public relief agency, gave grants to charities to provide "hospitals" for the poor (*beneficencias*), and the Portuguese promoted lay brotherhoods such as the Misericórdia.

The first general social insurance scheme was introduced in Germany in 1883. The scheme drew upon three types of precedent. The first was the ancient system of guild collection boxes—funds to which each member of a particular trade was required to contribute at regular intervals; such funds were originally used for hospital and funeral expenses and for food and lodging for aged and disabled members. By the middle of the 14th century these arrangements were covered by statutes and regulations. Relief funds were later established by associations of miners. The second precedent was a Prussian ordinance of 1810 that placed on masters a duty to ensure that their servants were given medical attention in case of illness. From 1849 communities could make bylaws requiring both employers and employees to contribute to relief funds, and a law of 1854 introduced compulsory health and accident insurance for miners. The third precedent was the employer's legal liability to pay damages for accidents caused by negligence. As a result of this liability, which was widened in 1871, many employers took out private insurance. The system did not work well because the burden of proof lay with the worker, who normally had to incur high legal costs and delay before he could hope to obtain lump-sum compensation.

Chancellor Otto von Bismarck's 1883 sickness insurance law provided to employees in defined types of industry both medical care and cash benefits during a period of sickness, to be paid for out of contributions from both employees and employers. This was followed by a law of 1884 making accident insurance compulsory. The schemes were operated by numerous funds controlled by the insured and their employers. Finally a law establishing a pension for all workers in trade, industry, and agriculture from the age of 70 was passed in 1889. This was directly administered by the Imperial Insurance Office. Austria followed part of the German example in 1888, Italy in 1893, and both Sweden and The Netherlands in 1901.

Friendly societies

Bismarck's political aim in introducing social insurance had been to address the legitimate grievances of workers so as to check the growth of socialism and avert revolution. A proportion of previous earnings were to be paid in cases of sickness, injury, widowhood, and old age. Employers and employees were to work together in implementing the scheme. In Austria part of the driving force was the Christian Socialists' aim of improving the worker's position. Although Britain had been the first country to industrialize, the developments in Germany and Austria originally attracted little British interest because of an aversion to state intervention, an apparently lesser likelihood of revolution, and the slower development of British socialism. In Britain self-help through friendly societies and savings banks was seen as the solution. The friendly societies were run by skilled workers with no employer participation and provided flat-rate cash benefits for sickness as well as treatment by the society's doctor, who was normally paid a flat rate per member insured—a so-called capitation payment. By 1870 membership had grown to 1,250,000 and by the early 20th century to 7,000,000. Apart from the regulation of friendly societies, the only social security legislation passed in the United Kingdom during the 19th century was to widen the liability of employers to compensate workers for personal injury arising out of work. By a law of 1897, compensation could be obtained whether or not the employer had been negligent.

Further action arose in the United Kingdom out of social concern about poverty, which was systematically investigated both in London and in York. In 1899 the government carried out an inquiry into the incomes of 12,000 elderly people. The influential precedents for action were those of New Zealand and Denmark, which had made provision for old age without establishing social insurance schemes, in contrast with Germany, where the scheme was based on insurance. In 1908 in Britain, pensions at age 70 were introduced in a noncontributory, income-tested basis, partly because such a scheme could bring immediate relief to the aged poor, as opposed to a contributory scheme, which could only pay pensions to those who had paid contributions. The social insurance approach was, however, applied to sickness and also to unemployment in certain occupations three years later. This compulsory scheme, including the first state scheme of unemployment insurance, again reflected Britain's concern to address the main causes of poverty. Benefits and contributions for sickness and unemployment insurance were flat-rate, building on the precedents established by the friendly societies and ensuring the maximum impact on the living standards of low earners. From 1925 the social insurance approach began to be extended to provide for widowhood and old age.

Unemployment insurance

Unemployment insurance was subsequently introduced in Austria and Belgium (1920), Switzerland (1924), Germany (1927), and Sweden (1940). In the case of health insurance, Denmark, Norway, and Sweden promoted voluntary health insurance before making such schemes compulsory, much later than in Britain or Germany. In France voluntary insurance had long been less developed, and mutual insurance societies had long been regarded by government with suspicion, and therefore suppressed. When they ultimately were allowed to expand, around the end of the 19th century, the bulk of their membership was middle class. During the second half of the 19th century larger employers established their own pension and welfare institutions. An employers' liability law was passed in 1898 for accidents at work irrespective of negligence, and in 1910 modest contributory pensions were introduced for industrial and agricultural workers. This law met with limited success, owing to opposition on the part of workers, noncompliance among employers, the loss of rights on change of job or bankruptcy of the employer, and the erosion of the value of pensions during inflation. Health insurance, though provided for in a law of 1920, did not come into effect until 1930, owing to the opposition of the medical profession.

A major innovation came in Belgium (1930) and France (1932) with the introduction of family allowances, although New Zealand had introduced a limited means-tested scheme in 1927. These derived from the ideas of social Christianity regarding "the just wage" and had originally been introduced by Christian employers on a private basis; special funds were later set up to equalize financial burdens among employers. Family allowances became relatively generous in France, partly because of concern to increase the birthrate after the heavy loss of men in World War I. (There is, however, no clear evidence that family allowances have any impact on birthrates.) France later introduced family allowances in many of its colonies during the 1950s.

During the interwar period social insurance schemes were introduced in more and more countries in Europe and Latin America. The most common model was that established in Germany—autonomous funds paying earnings-related benefits. The first group to benefit in Latin America was civil servants, followed by those working in railways and public utilities. There were separate schemes for hospital personnel in Argentina (1921), shipbuilders in Uruguay (1922), merchant seamen in Chile (1925), and dockworkers in Peru (1934). Thus the foundations were laid for the complex social security schemes in Latin-American countries that later reformers tried to amalgamate. The first comprehensive scheme for industrial workers was established in Chile in 1924. In African colonies many schemes of social security were originally introduced only for expatriate Europeans.

U.S. Social Security Act of 1935

The Great Depression of the 1930s finally overcame opposition in the United States to federal intervention in social security. Earlier government activity had consisted of piecemeal initiatives at the local or state level. The Social Security Act of 1935 not only provided federal grants for state public assistance to the aged, blind, disabled, and dependent children but also established a federal old-age insurance scheme and federal financial backing for state unemployment insurance plans that met federal guidelines. Provision for survivors was added four years later and for disability later still. A quite different approach was taken in New Zealand, which introduced in 1938 the first universal non-means-tested pension from age 65, available only on a test of residence and financed in part from a special social security tax on income.

A major influence on world developments was the British government's report by Sir William (later Lord) Beveridge in 1942, which argued for the maintenance of full employment as a responsibility of government, family allowances for all children after the first, comprehensive health care for the whole population, and a unified national scheme of social insurance run by the state with the safety net of a unified national scheme of social assistance. The aim was to eliminate want or poverty. By 1948 the scheme had been introduced in the United Kingdom with some compromises and modifications. A drive, inspired by Pierre Laroque, to unify social insurance in France after World War II was less successful.

During the period of rapid world economic growth from 1945 to 1973 there was a further major expansion of social insurance to more countries, covering higher percentages of population and wider risks. The expansion was particularly notable in Latin America and in certain French colonies in Africa, where comprehensive social insurance schemes were introduced following the original schemes for family allowances. In the British colonies a different approach was taken: provident funds (see below) were widely developed for particular categories of workers. Discrimination on racial grounds was widely prohibited but still persisted in South Africa.

Innovations after World War II

The major innovations in social insurance after World War II were the protection of pensions by linking them to the inflation rate; the development of dynamic pension formulas that indexed past pension contributions to the level of earnings at the time of retirement; the introduction of flexible retirement providing for part pension and part-time earnings in the last few years before full retirement; the movement toward equal rights for men and women; attempts to provide for all disabled people on the basis of the degree rather than the cause of disability (*i.e.,* whether or not work-related); the growing recognition of extra needs arising from disability and of the needs of persons caring for the disabled; special provisions for one-parent families; the development of parental allowances in addition to family allowances; the integration of child tax allowances with family allowances; and the extension of the same health-care rights to all citizens.

METHODS OF PROVISION

Legal liability. Many countries that once held employers themselves legally responsible for compensating victims of work accidents and for paying for their medical care have now adopted state schemes of compulsory insurance. From the point of view of the worker the problems with the former system include the delays and costs of going to court and the possibility that the employer may be uninsured, unable to pay, or bankrupt by the time the case is heard. Moreover, a lump sum awarded by a court cannot be invested so as to provide a secure inflation-protected income for life. And when the employer is privately insured, the insurance company is in a position to offer the worker a small lump sum soon after the accident, knowing that the worker may well accept it rather than incur the delay, costs, and uncertainty of a court case to obtain the full value of the claim. From the national point of view such a system is wasteful because of the legal costs and the high administrative costs incurred by the insurer and passed along to the insured by way of higher premiums. The argument in favour of this approach is that insurers quote premiums for individual employers according to their experience of risk, which provides financial incentives for industrial safety. But insofar as such incentives are effective, premiums for a national program of accident insurance can also be risk-rated.

Private versus statutory compensation

In some countries, when a statutory insurance scheme of occupational injury has been introduced, the right of the employee to sue the employer for negligence is removed. In other countries the employee is free to supplement industrial injury benefits by making a claim for negligence.

The legal liability approach is still used in many developing countries for the general provision of medical care. Thus large employers or employers of labour in mines or specific agricultural estates (*e.g.,* sugar, tea, and rubber) are required to provide clinics and hospitals for their employees and dependents. This is one way of ensuring that health services are provided to people working far from the main urban health services. It is, however, difficult to ensure that employers comply with the spirit of the law. Moreover, employees may suspect that the doctors and nurses working in such services owe their primary loyalty to the employer and thus tend to economize on the treatment or are reluctant to certify time off for sickness. A further problem is that it is uneconomic to provide government services in these areas for the remainder of the population who are not employed by the major local employer and is difficult to integrate employers' services with government services.

In several countries employers are required to provide defined levels of cash benefits during short periods of sickness (*e.g.,* six to eight weeks). This avoids the administrative complexity of a social insurance benefit paid by a national scheme or a sick fund supplemented by an employer's scheme. Provision may be made to protect the workers' rights if the employer goes out of business.

While social insurance is preferred to the employer liability approach by social security experts because it can give better protection, employer liability is still widely used in developing countries not only for employment injury but also for sickness and maternity benefits and employer severance payments.

Lump-sum benefit

Provident schemes. Many developing countries require certain employers to contribute to a provident scheme providing a lump-sum payment in the event of death or disability or on retirement. Such a scheme differs from a social insurance scheme in that each worker usually has his own personal account from which he or she can draw if certain contingencies arise; there is no pooling of risks among members as there is in a social insurance scheme. Such schemes, which avoid the administrative complexity of paying a regular cash benefit, may be a step toward a full-fledged social insurance scheme. There are three disadvantages of such schemes from the point of view of the beneficiary. First, provision is inadequate for risks occurring early in working life. Second, the funds are generally invested in government stock with a rate of interest fixed in money terms that may be below market rates; the real value of the accumulated savings may thus be substantially eroded by inflation by the time of retirement. Third, a lump sum once received cannot normally be securely invested to provide an income protected against inflation. Moreover, it may be frittered away or unwisely invested. From the point of view of governments, however, such schemes are attractive in that they generate forced savings that can be used to finance national development plans.

Social insurance. The use of compulsory insurance as a mechanism to provide medical benefits and cash benefits in the case of sickness, disability, widowhood, and old age became acceptable to legislative bodies fearful of accepting extended state intervention that would require higher taxes to finance pensions or other benefits. In societies where self-help by voluntary insurance had been widely supported, the further step of compulsory insurance was seen as a means of making workers "good" by legislation. Because the schemes were financed by contributions levied on both employers and employees with, in some cases, modest state subsidies, unacceptable levels of national taxation were avoided; indeed, as such schemes reduced the need for social assistance or poor relief, the burdens on local taxation were reduced.

Reduction of tax burden

Compulsory insurance contributions are essentially a tax on earned income. Employers try—and probably succeed in most circumstances—to shift the burden of their share of the contribution either to consumers in higher prices or more probably, in the long run, to their employees by paying them less in cash. Thus employers' contributions are in most cases not paid at the expense of profits. However, the fact that the worker is told that the employer has to pay a proportion of the total contribution helps to make such schemes acceptable to employees, quite apart from the clearly defined benefits that flow from paying their share. Compared with the complexities of an income tax, a social insurance tax is a simple one to collect. But if the level of contributions is high, it creates incentives for workers to become self-employed in what has come to be called the "black," or "underground," economy and for employers to avoid contribution liability by employing contract labour rather than full-time staff.

In terms of meeting social needs or reducing poverty the social insurance method of provision has a number of disadvantages. Over the years many countries have tried to find means of countering these. First, the analogy with private insurance, which made such schemes politically salable, carries with it the social disadvantage that benefits should be paid to those who have contributed. Thus such schemes cannot provide benefits to persons who have never worked, for example, persons who have become disabled before reaching the age to enter employment, those incurring risks very soon after entering employment, and women (or men) who do not enter the labour force because of family responsibilities. Second, the expectation that benefits should be related to the amount paid in discriminates against individuals, usually women, who because of family responsibilities have fewer years in paid employment. Moreover, workers with dependent spouses and children have greater needs than single persons, though the assumption of marital responsibilities—or the converse assumption of marital dependency—is not strictly speaking an insurable risk. Third, where contributions are related to earnings, the benefit will be low for low earners, thus failing to protect them from poverty. The alternative approach, which some countries have adopted, of flat-rate contributions and flat-rate benefits can impose heavy burdens on low earners with family responsibilities. Fourth, it is difficult to bring the self-employed and those working for small employers (*e.g.,* agriculture or domestic work) into such a scheme.

Over the years many countries that started with a purist insurance approach have modified their schemes to try to overcome many of these disadvantages. For example, extra benefits have been provided to persons with dependents. Contributions have been credited to persons outside the labour force for reasons of family responsibility, sickness, or disability. Minimum benefits have been introduced above those strictly warranted by low earnings-related contributions, or the benefit formula has been weighted in favour of lower earners. And some countries have made contributions earnings-related or integrated them with income tax while still paying flat-rate benefits.

Demogrants

Benefits to all residents. Because of the disadvantages of the social-insurance approach, some countries have made certain benefits available to all residents and financed them out of taxation. When the benefit is paid on the basis of age it is sometimes called a demogrant. The most common benefit selected for this approach is the family allowance. The underlying philosophy is that provision for children should not depend on whether the parent is or has been in paid employment. Some countries have adopted this approach for pensions or at least for a minimum pension. In some cases this evolved from an earlier provision of an income-tested pension. In other cases this was the only way forward for governments in which the power to levy social insurance contributions did not rest at the federal level. Some countries have more recently applied this approach to provision for the disabled in the form of a minimum benefit based only on the extent of disability. It is increasingly applied to medical benefits on the grounds that all citizens have a right to health care.

Means and income tests

Social assistance. The development of social insurance and demogrants has not removed the need for social assistance to fill gaps in provision in advanced societies. Social assistance is based on need and thus requires declarations of income, family size, and other circumstances. Thus it is provided on the basis of a means test that takes into account not only income but also capital; persons with a specific level of savings may be ineligible. Alternatively it may be only income-tested, the income from capital being assessed in the same way as other income. Often those who have been given the task of operating the scheme (*e.g.,* social workers) have been allowed considerable discretion in deciding whether to give assistance and how much to give in certain types of cases. Not all basic rules are known to claimants. The tendency in industrialized countries has been to try to transform assistance into a right with published scales and regulations and opportunities for appeal. With codification has often come standardization and the unfortunate removal of some of the flexibility available under discretionary systems.

In some countries social assistance plays a residual role, providing a less favourable level of support than is normally available from social insurance benefits. In other countries (*e.g.,* the United Kingdom) social assistance plays a considerable role in supplementing social insurance benefits for those without other sources of income such as sick pay or employers' pension schemes as well as providing for those without rights to benefits (*e.g.,* one-parent families other than widows) or those whose benefits have run out because they are paid only for a specific number of months (*e.g.,* unemployment benefits).

Criticisms of social insurance

There are disadvantages of the social assistance approach. First, it penalizes saving and earning because income from any source is normally deducted from the assistance that would be payable, and persons with a certain level of savings may be ineligible until they have used them up. Second, it tends to stigmatize the recipient; and third, partly for this reason and partly because of the difficulty of knowing detailed rules of entitlement, there are considerable numbers of people who would be eligible but do not make claims. Partly because of this problem of stigma, social assistance programs are called by a variety of different names in the hope that they will be more acceptable to applicants. For example, the term used is supplementary benefit in the United Kingdom and GAIN (guaranteed income) in British Columbia. Eligibility rules differ considerably from country to country and are usually determined locally rather than centrally. Moreover, schemes are generally financed wholly from taxes—often local taxes. In the United Kingdom, where rules are determined centrally, persons in full-time work are not eligible. In the United States only households headed by a single parent are eligible for the Aid to Families with Dependent Children program, which creates incentives for desertion or fictitious desertion. There are, however, further programs for the blind, the disabled, and the aged.

The United States uses what is essentially the social assistance approach for meeting the medical care needs of low-income persons under the Medicaid program. Ireland operates a scheme by which persons with low income can apply for a medical card that gives them more extensive rights to free health care than are available to other income groups. Those with low incomes in South Korea can also apply for cards giving rights to free or nearly free health care.

A number of countries in Europe have developed separate income-tested provisions to help persons with low incomes meet the cost of rent or property taxes. Such housing allowances are available to persons whether in work or not and take account of family composition as well as rent payable.

Negative income tax. Partly because of the stigma attached to social assistance, the difficulty the potential beneficiaries have in understanding eligibility, and their reluctance to apply, it is often proposed that the information provided to the state from income tax returns should be used by the state to determine the need for cash payments to persons with low incomes. The ability to do this depends on persons' being required to make income tax declarations by a certain date however low their incomes, which is not the practice in every country. Canada has a program to supplement on the basis of this information the incomes of persons drawing pensions. This approach is much less appropriate for younger people whose financial circumstances change considerably from year to year and month to month due to sickness, unemployment, job changes, marriage breakdown, remarriage, and so on. People need money when poverty strikes, not after the end of the income tax year.

CASH BENEFIT PROGRAMS

Provisions for cash benefits change from time to time in all countries. Thus no description can be fully up-to-date. The information presented here is chiefly based on the returns made by 140 countries to the Social Security Administration of the United States and published in 1985 as *Social Security Programs Throughout the World.*

Pensions. Three basic types of state pension schemes predominate. The first is a flat-rate pension with no income test. This may be available on a test of residence only or with the stipulation that the person has been employed for some specific period and has paid requisite contributions. This approach is found mainly in Scandinavia and the Commonwealth countries. The second is an income-tested pension. The third, and most common, type is a pension related in some way to earnings during working life. A further complication is that most countries with a flat-rate pension later developed a second tier of pension rights based on earnings during working life. In other words, the first and third principles are combined.

New Zealand pays a flat rate pension; financed from general taxation, to all who meet residence requirements at age 60. The rate for qualified married couples is twice the rate for single people. The rate of pension is quite a high percentage of average earnings. The Netherlands also provides all residents with a substantial pension but at age 65; it is financed from contributions and reduced if contributions due have not been paid in any year. The supplement for a wife of any age is less than half the rate paid to the husband. In Ireland the pension is less generous and only available to employed persons with minimum contributions paid. Australia combines the first and second approaches with a flat-rate pension from age 70 and an income-tested pension from age 60 for women or from age 65 for men.

Two-tiered schemes

Several countries in Scandinavia abandoned an early means-tested pension in favour of a flat-rate pension after World War II because of the unpopularity, complexity, and discouragement to savings of the means test. Later the level of the pension was regarded as inadequate for all except the low-paid, and an earnings-related tier was established on top. Thus flat-rate pensions are provided on tests of residence in Denmark, Finland, Norway, and Sweden. In three of these schemes a married couple receives substantially less than two single persons. In each case the schemes are supplemented by earnings-related pensions. Canada followed a pattern of development similar to those of the Scandinavian countries. The United Kingdom also gradually moved over to a two-tiered pension, but rights to both tiers depend on contributions paid with credits for absence from work for approved reasons; employers' schemes can be used to provide a specified minimum upper tier of pension.

The Scandinavian and Canadian two-tiered approaches have a number of advantages. First, non-means-tested basic pensions can be provided to persons without a contribution record—including the disabled, those who have not worked because of family responsibilities, and divorced or separated wives. There is a similar advantage in New Zealand's scheme. But, in addition, those who have had higher earnings and thus paid higher contributions receive higher pensions with the value of these guaranteed in terms of purchasing power by the government. This reduces the scope for employers' pension schemes in which the purchasing power of the pension finally paid depends on how far the yield of investments has managed to keep pace with inflation.

Capitalization

Contributory pension schemes, when they were first established, were run on much the same basis as private pension schemes. The level of contributions was calculated by an actuary, and a capital fund was built out of which the pensions could be paid. Even if there were no further contributions, the money was intended to be available to pay out pensions to contributors from the accumulated value of their contributions. This arrangement is known as capitalization or fully funded financing. The first scheme in Germany, enacted in 1889 and based on capitalization, covered most employed persons with earnings up to a specified level. The earnings-related contributions were equal for employees and employers, and there was a subsidy from the state to provide the low-paid with somewhat higher pensions than their contributions warranted. A breach with the principles of private insurance was made to allow workers close to retirement when the scheme went into effect to receive higher pensions than their contributions had earned them. This system of "blanketing in" older workers has frequently been used in other countries when new pension schemes have been established.

It was the experience of rapid inflation after World War II that led to a fundamental change in the financial basis of pensions. Instead of the contribution level's being sufficient to build up a large capital fund, it was calculated according to the expected cost of pensions due to be paid over the next few years. This pay-as-you-go method of financing statutory pension schemes, which became the normal arrangement, contrasts sharply with private pension schemes. The latter still have to accumulate capital funds because, unlike state schemes, they have no power to compel future generations to join them. Thus state pension schemes are essentially a "compact between generations." Those at work are compelled to pay to the pensioners of today in expectation, written into the law, that their pensions will be paid by the next generation of workers.

A second major development in pensions began in the late 1950s in response to rapid economic growth. It became recognized that if pensions were paid out on the basis of the money value of contributions paid in over a working life during which real earnings had been growing rapidly, pensions would amount to a low proportion of earnings at the time of retirement and a still lower proportion of what those at work would be earning 10 or 20 years later. Thus complex formulas were introduced to adjust pensions to the general level of earnings at the time of retirement. West Germany set the pattern in 1957 and was followed by several other European countries—for example, Austria, Switzerland, and the United Kingdom. An alternative approach (*e.g.,* in Italy and some eastern European countries) is to base the pension on the last few years of earnings. As this can be unfavourable to workers whose earnings decline in the later years of working life, some countries (*e.g.,* France) base pensions on the best few years of earnings. The Soviet Union offers an option of the last earning year or the best five consecutive years out of the last 10.

Minimum pension

The practice of giving low-paid workers higher pensions than were earned by their contributions and those of their employers, which was built into the original German scheme but later abandoned, has been copied in a number of later schemes (*e.g.,* those of the United States and the Soviet Union). An alternative or further way of helping low-paid workers is to provide a minimum pension, as in West Germany or the United States (though in 1981 the U.S. provision was removed for people not yet retired). This particularly helps women, whose average annual income, despite legal inroads against discrimination, remains well below that of men, and whose pension contributions are now likely to be interrupted by leave from work for family responsibilities. A much more common provision is an income-tested social pension as, for example, in Belgium or France.

Another development mainly of the period after World War II has been the automatic adjustment of pensions according to an index of prices or in some cases to the average level of earnings, or whichever is more favourable. Some countries have postponed adjustments or modified their formulas, particularly when prices were increasing faster than earnings.

Pension age

The age at which full pension can normally be drawn varies considerably between countries. In Europe the normal age for men can be as high as 67 and as low as 60 and for women as high as 66 and as low as 55. Some developing countries have still lower pension ages. To some extent pension age tends to reflect the expectation of life in the particular country. The pension age for women, however, is often lower than for men; one reason often cited for this is that husbands tend to be older than their wives, and so the disparity in pension ages permits simultaneous retirement. The arrangement, however, is disadvantageous for women who are retired compulsorily at the lower age after having had less time to accumulate a record of contributions. There is, therefore, a trend to equalize pension ages between the sexes. To do this by lowering the male pension age is expensive; it is for this reason that the European Communities has not made this binding on member states in its directive on equal rights to social security.

Some countries have long had provisions allowing the pension to be drawn a few years earlier than the stipulated age of retirement with an actuarially calculated reduction in the pension paid. Such provisions are suited to the more generous earnings-related schemes in which a reduced pension would not normally cause poverty. Ill health is a common reason for early retirement, though many choose this option in order to enjoy retirement while still in good health. There commonly are also provisions by which people who wish to postpone their retirement and continue to contribute can draw a larger pension. In some cases these arrangements have been introduced in the hope of encouraging later retirement, thus modifying the deterioration of the ratio between the employed population and the retired population, which necessitates higher levels of contribution by those still working as the proportion of the pensioned population increases.

Despite the logic of raising the normal pension age in line with an improved expectation of life, changes in schemes of industrial countries in the 1960s tended to lower the age. This trend has continued as the level of unemployment has increased, despite the financial burden this places on the schemes, particularly in the long run. The political objectives of reducing the number of persons recorded as unemployed and creating jobs for younger people have taken priority. Thus a wide variety of complex provisions have been written into pension schemes defining the circumstances in which full pensions can be drawn a few years earlier than otherwise stipulated. This may be allowed to those with many years of insurance (*e.g.,* 35), to those who have been unemployed for a substantial period (*e.g.,* a year), to those who are disabled, to those with arduous or unhealthy occupations, and to those whose jobs are being released for younger persons. In some countries the pension is income-tested below the normal age. Contrary to this trend for earlier pensions, the United States has raised the future pension age in two steps from 65 to 67 in response to the long-run financial prospects for the pension scheme.

Partial pension

A development pioneered by Norway in 1972, and since followed by more and more countries, was to allow persons aged 67 to 69 to reduce their working hours and receive at the same time a partial pension. This enabled older people to make a gradual transition between work and retirement. The change was made when the pension age was lowered to 67. Sweden followed in 1976 with a provision for those aged 60 to 64. Partial pensions have also been introduced in Spain and, on a much more restricted basis, in the United Kingdom.

In most schemes in industrialized societies there is a limit on what the pensioner can earn without leading to a reduction in pension. Some schemes specifically require the pensioner to leave his or her job on receipt of the pension. These provisions add to the wider pressures leading to the steady fall in the proportion of persons in full-time work above the normal pension age. But the main reason for this trend is the increasing generosity of pensions, both public and private.

Early pension schemes made extra provision for a dependent wife, and more did so between World Wars I and II. This can mean that women's contributions are "wasted" in the sense that the pensions they earn are less than they have a right to as a dependent wife. The greater frequency of divorce and cohabitation has meant that more women are wholly dependent on the pensions they earn in their own right. Moreover, some pension schemes make no provision for a dependent wife (*e.g.,* in West Germany, Austria, and Italy). The issue of women's rights to pensions is particularly important in the context of poverty, as women on average live longer than men.

A few countries allow housewives to contribute to pension schemes on a voluntary basis, but few women do so in practice. Others have adopted provisions for the dividing of pension credits between spouses. In the United Kingdom a man or woman can be credited with a full year's pension rights for each year up to a maximum of 20 during the whole of which he or she is caring for a dependent child or disabled relative. These rights are based on the individual's previous record of contributions.

In the early schemes widows over pension age were entitled to a proportion of the pension of their husbands. More and more schemes have been amended to make similar provisions for widows and widowers. Usually survivors can choose between their own personal rights and a proportion of the rights of their deceased spouse, but in Sweden widows can receive both earnings-related benefits on top of one flat-rate pension. This right is available, up to a maximum, to a widower as well as to a widow in the United Kingdom.

Disability and sickness benefits. In most countries provision for occupational injury is the oldest form of social security. The original German law of 1884 provided for workers to receive half pay for four weeks followed by two-thirds pay during temporary disability. In cases of permanent disability two-thirds of earnings from the year preceding the accident were paid out, with a proportion of this pension paid in cases of partial incapacity. Extra provision could be made for persons needing constant attention. The scheme was wholly financed by the employer, who paid insurance contributions, assessed on the degree of risk involved in the employee's occupation, to statutorily established associations. The associations then paid out any benefits.

The British law of 1897 made employers liable for compensation but did not require employers to insure against the risk. Compensation was half the basic pay for up to six months at which point the claim could be settled by a lump sum. These two very different precedents influenced developments in other countries. Continental European countries tended to follow the German model and the Commonwealth and the United States that of the United Kingdom. An act modeled on the British law of 1897 was passed in India in 1923, though the coverage was small. Moreover, the Belgians, Dutch, and French as well as the British preferred to introduce in their colonies laws imposing liability on employers rather than funded insurance schemes. By the end of World War II most colonies had such laws. These laws were often later augmented or replaced by insurance schemes. Some Scandinavian countries require the employer to insure but allow him to choose his own insurer.

Under an act of 1946 the United Kingdom introduced compulsory insurance through a state scheme with the same rate of premium for all employers. Benefits for incapacity were at a flat rate followed by a disablement pension based on degree of disability to which were added other allowances depending on the situation of the pensioner, including the loss of earnings and need for attendance.

Insurance is now compulsory in most industrialized countries, but the use of private insurance continues in a few countries (*e.g.,* Denmark and Finland) and the majority of U.S. states, while some countries give the employer the right to choose between a public or private insurer. Work-related injuries and an increasing number of occupational diseases lead in nearly all countries to higher benefits and more generous provision than are paid for sickness or injury not arising from work. For example, some countries in western and eastern Europe provide 100 percent of previous earnings as a temporary disability benefit. These benefits normally continue until recovery or the award of a long-term benefit. In most countries loss of earning capacity is a major consideration in the assessment of long-term benefits, and partial disability is more generously treated than in other social insurance programs. The long-term benefits in some countries can also amount to 100 percent of earnings. But the procedure of seeking lump-sum settlements from the courts still remains in some countries and some states of the United States, with all the associated costs, delays, and uncertainties and the difficulty of turning a lump sum into a secure income.

There are three special features of most occupational injury schemes that reflect their historical origins in the employer's liability. First, schemes are normally financed solely by employers' contributions. Second, the right to benefits operates from the very first day of employment. Third, a cash benefit is seen as compensation rather than income maintenance. For this reason dependents' benefits are not normally provided, but there is provision for surviving dependents. In addition, a compensatory benefit may sometimes be paid in addition to earnings or pensions. These features are found only in provisions for sickness or disability that are of occupational origin.

Both employers and employees normally contribute when the scheme is based on social insurance. Some minimum period or amount of contribution is generally required before there is entitlement to benefit. The amount of the benefit may depend on how long contributions have been paid, as for pensions, which is disadvantageous for those disabled early in working life. The main benefit is intended for income maintenance and thus cannot be drawn at the same time as other benefits or pensions with the same purpose. Finally, there is more likely to be provision for a dependent spouse.

Increasing uniformity of disability schemes

There has been a tendency in the period since World War II to bring occupational injury schemes into a closer relation to other social security schemes. Switzerland has always covered work-related and other accidents in the same scheme, established in 1911; New Zealand later adopted the same practice. The separate provisions for occupational injury and other disability raise difficult problems in specifying the distinction. Occupational injury normally has to "arise out of and in the course of employment." Some schemes allow travel to and from work to be considered within the "course of employment," while others do not. There are considerable difficulties in identifying whether certain disabilities (*e.g.,* deafness or arthritis) arise from work, and there are instances in which an injury is only partially attributable to the work situation. Part of the justification for combining the provisions for occupational injury and other disability is to eliminate such ambiguities. In addition there is the social argument that it is wrong to pay different benefits to different people, all of whom have the same degree of disability no matter how or when the respective conditions were caused. The Netherlands is the only country that has responded to this argument. From 1976, unified provision for disability has been made irrespective of cause. Costs of such a program can be substantial if all disability coverage is raised to a level approaching that of the previous, often considerably higher, occupational injury coverage.

In the case of sickness that is not associated with any occupational factors, most industrial countries pay a short-term benefit followed by a long-term pension after periods varying from about six months or less to a year or more. Some countries, such as Austria, Belgium, West Germany, and the United Kingdom, place responsibility for paying a benefit on the employer for the early weeks of sickness (though he may be reimbursed), after which the social security fund assumes payment. In some countries benefits may not be payable for an instance of illness lasting less than, for example, three days. In longer periods payment for the first three "waiting" days may be included in the benefit. Doctors' certificates may not be required for short spells of sickness. Benefits may be as high as 100 percent of earnings (*e.g.,* Austria and Belgium) in the early weeks of sickness or subject to a maximum, as in Norway, or for the full 52 weeks, as in Luxembourg. Or the rate may be 90 percent, subject to a maximum (Sweden and Denmark). Other countries normally pay only 50 or 60 percent (*e.g.,* France, Canada, and Greece). The benefit is flat-rate with extra for dependents in Ireland and the United Kingdom (after the eight weeks paid by the employer). The benefit is also paid on this basis in Australia and New Zealand, but it is means-tested. The United States is the exception among highly industrialized societies in that in most states there is no provision for short-term sickness apart from a special scheme for railway employees and social assistance, or welfare. In practice, provision is left for bargaining between employers' and employees' representatives.

Long-term invalidity pensions

Long-term invalidity pensions were included in the original 1889 German pension law (which was the first piece of legislation of this kind) for those who had lost two-thirds of their earning capacity. Many countries followed this model as part of (or as a later development of) their pension laws. In European countries invalidity pensions became payable after full short-term sickness benefit rights had been received. After World War II provisions were made in some countries for those who had considerable partial invalidity. Some countries require persons to have been insured for five or more years in order to be eligible for an invalidity pension, though generally there are means-tested pensions in industrialized countries for those who do not meet these requirements. In Australia and New Zealand those who meet residence requirements are eligible for income-tested pensions.

In countries with earnings-related pension schemes the invalidity pension is often calculated in the same way as the old-age pension. This means that the level depends on the number of years of contribution, though some countries have special concessions to enhance the pensions of those drawing them early in working life. Invalidity pensions may be supplemented by allowances for dependents and for constant attendance and other special needs.

Some countries make special provision for housewives who lack the contribution records that would qualify them for an invalidity pension. One such country is the United Kingdom, though the benefit is low and flat-rate. In Denmark a housewife can receive a substantial income-tested pension in her own right. Another group for which some countries have begun to make special provision is those who have been disabled from birth or before entering the labour market. These groups are provided for in the unified scheme of The Netherlands.

Most countries, however, are far from the position in The Netherlands, where all disabled persons are treated on a similar basis irrespective of the cause of disability. Those whose disability arises out of the work situation are generally most favourably treated; some countries provide full compensation for loss of earnings plus special allowances when required. Those who have paid contributions are generally treated better than those who have not, and often benefits depend on how long contributions have been paid.

Unemployment benefits. While sickness and disability are actuarial risks in that the incidence does not vary greatly from year to year, this is not the case with unemployment. It is partly for this reason that the duration for which unemployment benefits can be paid is limited in most countries or that benefits are reduced after a designated period. A further reason is to induce the unemployed to seek and accept work after benefits end or when they fall, although such work may be less well paid than the individual's earlier work and may provide an income lower than the unemployment benefit that has ceased.

Limits of coverage

The payment of contributions plays a critical role in policing eligibility for unemployment benefits; as a result the benefit is not payable to all persons who are involuntarily unemployed. The school dropout who has never had a job or only held one for a short period is normally ineligible for unemployment benefits. Women seeking to return to work after child-rearing are also ineligible, even though contributions were paid before leaving work. A prospective recipient must normally have held a job from which he has been released immediately before the benefit is claimed, and the individual must establish that he is available for work by registering at an employment office. Normally anyone who has voluntarily left a job or been discharged for misconduct is denied benefits or penalized.

In some countries the level of unemployment benefits is deliberately set at the same rate as the benefit for short-term sickness (*e.g.,* Canada, Denmark, and The Netherlands) so as to create no incentive for the beneficiary to try to establish eligibility for the higher benefit. In other countries the benefit for unemployment is at a lower level than the benefit for short-term sickness (*e.g.,* West Germany, Greece, and Hungary). Some countries that pay an earnings-related benefit for sickness pay a flat rate for unemployment (*e.g.,* Bulgaria and Italy). In Australia and New Zealand unemployment benefits, like sickness benefits, are subject to a test of income. The duration of the benefit varies from 13 weeks in Bulgaria to six months in

Hungary, Italy, and The Netherlands and a year in France, West Germany, Luxembourg, and the United Kingdom; in Belgium benefits can be continued indefinitely. In many, but not all, countries the unemployed can claim social assistance after their unemployment benefits cease. In several countries in northern Europe unemployment benefit schemes are operated by trade unions, though with substantial government subsidy.

Family, maternity, and parental allowances. While only a few countries had family allowances before World War II and several of the schemes covered employed persons only, with financing by the employer, there was a rapid extension of schemes in the 1940s and 1950s. The extension was in large part attributable to the influence of the Beveridge Report in the United Kingdom. Following the British example most of the new schemes in Europe, Canada, and Australasia included all resident children. A second influence was France, which introduced flat-rate family allowances for all children of employed persons in its African colonies—a system also introduced in some Latin-American countries (*e.g.,* Bolivia, Brazil, and Chile). The majority of schemes cover only employed persons, but a minority, particularly to be found in industrialized countries, pay allowances to all residents. The United States is exceptional among the latter countries in making no provision at all except in aid to dependent children paid on a means-tested basis.

Variations in dependent children allowances

Some systems of family allowances are intended to reduce poverty in large families or, particularly in eastern Europe, to increase the birthrate; the rate paid per child increases with the number of dependent children, reaching a maximum rate with the fifth or sixth child and subsequent children in, for example, Australia, Belgium, France, East Germany, Ireland, and Norway, or the eighth child, as in The Netherlands. In the Soviet Union family allowances start with the fourth child and reach the maximum rate at the 11th. Some systems seem to suggest a desired maximum family size insofar as the rate falls for subsequent children once there are three (*e.g.,* Bulgaria and Czechoslovakia) or two (*e.g.,* Greece and Hungary), or allowances may be payable for a maximum of six children (*e.g.,* Morocco). Finland recognizes that a mother is less likely to go to work if a child is under three years of age and thus pays a supplement. On the other hand, Austria pays higher rates for older children because they are more costly to maintain. Entitlement to family allowances ceases when a child reaches a particular age—often the age when compulsory education ceases, though allowances may be continued when a child continues in full-time education or is disabled. Allowances are payable only up to age five in the Soviet Union.

During the 1970s a number of countries decided to abolish income tax allowances for children and make a corresponding increase in the level of their family allowances. It was recognized that the largest beneficiaries from the tax allowances were high-income families with high marginal tax rates, and it was decided that this indirect benefit for children should be fairly shared among all families so as to increase the efficacy of family allowances in reducing poverty. Changes of this kind were made in Australia, Canada, Denmark, West Germany, Israel, New Zealand, and the United Kingdom. Denmark has gone one stage further and removed family allowances from the higher income groups by means of an income test. The United Kingdom has an additional income-tested allowance called the family income supplement which gives further help to low-income families.

It is the general practice for schemes that provide sickness benefits also to provide a maternity allowance starting before the birth of a child and extending for a number of weeks afterward. In some cases the rate of benefit is the same as for a sickness benefit, but in many cases the rate is higher—66 to 100 percent of previous earnings. Sweden has pioneered a parental allowance that can be drawn by the father as well as the mother to encourage fathers to take their turn in staying at home to look after the young child. In some cases a lump sum is also paid on the birth of a child to help pay for nursery goods and clothing.

Parental allowance for fathers

During the 1970s there was a concerted effort in eastern Europe to try to increase the birthrate by increasing the period for which a maternity benefit was paid and by giving credits in the social insurance scheme to mothers who stayed at home to look after a young child. Similar credits are provided in the United Kingdom for persons who stay at home to care for a child or a disabled relative, but the motive in this case is to increase the personal pension rights of those, particularly women, who have accepted family responsibilities.

Benefits for survivors and single parents below pension age. Provision is normally made for a widow below pension age left with a dependent child. Where pensions are earnings-related, the pension for a widow typically amounts to one-half to three-quarters of her husband's pension rights. In some countries the benefit is income-tested or time-limited (*e.g.,* three years in France). Other schemes vary considerably in the extent to which provision is made for widows. Some countries pay benefits providing widows are of a certain age when their husbands die. The age may vary between 40 (The Netherlands) and 55 (France). Some countries only pay the benefit providing the marriage has lasted for a specified period (six months in Greece; two years in France). Other countries pay the benefit to any widow who is disabled or to widows of any age for a short period or indefinitely. Widows' benefits normally cease on remarriage. A widower may be able to claim rights similar to those of a widow if he was dependent on his wife. Some countries extend widows' rights to divorced women. Increasingly, long-term provision for widows without dependent children is being questioned in societies where the trend has been for more and more married women to engage in paid work.

Provision for single parents, other than widows, is normally left to social assistance where such a scheme exists. Some countries have a special income-tested benefit. In Australia this is at the same level as an old-age pension for a person aged at least 65 but less than 70. In New Zealand it is less than half this rate for a single parent with one child. The problem with either of these arrangements is that a less skilled woman is unlikely to be able to improve her position by taking paid work because earnings lead to a reduction of the benefit or assistance. Denmark pays an extra family allowance higher than the normal rate per child for a single parent. Norway pays an extra allowance as if for one more child. The United Kingdom pays an extra allowance at just over half the level of child benefit.

Variations in provision between countries. All of the industrialized countries have social insurance schemes, and nearly all of them cover the main contingencies discussed above. The United States is exceptional in not providing family allowances, in not providing short-term sickness benefits in the vast majority of states, and in having no general scheme of national health insurance other than for the aged and the poor. The extent of provision in developing countries varies between those that still make provision by employers' liability and those that make provision by social insurance.

BENEFITS IN KIND

Systems of organizing health services or health insurance systems and of paying providers are changed occasionally but less frequently than the detailed provisions for cash benefits.

Introduction in Germany

The first national compulsory health insurance scheme, introduced in Germany under Bismarck's law in 1883, built upon precedents going back many years in the separate German states. Health insurance had developed mainly on an occupational basis and was a requirement for that occupation. The feudal obligation of the employer to his workers was given legislative substance in a society developing national markets, in which the employer without an obligation to pay to a sick fund might undercut the employer who had such an obligation. But the main reason for the scheme, as mentioned earlier, was to try to contain socialist tendencies.

The administration of compulsory health insurance was left in the hands of numerous local sick funds operating under legislative regulations. They became jointly controlled by employers and employees and made their own contracts with particular doctors and hospitals for the provision of services. All lower earning workers were eventually required to be members of a fund. Doctors were paid in a variety of different ways, including salary and capitation. In the course of time there were major protests from doctors excluded from contracts with the funds, and the profession demanded the right for any doctor to undertake health insurance work. The substitution of payment per case and later fee-for-service payment, which the German medical profession fought for and eventually won, was a means of establishing open competition between all doctors wishing to take part in the scheme.

Health insurance was enacted in Austria in 1888 and Hungary in 1891 on a similar basis. A bill to introduce such a scheme in Switzerland was, however, decisively rejected by a plebiscite in 1900. The British Radical politician David Lloyd George visited Germany in 1908 to see the scheme firsthand and subsequently introduced compulsory health insurance for persons with earnings below an upper limit in Britain by a law of 1911. However, the scheme provided only for the services of the general practitioner and the drugs he prescribed; hospital benefits were excluded except for some provision for tuberculosis, partly so as not to disturb the charitable hospitals that provided free care to those in need. Moreover, as a result of pressure from the medical profession, the benefit in kind was administered by statutory committees for each area, which enabled every general practitioner to participate who wished to do so, rather than by the large number of friendly societies that had previously provided medical benefits under voluntary insurance and had made their own contracts with particular doctors. Payment was on a capitation basis, as in the previous friendly society schemes.

The introduction of compulsory health insurance was considered in Sweden in 1884 and Denmark in 1885, but both countries decided instead to encourage voluntary insurance by government subsidy. Whereas Norway introduced compulsory health insurance in 1911, Denmark did not follow until 1933, and Sweden not until 1955. In these countries public hospitals were well developed and heavily subsidized. A compulsory health insurance law was passed in France in 1920 but, as mentioned earlier, did not come into effect until 1930 owing to disagreement about the local control of the scheme and a dispute with the doctors about the method of payment. Fee-for-service payment was finally substituted for the capitation system originally proposed. Moreover, as the doctors refused to accept the intrusion of any third party between them and their patients, the scheme operated on a reimbursement basis; the patient paid the fee and claimed a refund for the major part of it from the relevant insurance fund. This reimbursement system was adopted later in Sweden, Finland, and Australia (under subsidized voluntary insurance).

Reimbursement system

When health insurance was established in Russia in 1912, insurance doctors were paid salaries and practiced from government-owned premises. This was the pattern adopted in Chile in 1924. Payment of doctors on a part-time salaried basis for work performed on premises owned by the sick fund became the general pattern in Latin-American countries and in Spain, Portugal, and Greece. In most of Europe and Australasia, however, existing hospitals began accepting insured patients when hospital care became covered by insurance; special hospitals for insured persons were built in Spain, some parts of Italy, and a number of countries in Latin America.

The next major step in the evolution of medical benefits was for countries to make them available to the entire resident population, financed wholly by taxation or financed in part by social insurance contributions. This step was taken first by Hungary in 1920 and then by the Soviet Union in 1937. New Zealand implemented similar coverage in a series of steps—free inpatient treatment for the whole population in 1939 and outpatient treatment, free pharmaceuticals, and part payment of general practitioners' bills in 1941, with further steps later on. The United Kingdom established its National Health Service in 1946. Norway made services available to all residents in 1956, Sweden in 1962, Denmark in 1973, Portugal in 1979, and Italy in 1980. By the 1980s, more than 20 countries had adopted this system. This does not necessarily mean that all services are free at the time of use. Nor does it necessarily follow that all services are government-owned. Of the eastern European countries, some (Bulgaria, Czechoslovakia, Hungary, and Romania, as well as the Soviet Union) have adopted this approach; others (such as East Germany and Poland) have not. About half of all countries retain an element of financing by social security contributions after adopting this approach.

Canada was relatively late in establishing compulsory health insurance. The first province to do so was Saskatchewan in 1962. By 1971 all provinces had done so, spurred on originally by a 50 percent grant from federal funds; the provincial schemes became available to all residents. The not-for-profit general hospitals were given budgets by the provinces to provide this care. Australia was also late in changing from subsidized voluntary insurance to compulsory health insurance. The United States and Switzerland are left as the only highly industrialized countries without general compulsory health insurance or a health service available to all residents. There have been many attempts, against strong opposition from the American Medical Association, to introduce compulsory insurance in the United States. However, a limited scheme of compulsory health insurance for the aged (Medicare) was finally introduced in 1966 along with a system of means-tested medical care operated by each state for the indigent and medically indigent (Medicaid).

Developments in the United States

Some countries in Europe have succeeded in securing high coverage of the population under compulsory health insurance without switching to a service available to all residents. Schemes cover the employed, the self-employed, and all social security beneficiaries and their spouses and dependent children: this can amount to 99 percent of the population. Other European countries (West Germany and The Netherlands), nearly all of which have some system of private insurance, exclude the higher income groups from statutory health insurance. Alternatively, some benefits (*e.g.,* hospital care) are available to the whole population, while higher income groups must make their own arrangements for certain other benefits (Ireland).

Apart from Cuba, which has a national health service, only three countries in Latin America (Argentina, Brazil, and Costa Rica) have managed to cover 80 percent or more of the population by health insurance. Moreover, coverage may not necessarily mean that services are equally available. Coverage extends to about half the population in Mexico, Panama, and Uruguay, and more than a quarter in Bolivia and Venezuela. In the remaining countries coverage is 10 percent or less. By no means do all of these countries extend the same rights to the spouse and children of the insured person. Several provide only maternity care and pediatric care for dependents. Coverage is more easily provided for the employees of larger establishments, which tend to be concentrated in urban areas. Even in urban areas those excluded tend to be the self-employed, domestic servants, and itinerant workers. The obstacles to expanding rural coverage include the much lower levels of earnings, the geographic dispersion, the less formal employment conditions, and more extensive self-employment and seasonal employment. Most important of all, some schemes have become too costly to extend on the same basis with tax subsidy to cover the whole population. Thus the remaining population must depend on poorly financed and staffed services provided by ministries of health. Health insurance is, therefore, increasingly criticized for exacerbating inequality in health care by outbidding government health services for trained manpower and for creating a heavy emphasis on sophisticated and expensive curative services in urban areas while the main health need is for preventive services to cut the incidence of infectious diseases in both urban and rural areas.

Obstacles to rural coverage

Japan has managed to avoid the worst of these effects and to achieve high coverage. India, conscious of the damage that health insurance could do to government services, has developed health insurance slowly as resources have become available for doing so in particular states. South Korea has introduced health insurance for the urban employed population and also has provided rights to those with low incomes in urban areas; the problem of covering the remaining half of the population in rural areas remains to be solved.

Many developing countries, particularly those that were previously British colonies, have made health services available to the whole population, providing free or nearly free services. This is the pattern, for example, in the West Indies, Kenya, Zimbabwe, India, Sri Lanka, Malaysia, and many Arab states in the Middle East. In most cases services were originally developed for the expatriate colonialists and extended in the course of time to local residents. The services tend for this reason to be heavily concentrated in urban areas, with little or no coverage of the rural population. With their limited resources, these countries are striving, as part of the World Health Organization's program Health for All, to extend rural coverage with primary health care to all areas by the year 2000.

Methods of provision

Among the various national health schemes, benefits are provided in three ways. First is the direct service approach in which the government or insurance fund owns the facilities (hospitals and clinics), pays for supplies, and remunerates the staff on a full- or part-time basis. This is the approach used in the United Kingdom for hospitals and community services and in Scandinavia, where local authorities provide hospitals and clinics, though there may also be a parallel system of doctors working from their own offices. It is also generally used in eastern Europe, in Greece, Spain, and Portugal, in most countries in Latin America, and in most other developing countries. The hospital system in Canada is exceptional; the scheme determines budgets for general hospitals that remain in the hands of not-for-profit agencies.

The second method is the indirect contract with providers. The providers may be private entities (hospitals or practitioners) or public hospitals, but the health insurance scheme makes a contract with the provider and pays each provider for services used according to rates established in a negotiated contract. This is the system used for all services in such countries as Belgium, West Germany, Luxembourg, and The Netherlands.

The third method is reimbursement, in which the patient pays the bill and submits the receipted bill for reimbursement. The provider may be public or private. This approach is widely used in France, some northern European countries for the parallel system using practitioners in the private sector, and to some extent in Australia and Sweden. The patient may be left to pay part of the bill, as, for example, in France. A fee schedule may be established for rates of reimbursement, but unless strong measures are taken to prevent it, some practitioners may charge more than the established fee.

In practice many countries use a combination of these systems. Thus, for example, the National Health Service in the United Kingdom, with its direct service provision of hospitals and community services, uses indirect contracts for general practitioners, community pharmacists, opticians, and most dentists. Moreover, where private hospitals are used they are paid under contract, as is also the case in Greece, Italy, and Portugal. In a number of countries in Latin America health insurers use the direct service approach in urban areas but service dispersed populations in rural areas by using indirect contracts.

Health insurance schemes vary in the method by which providers are paid, and this can have a substantial impact on costs. Where doctors and dentists are paid on a fee-for-service basis this provides incentives for the provision of further services—even in France where the patient has to pay a proportion of the cost. In the Common Market countries about twice as many prescriptions are issued to patients when the doctor is paid on a fee-for-service basis as when he is paid on a capitation basis. More surgery is performed where doctors receive fees rather than salaries. Moreover, the patient normally has direct access to specialists and can visit several different doctors in the course of one illness; this also adds to costs. When hospitals are paid on the basis of an itemized bill, more items are often provided. Where hospitals are paid per day of care, there are incentives for the hospital to keep patients for longer than necessary. For this reason, some countries in Europe (Belgium, France, and The Netherlands) have required hospitals paid on this basis to adhere to a predetermined budget. Where hospitals are given a budget from the local or central government, costs are kept under control. Financial incentives for the provision of further services are avoided where doctors are paid on a salary or capitation basis (The Netherlands and the United Kingdom). But this can lead to delays in receiving treatment either as an inpatient or as an outpatient. A provision permitting access to specialists, normally only on the basis of referral by a general practitioner, can be enforced where the patient normally has access to only one practitioner; this helps to limit costs. The system of paying doctors part-time salaries, leaving the doctor free to undertake practice, as in Greece, Portugal, Spain, and most countries in Latin America, can lead to what patients see as poor quality in services—a lack of courtesy and limitation of time devoted to the consultation. For this reason many countries are beginning to offer full-time salaries without rights for the doctor to undertake private practice.

The right to free medical treatment was included in the original German scheme for industrial injury, and provision for rehabilitation was added in 1925. In the course of time more and more emphasis came to be placed on efforts to restore working capacity, and specialized institutions were created for this purpose. Many countries have copied the German example and developed highly specialized institutions owned by sick funds or under the control of the agency responsible for national health insurance for both physical and vocational rehabilitation.

ADMINISTRATION AND FINANCE

Administration of social security. Countries vary considerably in the extent to which social security is centralized and unified. A high degree of centralization obtains in the Commonwealth countries and Scandinavia (except for health care and social assistance, which are decentralized to lower levels of government). A centralized scheme may be run by a ministry or by a semiautonomous agency. In other countries schemes are more often run by separate occupational funds or by funds providing for different risks, as tends to be the pattern in continental Europe and Latin America. The control may rest with boards composed equally of employers and employees. Or it may be tripartite, with the government participating as the third party. In the United States social security as defined here is divided between federal and state responsibility. There have been attempts in some countries to secure greater unification, but such efforts have often run into strong resistance from particular occupational groups with better benefits or lower contributions attributable to lower risks.

Complexity of regulations

Social security regulations have become extremely complex and difficult to understand. Where there are separate funds, each may have a national office, with no branch offices to which the public has access. Disputes often arise over which fund is responsible for paying benefits to particular claimants. It is, therefore, not necessarily the case that all claimants obtain what they are entitled to receive, and substantial delays can occur while entitlements are sorted out. Problems of this kind are not, however, unique to the public sector. Some private insurance companies are resistant to paying out claims. Unified social security systems with local offices are more accessible to the public, but the offices are not always adequately staffed to give the public prompt and efficient service.

Social assistance regulations are inevitably even more complex to operate than other parts of the social security system. Moreover, they frequently contain a considerable element of discretion. Where schemes are administered by social workers there can be what beneficiaries see as potential coercion; failure to follow the social worker's advice may be thought to lead to the reduction or removal of benefits. Some have argued that all social assistance regulations should be published so that claimants can know their rights and thus be in a position to appeal against decisions to refuse benefits or extra allowances. Adoption of this approach has led in some cases to regulations that are too complex for the staff to operate efficiently, or in others to regulations that have been streamlined at the expense of former provisions for discretion. Particularly contentious is the question of cohabitation. If an unemployed married woman living with her wage-earning husband is not entitled to social assistance, it would seem at first sight only fair that an unemployed woman cohabiting with an employed man should be treated in the same way. But cohabitation may not be accompanied by maintenance and is anyway extremely hard to define. The borderline between a lodger and a cohabitant is by no means clear cut in all cases nor readily established by any outside agency. Attempts to do so can involve considerable invasion of personal privacy.

The issue of cohabitation

Financing of social security. In most countries the major part of the cost of social security is paid for by proportional contributions of earnings from employers and employees. The contributions may be divided equally between employers and employees, except for the whole cost of the occupational injuries scheme, which falls to the employer. Alternatively the employer may pay about twice the amount falling to the employee. There is usually a "ceiling," or level of earnings, beyond which the contribution becomes flat-rate at the level of contribution due on this maximum of earnings, though this is not the case in either Sweden or Switzerland. The maximum varies from around 50 percent above average earnings (*e.g.,* France, Ireland, and Italy) to twice average earnings (*e.g.,* West Germany, the United Kingdom, and the United States) or higher (Norway). The reason for this may be to prevent insurance contributions from overlapping with high marginal rates of income tax or to leave the replacement of high earnings to the private sector. Some countries also exempt very low earners from contributions or make the employer pay them instead of the employee.

Usually some portion of costs is left to be met from taxation. At the very least the government will stand by to meet any deficit between benefits and contribution income. During the 1970s there was a trend in most countries in western Europe for costs to be shifted away from employers and onto taxes (*e.g.,* Denmark, Ireland, Italy, The Netherlands, Portugal, and the United Kingdom) or to employees (Austria, France, and West Germany). One reason for the trend toward tax financing was the growth of unemployment financed by social assistance payments.

The role of taxes

Countries where no costs at all fall on taxes include the small schemes in Burundi and Ethiopia and the wider schemes in Malaysia, the Philippines, and Singapore. At the other extreme, countries where contributions play a very small role and by far the bulk of costs is covered by taxation are Australia, Czechoslovakia, Denmark, New Zealand, and the Soviet Union. In the United Kingdom, where the national health service is primarily financed from taxes and social assistance plays a major role, roughly half of the costs are borne by taxes and half by contributions. Several eastern European countries have no employee contributions; their schemes are mainly financed by employers.

Contributory versus tax-financed schemes

The relative merits of financing by contributions or taxes have long been debated. In favour of contributions it is argued that making beneficiaries pay prevents irresponsible increases in benefits and, where there are separate funds, encourages participation by both employees and employers. The payment of contributions also helps to ensure that commitments are honoured. Contributions are administratively easy to collect since the employee has an interest in securing compliance by the employer. The benefits to the employee of paying are clearly identified, while the cost falling on employers may create some incentive to prevent certain occupational risks from arising. Finally, only by earmarking contributions can earnings-related benefits be justified.

The critics of contributions argue that where they are flat-rate or where earnings-related contributions are only payable up to a low ceiling of income they are regressive and constitute a heavy burden on the poor; progressive taxes on income would be preferable, as they vary according to ability to pay and are also levied on investment income. It is also argued that tax-financing enables governments to judge priorities among all fields of public expenditure, and, where it leads to administration by government, this secures closer coordination between social security and other services. In addition, high contributions lead to the growth of the black, or underground, economy. This is a major problem in France and Italy with their high employers' contributions and leads to a widespread lack of social insurance coverage.

An argument that became more strongly pressed when levels of unemployment rose in the 1970s was that high employers' contributions made products uncompetitive in world markets, particularly in the case of labour-intensive industries, compared with products from Third World countries where social security is less developed. This was said to sharpen the recession and aggravate unemployment in highly industrialized countries. While it is true that employers might gain a short-term advantage if contributions were lowered, it is much less certain that this gain would be sustained in the long run. What was gained in lower contributions might sooner or later have to be conceded in higher wages and salaries or in other wage costs. If the argument were valid, such countries as Australia, Denmark, or New Zealand, which make little use of employers' contributions, would be seen to be cornering a heavy share of world trade. The fact that this has not happened reinforces the argument that it is total labour costs, of which social security contributions are only a part, that affect competitiveness.

It has been claimed that high employers' contributions particularly damage labour-intensive firms and encourage the replacement of labour with capital. In examining this assertion it is relevant first to remember that firms making capital goods also have to pay the same high employers' contributions and that capital-intensive firms pay them indirectly on raw materials, facilities and equipment, and energy. Second, high employers' contributions may well cause cash wages to be lower than would otherwise be the case so that total labour costs are not, in fact, increased by employers' contributions. Third, insofar as high employers' contributions encourage all firms to use more capital-intensive methods of production, this applies to labour-intensive firms as well. This encouragement of investment may lead to production at lower cost and thus a more competitive position in world markets in the longer run.

While there is a lack of convincing evidence that employers' contributions are bad for employment, a low ceiling on contributions may itself damage employment. It may discourage offers of part-time work and lead employers to prefer offering overtime to taking on additional workers. This was the view of international experts appointed by the International Labour Office, who recommended in their report of 1984 that contribution ceilings be abolished.

The substitution of taxes for contributions may not relieve poorer workers if the extra taxes come from goods such as tobacco that are consumed more heavily by those with low incomes than those with high incomes in industrialized countries. There is no guarantee that governments would raise the extra revenue from progressive taxes; they may, for example, lower the threshold at which income tax is paid.

The argument in favour of contributions

The strongest case for contributions is that they justify earnings-related benefits. The strongest case for taxes is that they are used in many countries to make benefits available to all residents—whether the benefits be health care, family allowances, or minimum flat-rate pensions. Solutions to the problem of persons not currently covered or inadequately covered by social insurance programs normally require a greater element of tax financing. This has been the trend in many countries.

The rising cost of social security. The cost of social security rose substantially in the period after World War II both in real terms and as a proportion of rising gross domestic product. While social security spending amounted to less than 10 percent of the gross national product in nearly all countries in 1950, it had risen to 20 to 30 percent or more in many European countries by 1980. Among the reasons were the extension of the coverage of social security, the widening of the risks covered, the indexing of benefits, and the greater generosity of benefits, which moved up to or near 100 percent replacement of earnings for certain contingencies in some countries. But also of major importance was the maturing of pension schemes. Many of them were recast in the 1940s and 1950s; it was not until the 1980s that people had had the opportunity to contribute on the new basis for all or most of their working lives and thus draw pensions approaching or reaching the maximum for which these schemes provided. Three further factors were the increasing proportion of aged persons in the population, the decline in pension ages, and the lower proportion of working population.

The costs of health care also rose sharply after World War II. Several reasons contributed to this trend. First, the higher proportion of elderly in the population influenced health care costs as well as the costs of cash benefits. Persons over pension age require two to three times more health care than persons of working age, and the difference is still greater for those over 75, the fastest growing age group. A second factor was the decline in working hours, which meant that more persons (*e.g.*, nurses) were needed in order to staff 24-hour services. A third factor was the continuous development of medical technology, such as new equipment and labour-intensive procedures. Instead of replacing labour, as in industry, innovations in health care normally required more labour for their operation. A further reason was the removal of supply restraints with the provision of more doctors and dentists, a major growth of medical auxiliaries, and the construction of new hospitals, which were more expensive to run. A fifth reason was the financial incentives to supply more services, which underlay many of the systems of paying providers under health insurance.

Consequences of new technology

The final and critical factor that destabilized the finances of social security schemes was the rapid growth of unemployment beginning in the 1970s. In those countries that included unemployment benefits in their social insurance schemes, this phenomenon created both unpredicted higher costs for benefit payments and a loss of revenue from those who were unemployed. The burdens on social assistance programs were also substantial in some countries, coming at a time when unemployed persons were no longer in a position to contribute to tax revenue.

The rapid growth of social security expenditure attracted little attention during the period of rapid economic growth up to 1973. It began to cause concern after the steep rise in oil prices checked economic growth in oil-importing countries. The revenue that financed social security ceased to be buoyant at the same time as new major demands were made on the system. From the late 1970s there was talk of a crisis in social security financing.

By 1980 social security expenditure amounted to 32 percent of the gross national product in Sweden, between 25 and 30 percent in Belgium, Denmark, France, and The Netherlands, and between 20 and 25 percent in Austria, West Germany, Ireland, Luxembourg, and Norway. These figures were much higher than for Australia (12 percent), Canada (15 percent), Japan (11 percent), New Zealand (14 percent), the United States (13 percent), or the United Kingdom (18 percent). The cost was much lower in developing countries. High costs are associated with high levels of social security benefits and also with costly systems of providing health care. Some countries, such as Sweden, have allowed health care costs to continue to rise because of the capacity of this service sector of the economy to provide further jobs and thus avoid high rates of unemployment.

Attempts to contain costs

The aim in many industrialized market countries came to be the containment of the costs of social security. This requires that program costs not grow faster than the yield of contributions. Various devices were introduced to help secure this result. Systems of indexing benefits and pensions to prices or earnings were revised downward, or adjustments were made less frequently. Pensioners were made to pay contributions toward health-care benefits. In France tax income was brought in to supplement the yield of contributions. In the United Kingdom the earnings-related additions to short-term benefits were abolished.

A series of measures was introduced to limit the cost of health care. Charges and copayments were increased or new charges were introduced. Payment for drugs was introduced in West Germany (1977), Italy (1975), and Portugal (1982). Portugal and Luxembourg joined France and Belgium in charging for consultations with doctors. Charges for hospital care were introduced or extended in Belgium, West Germany, Portugal, and France. By 1984 there was no country in western Europe that provided free care to all its insured population.

Payment systems under health insurance were revised to reduce incentives for overservicing. The aim in West Germany was to pay the doctor more for the consultation and less for medical procedures. Payments for diagnostic tests were sharply reduced in Belgium. As part of the introduction of a national health service in Italy, payment to all general practitioners was changed from fee-for-service to capitation, and the bulk of specialists began to receive full-time or part-time salaries. Budgets for each hospital were introduced in Belgium, France, and The Netherlands, in part to discourage unnecessary retention of patients paying per day of care. Countries in which hospitals were already paid on a budget basis reduced the budgets. In the United States hospitals began to be paid under Medicare and Medicaid according to a schedule of costs for various groups of diagnoses.

Countries maintained strong controls over new hospital construction or expansions, and incentives were created in a number of countries to transfer beds from general use to the care of the long-term sick. Several countries took measures to develop alternatives to hospital care, such as outpatient surgery, outpatient hospitals, nursing homes, residential homes, and home care by domiciliary teams. The United Kingdom closed some 400 hospitals over a period of 10 years. Restrictions on the installation of major new medical equipment went into effect in Belgium and France. By 1955, 10 of the 12 countries of the European Economic Community had instituted quotas for medical schools. In Denmark, France, Ireland, Portugal, and Spain the number of medical students was cut substantially.

Most countries in western Europe introduced restrictions as to what medications a doctor could prescribe under the health service or health insurance system. Most of these countries exercised tight control over pharmaceutical prices and pharmacists' margins. New measures were introduced in the effort to control overprescribing.

Variations in social security spending

Social security spending tends to vary between countries in direct proportion to their respective standards of living; in other words, the more affluent a country is, the more it is likely to spend on social security. Spending also tends to vary according to the proportion of elderly people in the population. Third, it varies according to the year in which the first legislation was adopted: countries with older social security programs tend to spend more. There are, of course, exceptions to this pattern. For example, the United States and Japan are low spenders both for their standard of living and for their proportion of elderly, and New Zealand is a low spender for a country that introduced pensions as early as the end of the 19th century.

This type of analysis has been criticized, however, for ignoring private arrangements, particularly employers' provisions established as part of collective bargaining. Thus, for example, the large role of fringe benefits in Japan helps to explain the relative lack of development of statutory social security. Similarly, the large role of occupational pensions and health insurance negotiated between employers and employees helps to explain the underdevelopment of statutory social security in the United States. Hence it is argued that private and public social security must both be taken into account in any comparison of national programs. In federal countries such as Australia, Canada, Switzerland, and the United States there were constitutional obstacles to adopting social security that led to the private sector's playing a larger role.

Political orientation also plays a role in explaining the extent to which social security has been developed in the public sector. After some initial opposition, political parties drawing substantial support from the working classes and the trade unions have promoted the expansion of social security. This includes European Catholic Workers' movements. Extensions of social security may be introduced by coalition governments with a conservative majority as the price needed to keep the coalition together. The high spending in Scandinavia can be explained by the strong influence of social democratic parties in the period following World War II. Trade unions have had less influence in this direction in Australia and New Zealand. The absence of a working-class party in the United States is part of the explanation of the relative underdevelopment of its social security program.

Implications

Some of the trends leading to increased costs are bound to continue. While the number of aged persons in most highly industrialized societies is likely to stabilize during the later years of the 20th century and the early years of the 21st century, the proportion within it of those over 75 will continue to increase substantially. This has major implications for further increases in the cost of health care. Moreover, pension schemes are still maturing, and there are pressures for further improvements of benefits, particularly to provide sex equality, lower pension ages, and better assistance for persons, particularly women, inadequately provided for previously. On top of all this, costly developments in medical technology continue. If the trend to shorter working hours continues this will also have a further major impact on the cost of health services.

Looking further ahead, the proportion of aged in the population is expected to start to increase substantially in the second and third decades of the 21st century as the increase in births after World War II becomes reflected in an increase in pensioners. It is this prospect that has led the United States to plan for increases in pension ages and the United Kingdom to decide to scale down its second tier earnings-related pension scheme.

The level of contributions and taxes needed to sustain present plans for social security cannot be predicted. While the continuing trend toward a higher number of aged in the population can be safely predicted, the birthrate is much harder to forecast. Of vital importance is the level of unemployment because of its impact on both sides of the balance sheet; reduced unemployment would add to contributions and tax income as well as lower the cost of benefits. Nevertheless, the prospect of a substantial increase in pensioners in the 21st century has led to fears in some quarters that the "compact between generations" may not perpetually be honoured. Hence it is argued that the pay-as-you-go method of pension contribution should be replaced by the capitalization method used in early pension schemes and in the private sector. Alternatively it is argued that the privatization of social security pensions would lead to higher savings and investment out of which future pensions could be paid. The disadvantages of either of these approaches are that there would need to be an immediate increase in contributions to provide the planned level of pensions. This could lead to pressure for higher cash earnings. Moreover, the level of pensions would no longer be indexed but would depend on the yield of investments.

CRITICISMS

It has been argued that the high cost of social security is in part responsible for the low levels of economic growth in industrialized societies since 1973. The argument takes three forms. First, it is said that high levels of unemployment benefits reduce the incentives to take paid work. Second, resistance to the payment of taxes and contributions leads to wage demands, inflation, and government deficits. Third, it is argued that because people have rights to social security benefits they are less likely to save; this lowers investment and thus economic growth. For all these reasons social security is said to have contributed to or even to have been responsible for not only low growth but also for high levels of unemployment.

In response to these criticisms it has been pointed out that empirical investigations lend little support to the contention that people prefer benefits to work, though the availability of benefits may make them less willing to take low-paying jobs. Second, it is argued that tax resistance would apply whether pensions were provided in the private or in the public sector. Indeed, if pensions were provided in the private sector they would have to be capitalized, which would require higher contributions and therefore lower cash earnings, leading to still greater pressures for higher pay. Third, the evidence that social security reduces savings is by no means conclusive; indeed, in many countries there has been a boom in the variety of different forms of saving following the establishment of pay-as-you-go systems of financing social security. Moreover, investment is limited by the availability of profitable investment opportunities rather than by any shortage of savings. And if low savings does limit investment, governments can generate a budget surplus out of which investment can be financed.

Effects on savings

Some critics argue that social insurance benefits should be replaced by a negative income tax. As countries get richer, it is argued, an increasing proportion of the population is in a position to take out private insurance against the risks for which social security provides. If social security were concentrated exclusively on the lower income groups, provision could be more generous and the burden of public provision could be reduced. The administrative and other problems of using annual tax returns as the basis of making cash payments to individuals whose circumstances are constantly changing as they go in and out of employment, marriage, and cohabitation are considerable. But in the context of saving, it was because income-tested pensions were thought to be damaging to thrift that many social insurance programs were established in the first place. Low earners are unlikely to save if the yield of their savings leads to a dollar-for-dollar reduction in pension. Moreover, many countries in Europe already have income-related housing allowance schemes that serve much the same function even though they are separate from income taxes. Where this is the case, there is no room for a further negative income tax or other income-tested scheme without imposing extremely high tax rates on increased earnings. Most important of all, it is by no means clear that the economically securer members of the population would be willing to accept anything like the existing level of contributions and taxes if they stood to gain nothing from social security. As a result, provision for the poor might be no better or, more probably, it might even be worse than it is as part of a scheme to which all contribute and on which all are in a position to make claims. Historically, services for the poor have always tended to be poor services.

Empirical studies have shown some small association between higher levels of social security spending and lower rates of economic growth. But it is not clear that one necessarily causes the other. Many other factors are at work. Countries that had a high proportion of the population in agriculture were in a favourable position to achieve high growth in the postwar period as their agricultural populations declined. Moreover, countries that have had relatively low rates of economic growth, such as the United Kingdom and the United States, are relatively low spenders on social security, while countries that have had high rates of economic growth, such as Belgium, Denmark, and The Netherlands, are high social security spenders.

Low target efficiency

Critics of social security are not confined to those concerned about its effects on the economy or about personal freedom regarding the extent to which and methods by which individuals provide for their security. There are also those concerned about the "target inefficiency" of social security, or its limited redistribution to the poor. This attack is generally directed at earnings-related benefits. Proposals have been made to use the yield of social security contributions, supplemented by taxes, to provide everyone with a minimum income on which they could live at a modest level supplemented by earnings if they wished to take paid work. Social dividend schemes of this kind are seen not only as a way of redistributing income but also of reducing unemployment. A major problem with such schemes is the high level of contributions and taxes needed to finance the minimum income if it is to go to those with jobs as well as to those without them and be sufficient to live on, if only at a minimum level.

It is true, however, that high spending on social security has failed to solve the problem of relative poverty in industrialized societies. Yet abolishing poverty was never the original intention, or at least it was not in many countries.

Social security was seen as a system of maintaining income by redistribution from the well to the sick, from the young to the old, and from those with jobs to those without them. This involves substantial redistribution but not necessarily redistribution from rich to poor. For instance, while there is more illness and unemployment among low earners, higher earners tend to live longer and thus draw pensions for a longer period.

A study financed by the European Economic Community showed the extent to which poverty remained in the Common Market countries around the year 1975. Poverty was defined as half the average level of living for each country. The study showed that the greatest success in combating poverty was achieved in The Netherlands and, second, in the United Kingdom, though Belgium and West Germany came close to the United Kingdom. Both of the first two countries have primarily flat-rate benefits. Poverty, however, was at its greatest in Ireland and Italy—both countries with substantial agricultural populations—though Ireland also relies on flat-rate benefits.

A variety of reasons explain why poverty persists in industrialized countries despite their elaborate provisions for social security; the precise reasons and their relative importance differ from one country to another. One reason may be that the arbitrarily selected poverty line is just above the level of living provided by social assistance. A second may be the failure of all those entitled to benefits to claim them. A third may be that certain categories of people in some countries are not entitled to claim social assistance (*e.g.*, the long-term unemployed). A fourth reason may be that earnings-related benefits do not secure a sufficient income for low earners to rise above the poverty line or that their record of contributions is insufficient to achieve this result. But considerable poverty may also persist among families headed by a full-time worker. This is more likely to occur in one-parent families when the earner is a female with limited skills and low earnings. But it may also occur in two-parent families with one earner and several children, where family allowances fall below the cost of maintaining a child at the minimum level or the cost of rent is considerable.

In the case of developing countries social security is criticized for reinforcing the dichotomies between urban and rural populations in general and between employed and unemployed persons in the urban sector. Social insurance contributions, which are in effect taxes specifically tied to providing benefits to the members of the schemes, cannot be used by governments for the benefit of the community as a whole. This limits the capacity of governments to raise tax revenues for broader purposes. Moreover, different health benefits as well as cash benefits may be provided for different occupational groups, thus perpetuating and accentuating inequalities. Those who defend social security argue in reply that requiring part of earnings to be put into an insurance fund is robbing no one outside the scheme, since only by providing the benefits could the particular taxes be justified and compliance with paying them be secured. Criticisms regarding inequality should be directed at the pattern of original earnings, not at social security, which mobilizes part of them for good purposes. (B.A.-S.)

BIBLIOGRAPHY

Social welfare: Key texts from the extensive British literature include the *Report of the Committee on Local Authority and Allied Personal Social Services* (1968), known as the Seebohm Report; ERIC SAINSBURY, *The Personal Social Services* (1977); *Social Service Teams: The Practitioner's View* (1978); and EILEEN YOUNGHUSBAND, *Social Work in Britain, 1950–1975: A Follow-Up Study*, 2 vol. (1978). The recent historical development of these services is analyzed in JOAN COOPER, *The Creation of the British Personal Social Services, 1962–1974* (1983); and *Social Workers: Their Role & Tasks* (1982), known as the Barclay Report, which brings together different views of the role of the personal social services. *The Future of Voluntary Organisations* (1978), known as the Wolfenden Report, gives a clear account of voluntarism in Britain; and HUGH W. MELLOR, *The Role of Voluntary Organisations* (1985), provides more up-to-date analysis. MARTIN BULMER (ed.), *Neighbours: The Work of Philip Abrams* (1986), studies the role of informal care and reviews the debate on the relationship between the formal and the informal sectors. See also RUDOLF KLEIN and MICHAEL O'HIGGINS (eds.), *The Future of Welfare* (1985).

There are a number of useful comparative studies, including BARBARA N. RODGERS, ABRAHAM DORON, and MICHAEL JONES, *The Study of Social Policy: A Comparative Approach* (1979), on the United Kingdom, France, Israel, and Australia; ALFRED J. KAHN and SHEILA B. KAMERMAN, *Social Services in International Perspective: The Emergence of the Sixth System* (1976, reissued 1980), which covers Canada, the United Kingdom, France, West Germany, Poland, Yugoslavia, Israel, and the United States; SHEILA B. KAMERMAN and ALFRED J. KAHN (eds.), *Family Policy: Government and Families in Fourteen Countries* (1978); JOAN HIGGINS, *States of Welfare: A Comparative Analysis of Social Policy* (1981); NEIL GILBERT, *Capitalism and the Welfare State* (1983, reprinted 1985); J.A. YODER (ed.), *Support Networks in a Caring Community: Research and Policy, Fact and Fiction* (1985); and ELSE ØYEN (ed.), *Comparing Welfare States and Their Futures* (1986). See also CATHERINE JONES, *Patterns of Social Policy: An Introduction to Comparative Analysis* (1985).

Comparative studies on socialist policies include VIC GEORGE and NICK MANNING, *Socialism, Social Welfare, and the Soviet Union* (1980); JOHN DIXON, *The Chinese Welfare System, 1949–1979* (1981); and BOB DEACON, *Social Policy and Socialism: The Struggle for Socialist Relations of Welfare* (1983).

Key issues on personal social services in developing countries are discussed in JAMES MIDGLEY, *Professional Imperialism: Social Work in the Third World* (1981); and MARGARET HARDIMAN and JAMES MIDGLEY, *The Social Dimensions of Development: Social Policy and Planning in the Third World* (1982).

Several of the following studies of social security, notably those of Wilensky *et al.*, Köhler and Zacher, Jones, Easton, and Midgley, include material on the personal social services.

Social security: Nearly all countries make periodic reports summarizing their social security provisions; these are published every few years by the UNITED STATES. SOCIAL SECURITY ADMINISTRATION. OFFICE OF RESEARCH, STATISTICS, AND INTERNATIONAL POLICY, with the title *Social Security Programs Throughout the World.* A second basic international source is *The Cost of Social Security,* published irregularly by the INTERNATIONAL LABOUR OFFICE, which contains comparative tables of the costs, benefits, and financing of social security schemes. For the Common Market countries provisions are periodically summarized in *Comparative Tables of the Social Security Systems in the Member States of the European Communities: General Systems,* published by the COMMISSION OF THE EUROPEAN COMMUNITIES.

A summary of the findings of comparative research is found in HAROLD L. WILENSKY *et al., Comparative Social Policy: Theories, Methods, Findings* (1985). For economic aspects see L.D. MCCLEMENTS, *The Economics of Social Security* (1978). The International Labour Office report by 10 experts referred to in the text is *Into the Twenty-First Century: The Development of Social Security* (1984).

The historical evolution of the schemes in Austria, France, Germany, Great Britain, and Switzerland is described against a wide economic, social, and political background in PETER A. KÖHLER and HANS F. ZACHER (eds.), *The Evolution of Social Insurance 1881–1981* (1982). For a penetrating comparison between Britain and Sweden see HUGH HECLO, *Modern Social Politics in Britain and Sweden: From Relief to Income Maintenance* (1974). For the United States see BRUNO STEIN, *Social Security and Pensions in Transition: Understanding the American Retirement System* (1980, reprinted 1983). Other historical studies of advanced societies include M.A. JONES, *The Australian Welfare State: Growth, Crisis and Change,* new ed. (1983); DENNIS GUEST, *The Emergence of Social Security in Canada,* 2nd ed. rev. (1985); and BRIAN EASTON, *Social Policy and the Welfare State in New Zealand* (1980). For the evolution and current role of social security in developing countries, see JAMES MIDGLEY, *Social Security, Inequality, and the Third World* (1984).

Descriptions of health insurance and health services in various western European countries are given in BRIAN ABEL-SMITH and ALAN MAYNARD, *The Organization, Financing, and Cost of Health Care in The European Community* (1978); and BRIAN ABEL-SMITH, *Cost Containment in Health Care: The Experiences of 12 European Countries, 1977–83* (1984). For other countries, see MICHAEL KASER, *Health Care in the Soviet Union and Eastern Europe* (1976); LEE SODERSTROM, *The Canadian Health System* (1978); SIDNEY SAX, *A Strife of Interests: Politics and Policies in Australian Health Services* (1984); and *Medical Care Under Social Security in Developing Countries* (1982), papers from a meeting of the International Social Security Association.

(R.A.P./B.A.-S.)

Modern Socio-Economic Doctrines and Reform Movements

Since the 17th century, and more markedly since the 19th, political life and thought have been affected by systems of belief often characterized as ideological, by which is meant that they derive programs of practical action in the political sphere (understood to include broadly socio-economic concerns) from more or less systematically argued views of the proper structure and ends of human society. The history and general nature of ideological thought is considered in some detail in the *Macropædia* article IDEOLOGY. This article describes the major categories of modern ideology. Individual ideologies associated with particular theorists or historical settings are discussed separately elsewhere in the encyclopaedia; *e.g.,* for a full treatment of Marxism, see the *Macropædia* article MARXISM; for information on Nazism, see in the *Micropædia* under NATIONAL SOCIALISM and NAZI PARTY.

This article is divided into the following sections:

Socialism

Socialism refers to both a set of doctrines and the political movements that aspire to put these doctrines into practice. Although doctrinal aspects loomed largest in the early history of socialism, in its later history the movements have predominated over doctrine, so much so that there is no precise canon on which the various adherents of contemporary socialist movements agree. The most that can be said is that socialism is, in the words of Anthony Crosland, a British socialist, "a set of values, or aspirations, which socialists wish to see embodied in the organization of society."

Problems of definition

Although it is possible to trace adumbrations of modern socialist ideas as far back as Plato's *Republic,* Thomas More's *Utopia,* and the profuse Utopian literature of the 18th-century Enlightenment, realistically, modern socialism had its roots in the reflections of various writers who opposed the social and economic relations and dislocations that the Industrial Revolution brought in its wake. They directed their critical shafts against what they conceived to be the injustice, the inequalities, the suffering brought about by the capitalist mode of production and the free and uncontrolled market on which it rested. To the acquisitive individualism of their age they opposed a vision of a new community of producers bound to each other through fraternal solidarity. They conceived of a future in which the masses would wrest control of the means of production and the levers of government from the capitalists.

Although the great majority of men calling themselves socialists in the 19th and 20th centuries have shared this vision, they have disagreed about its more specific ideas. Some of them have argued that only the complete nationalization of the means of production would suffice to implement their aims. Others have proposed selective nationalization of key industries, with controlled private ownership of the remainder. Some socialists insist that only strong centralized state direction and a command economy will suffice. Others advocate a "market socialism" in which the market economy would be directed and guided by socialist planners.

Socialists have also disagreed as to the best way of running the good society. Some envisage direction by the government. Others advocate as much dispersion and decentralization as possible through the delegation of decision-making authority to public boards, quasi-public trusts, municipalities, or self-governing communities of producers. Some advocate workers' control; others would rely on governmental planning boards. Although all socialists want to bring about a more equal distribution of national income, some hope for an absolute equality of income, whereas others aim only at ensuring an adequate income for all, while allowing different occupations to be paid at different rates.

"To each according to his need" has been a frequent battle cry of socialists, but many of them would in fact settle for a society in which each would be paid in accordance with his contribution to the commonwealth, provided that society would first assure all citizens minimum levels of housing, clothing, and nourishment as well as free access to essential services such as education, health, transportation, and recreation.

Socialists also proclaim the need for more equal political rights for all citizens, and for a levelling of status differences. They disagree, however, on whether difference of status ought to be eradicated entirely, or whether, in practice, some inequality in decision-making powers might not be permitted to persist in a socialist commonwealth.

The uses and abuses of the word socialism are legion. As early as 1845, Friedrich Engels complained that the socialism of many Germans was "vague, undefined, and undefinable." Since Engels' day the term socialism has been the property of anyone who wished to use it. The same Bismarck who as German chancellor in the late

1870s outlawed any organization that advocated socialism in Germany declared a few years later that "the state must introduce even more socialism in our Reich." Modern sophisticated conservatives, as well as Fascists and various totalitarian dictators, have often claimed that they were engaged in building socialism.

ORIGINS OF THE SOCIALIST IDEA

The utopians of the early 19th century

The term socialism, in its modern sense, made its first appearance around 1830. In France it was applied to the writings of Fourier and the Saint-Simonians and in Britain to those of Robert Owen.

Saint-Simon and Fourier. Comte Henri de Saint-Simon (1760–1825) was an erratic genius with a fertile and yet disorganized mind. His socialist writings revolved around the idea that his age suffered from an unhealthy and unbridled individualism resulting from a breakdown of order and hierarchy. But he held that the age also contained the seeds of its own salvation, which were to be found in the rising level of science and technology and in the industrialists and technicians who had already begun to build a new industrial order. The joining of scientific and technological knowledge to industrialism would inaugurate the rule of experts. The new society could not be equalitarian, Saint-Simon argued, because men were not equally endowed by nature. Yet it would make the maximum use of potential abilities by assuring that everyone would have equal opportunity to rise to a social position commensurate with his talents. By eradicating the sources of public disorder, it would make possible the virtual elimination of the state as a coercive institution. The future society would be run like a gigantic workshop, in which rule over men would be replaced by the administration of things.

Saint-Simon's followers bent the founder's doctrine in a more definitely socialist direction. They came to see private property as incompatible with the new industrial system. The hereditary transmission of power and property, they argued, was inimical to the rational ordering of society. The rather bizarre attempt of Saint-Simon's followers to create a Saint-Simonian church should not obscure the fact that they were among the first to proclaim that bourgeois-capitalist property was no longer sacrosanct.

François-Marie-Charles Fourier (1772–1837), a lonely and neglected thinker who was more than a little mad, was led to his anticapitalist vision by a loathing for a world of competition and wasteful commerce in which he spent most of his life as a salesman. Possessed by an inordinately wide-ranging imagination, he argued that the regenerated world to come would be characterized not only by social but also by natural and even cosmological transformations. The ocean would be changed into lemonade, and wild animals would turn into anti-lions and anti-tigers serving mankind.

Fourierist communities

With meticulous and obsessive care, Fourier set forth plans for his model communities, the *phalanstères,* the germ cells of the good society of the future. In these communities men would no longer be forced to perform uncongenial tasks but would work in tune with their temperaments and inclinations. They would cultivate cabbages in the morning and sing in the opera in the evening. Fourier's was an antinomian vision in which human spontaneity made outside regulation unnecessary. Whereas Saint-Simon called for the rule of experts, Fourier was convinced that love and passion would bind men together in a harmonious and noncoercive order.

Owenism. The Welshman Robert Owen (1771–1858) held more sober views. Early in his career he became known as a model employer in his textile works in Scotland, and as an educational and factory reformer. Despairing of his fellow capitalists he later turned to the emergent trade union movement. Acutely conscious of the evils of industrialism by which he had acquired his wealth, he thought that the new productive forces could be turned to the benefit of mankind if competition were eliminated and the effects of bad education were counteracted by rational enlightenment. He advocated cooperative control of industry and the creation of Villages of Unity and Cooperation in which the settlers, in addition to raising crops, would improve their physiques as well as their minds. Owenite communities established in New Harmony, Indiana, and elsewhere in America all failed. His attempts to join the cooperative and the trade union movements in a "great trades union" also proved a failure. Yet he left a lasting imprint on the British socialist tradition; his indictment of the competitive order, his stress on cooperation and education, his optimistic message that men could increase their stature if only the stultifying effects of an unhealthy environment were removed have continued to inform the socialist movement.

Other early socialists. The 1840s saw the rise of a number of other socialist doctrines, particularly in France. Louis-Auguste Blanqui evolved a radical socialist—or, as he called it, communist—doctrine based on a democratic populism and on the belief that capitalism as an inherently unstable order would soon be replaced by cooperative associations. Impatient with theorizing, given to a strong belief in voluntarism and the virtues of revolutionary action, he is remembered for his many attempts at organizing insurrections rather than for his theoretical contributions.

Étienne Cabet, in his influential utopian work *Voyage en Icarie* (1840), carried on the tradition of Thomas More as well as of Fourier. Louis Blanc is best known for *L'Organisation du travail* (1839), in which he advocated the establishment of national workshops with capital advanced by the government. These workshops would remain free from government control, with workers electing their management. The national workshops he organized in Paris after the revolution of 1848 were soon dissolved by a resurgent middle class. His plans for the "organization of labour" and his pleas for the recognition of the "right to work" were nevertheless a foreshadowing of the modern welfare state.

Pierre-Joseph Proudhon (1809–65) is best viewed as one of the founders of the anarchist tradition. But his attacks against private property and the institutions on which it rests, as well as his championing of a system of human relationships in which reciprocity, equity, and justice would replace what he saw to be rapacity, exploitation, and greed, powerfully stimulated the socialist imagination. His anti-statist and federalist vision of producers' communities provided a counterweight to the centralizing and statist impulses in the socialist tradition.

In England, the first half of the 19th century saw the emergence of a number of writers attacking the inequities of capitalism and basing their indictment of wage labour on radical interpretations of the thinking of an eminent economist, David Ricardo. Somewhat later, a Christian socialist movement led by Frederick Denison Maurice and Charles Kingsley attempted to combine radical economic views with political conservatism. The radical Chartist Movement of the 1830s and 1840s is better viewed as a political movement of the working class than as a specifically socialist formation, though anticapitalist ideas played a strong part in it.

MARX AND THE RISE OF SOCIAL DEMOCRACY

The idea of class struggle

In the perspective of intellectual history, all of these pre-Marxist socialist thinkers produced ideas of considerable intrinsic worth. But from the viewpoint of the subsequent development of socialism their ideas seem to be tributaries feeding the mighty stream of the Marxist movement that came to dominate the socialist tradition in the last third of the 19th century.

The Communist Manifesto. Karl Marx (1818–83) had a synthesizing mind. He fused German idealistic philosophy with British political economy and French socialism. Marx's earlier writings are discussed elsewhere (see MARXISM). In this section the focus is on his mature thought as first developed in *The Communist Manifesto* (1848), which he wrote in conjunction with Friedrich Engels, his lifelong intellectual companion.

To Marx, society is a moving balance of antithetical forces; strife is the father of all things, and social conflict is the core of the historical process. Men struggle against nature to wrest a livelihood from her. In the process they enter into relations with one another, and these relations differ according to the stage they have reached in their productive activities. As a division of labour emerges in human society, it leads to the formation of antagonistic classes that are the prime actors in the historical drama. In contrast to his predecessors, Marx did not see history as simply a struggle between the rich and the poor, or the powerful and the powerless; he taught that such struggles differ qualitatively depending on what particular historical classes emerge at a given stage in history. A class is defined by Marx as a grouping of men who share a common position in the productive process and develop a common outlook and a realization of their mutual interest.

Marx, like Hegel and Montesquieu, considered societies as structured wholes; all aspects of a society—its legal code, its system of education, its religion, its art—are related with one another and with the mode of economic production. But he differed from other thinkers in emphasizing that the mode of production was, in the last analysis, the decisive factor in the movement of history. The relations of production, he held, constitute the foundation upon which is erected the whole cultural superstructure of society.

Marx distinguished this doctrine, which he called scientific socialism, from that of his predecessors whom he labelled utopian socialists. He asserted that his teachings were based on a scientific examination of the movement of history and the workings of contemporary capitalism rather than simply on idealistic striving for human betterment. He claimed to have provided a guide to past history as well as a scientific prediction of the future. History was shaped by class struggles; the struggle of contemporary proletarians against their capitalist taskmasters would eventuate in a socialist society in which associated producers would mold their collective destinies cooperatively, free from economic and social constraints. The class struggle would thus come to an end.

The growth of organized parties

The First International. *The Communist Manifesto,* which had been written as a program for the Communist League, a group of continental workmen, failed to have an impact on the European revolutions of 1848. For a number of years thereafter Marx and Engels lived in complete isolation from the labour movements developing in England and on the Continent. Socialism in those years was only the creed of isolated sects, often of exiles. In 1864, however, after a gathering in London of continental and English workers' representatives and associated intellectuals, there emerged the International Working Men's Association, commonly known as the First International. Although it encompassed various tendencies ranging from simple trade unionism to anarchism, Marx dominated it from its inception and made it an instrument for the diffusion of his message. Its headquarters were in London, but it never exerted much influence in England, where the labour movement remained impervious to Marxist revolutionary ideology. On the Continent, particularly in Germany, Marxism spread rapidly and soon became the major doctrine of the emerging labour movement.

German Social Democracy. In Germany, Ferdinand Lassalle (1825–64), the architect of the German labour movement, agreed with Marx on the need for autonomous organization of the working class but differed from him in wanting the government to provide the necessary capital for the establishment of producers' cooperatives that would emancipate labour from capitalist domination. To Marx, any appeal to the bourgeois state was out of the question, and he proceeded to organize followers in Germany against Lassalle. In 1869 they created the Social Democratic Party. The division between the followers of Lassalle and those of Marx persisted until 1875, when the two parties united on the basis of a compromise program (which Marx sharply criticized for its Lassallean vestiges).

The German Social Democratic movement grew rapidly, despite Chancellor Otto von Bismarck's attempts to suppress it through anti-socialist legislation and to undercut its appeal by social reforms. In 1877 the Socialists obtained half a million votes and a dozen members in the Reichstag. In 1881 the party claimed 312,000 members, and, by 1890, 1,427,000. After the repeal of the anti-socialist laws the party adopted the so-called Erfurt Program of 1891, eliminating all demands for Lassallean state-aided enterprises and pledging itself to the orthodox Marxian goal of "the abolition of class rule and of classes themselves."

Revisionism. It soon became apparent that Marx's own thought had gone through a process of evolution so that different disciples could quote chapter and verse in support of fairly divergent political views. In particular, whereas Marx in the late 1840s and early 1850s had asserted that only a violent revolutionary overthrow of bourgeois rule and the emergence of the "dictatorship of the proletariat" would lead to the emancipation of the working class, by the late 1860s his views had considerably mellowed. Writing in England after the second Reform Bill (1867), which had given the vote to the upper strata of the workers, Marx suggested the possibility of a peaceful British evolution toward socialism. He also thought that such a peaceful road might be possible in the United States and in a number of other countries.

Although the leaders of German Social Democracy liked to speak in revolutionary Marxist rhetoric, they had in daily life become increasingly absorbed in parliamentary activities. Under the intellectual guidance of their theoretician Karl Kautsky (1854–1938) they developed a brand of economic determinism according to which the inevitable development of economic forces would necessarily lead to the emergence of socialism. The official Social Democratic platform remained ideologically intransigent, while the party's activities became increasingly pragmatic.

Eduard Bernstein (1850–1932), once a close companion of Engels, challenging prevailing orthodoxy in his famous *Die Voraussetzungen des Sozialismus und die Aufgaben der Sozialdemokratie* (1899; Eng. trans., *Evolutionary Socialism,* 1909), appealed to the party to drop its revolutionary baggage and recognize theoretically what it had already accepted in practice: namely, that Germany would not have to go through revolutionary convulsions in order to reach socialist goals. Ignoring the differences between political conditions in Germany and England, Bernstein urged the party to travel along the English road in hope of gradually transforming capitalism through socialist reforms brought about by parliamentary pressure.

Defeat of revisionism

The struggle between Kautsky's orthodoxy and Bernstein's revisionism shook the German party. Bernsteinian doctrine was officially defeated in 1903, but revisionism in fact permeated the party, especially its parliamentary and trade union leaders. At the outbreak of World War I practically all the leaders supported the government and the war, thus ending the party's revolutionary pretensions.

Other Social Democratic parties on the Continent. In France, the Marxists had to contend with rival socialist traditions that had profound roots in French working-class history. The followers of Blanqui and Proudhon played leading roles in the Paris Commune of 1871. In the years that followed, French socialism was torn by conflicting tendencies. The Parti Ouvrier founded by Jules Guesde in 1875–76 represented Marxist orthodoxy, but there were other socialist parties that reflected the influence of Blanqui, Blanc, and Proudhon, as well as the 18th-century revolutionary heritage. Even after the various parties amalgamated in 1905, the movement continued to be torn by dissension between its revolutionary and reformist wings. Nonetheless, it continued to grow. At its first congress the unified party claimed 35,000 members, and in the elections of 1906 it won 54 seats in Parliament. By 1914 it had more than 100 members in the Chamber of Deputies. As in Germany, however, revolutionary rhetoric usually went hand in hand with pragmatic action, and the party became in fact a skillful participant in the parliamentary games of the Third Republic. After Jean Jaurès, the great Socialist orator and a principled leader of the peace elements, was assassinated on the eve of World War I, most of the Socialists supported the French war effort.

In the last part of the 19th century, Social Democratic parties generally beholden to Marxist doctrine sprang up in most of the countries of continental Europe. A Danish Social Democratic Party was founded in the 1870s, the Swedish Socialist movement in 1889. The Norwegian Labour Party (first called the Social Democratic Party) was formed in 1887 but became a major political force only in the early 20th century. In central Europe, Social Democratic parties fairly rapidly assumed a major place on the political horizon. An Austrian Social Democratic

Party was founded in 1888. By 1908 it had gained about one-third of the vote cast in the parliamentary elections, to become the strongest Socialist party outside Germany. The Belgian Labour Party, formed in 1885 as an amalgamation of trade union, cooperative, and other groups, rapidly organized thousands of mutual aid societies, built a very strong trade union movement, and led a number of general strikes on behalf of more liberal suffrage laws. The Dutch Socialist-Democratic Workers Party, founded in 1894, became a significant force only in the years immediately preceding World War I. It held 20 percent of the seats in the lower house of Parliament in 1912.

All of the continental parties were torn by internal tensions. Proposals to enter liberal coalition governments often were defeated by only narrow margins; Marxist orthodoxy prevailed only after sharp struggles. In The Netherlands, for example, a proposal to enter a coalition government was rejected by the close vote of 375 to 320 at the party congress of 1913.

Anarchist tendencies. In the less industrialized parts of Europe, particularly in Italy and Spain, Marxism had to contend with anarchist tendencies mainly rooted in the precapitalist and peasant strata. European anarchism as a political force was created by Mikhail Bakunin, the highly influential Russian libertarian thinker. His Anarchist Federation had belonged to the First International, but quarrels with Marx led to the expulsion of Bakunin and his followers in 1872.

The Anarchist Federation

Bakuninist and other anarchist strains of thought remained powerful in Spain, despite the founding of the Social Labour Party in 1879. The Spanish socialist movement suffered from the competition of the anarchists throughout its subsequent history, and only after World War I did it become a political force to be reckoned with.

In Italy anarchist tendencies also impeded the growth of a socialist movement. The Italian representatives to the First International followed Bakunin's lead. Not until 1892 was a distinctly Socialist party formed under the leadership of Filippo Turati. In 1913, after the electoral franchise was broadened, the official Socialist Party secured 51 seats in Parliament, and two other Socialist parties that had split from its ranks gained 31 seats. Although it continued to suffer from internal dissension and from anarchist tendencies in the more backward areas of the country, by World War I the Italian Socialist Party had become one of the strongest Marxist organizations in Europe.

The Second International. The First International had brought into being a variety of Socialist movements throughout Europe. When these began to grow roots in their respective political systems, it became apparent that the international movement could no longer be controlled by a single directing centre. After the dissolution of the First International in 1876, Marx and Engels remained father figures whose counsel the movement eagerly sought; but they could no longer direct it. The history of socialism now became largely the history of separate national movements that, for all their ceremonial acknowledgment of Marxist orthodoxy, increasingly tended toward a revisionist and nonrevolutionary line. By the early years of the 20th century socialism had become a powerful parliamentary force in most European countries. Except in Russia, where autocracy still held sway, the Socialists were reformers seeking a transformation of the existing system rather than its violent overthrow. Only left-wing minorities within the various parties still stood for revolutionary orthodoxy.

The Second International, founded in 1889, reflected the changed character of the movement. It was a kind of international parliament of socialist movements rather than the unified and doctrinally pure organization that the First International had attempted to be. It was dominated by the German party. With traditional Marxist rhetoric, the German delegates stood adamant against proposals to sanction socialist participation in bourgeois governments, and thus appeared to favour a "left" course. But socialist participation in government was not a realistic option in Kaiser William's Germany, and so the German delegates could be intransigent at no cost to themselves. When the issue was put to a vote at the Amsterdam congress in 1904, the Germans sided with those who opposed participation, against Jaurès and those who condoned it. But Jaurès had the better of it when he pointed out that "behind the inflexibility of theoretical formulas which your excellent comrade Kautsky will supply you with till the end of his days, you concealed . . . your inability to act." As with the issue of government participation, so with the issue of war. The Second International, under its German leadership, issued many moving and stirring manifestoes against war, but when war broke out it disclosed its paralysis. Most of its national components sided with their own governments and abandoned the idea of international working-class solidarity. Almost all of them recognized what they may secretly have believed for a long time: the workers, after all, had a fatherland.

OTHER SOCIALIST TENDENCIES BEFORE WORLD WAR I

The non-Marxists

British Fabianism. Although Marxism triumphed in the continental Socialist movement, it did not do so in Great Britain. Henry Hyndman, a radical journalist, founded the Social Democratic Federation on strictly Marxist principles in the 1880s, but it ever remained marginal to the British socialist movement. The Socialist League, founded by the poet William Morris, propounded libertarian-syndicalist ideas and likewise failed to make headway. Fabian socialism, on the other hand, based on non-Marxist ideas, was to have an enduring influence in Britain.

The Fabian Society was organized in the 1880s by a number of young radical intellectuals among whom Sidney and Beatrice Webb, Graham Wallas, Sidney Olivier, and George Bernard Shaw were the most outstanding. It developed an evolutionary and moderate form of socialism. Convinced of "the inevitability of gradualness," the Fabians never endeavoured to become a mass organization but preferred to be a ginger group of intellectuals working to transform society through practical and unobtrusive advice to the men of power. The extremely influential *Fabian Essays* (begun in 1889) contained detailed blueprints for social legislation and reform that influenced policymakers whether they were socialists or not. Through "permeation," which Shaw defined as "wire-pulling the government in order to get socialist measures passed," the Fabians attempted to convince key politicians, civil servants, trade union officials, and local decision makers of the need for planned and constructive reform legislation. Basing their doctrine at least as much on non-Marxist economics as on the continental socialist tradition, they worked for a new order "without breach of continuity or abrupt change of the entire social tissue."

Syndicalism. The syndicalist movement grew out of French trade unionism when it was reconstituted after the bloodletting of the Paris Commune (1871). Convinced of the futility of parliamentary and political activity, the syndicalists stressed that only direct action by workers organized in their unions would bring about the desired socialist transformation. Under the leadership of Fernand Pelloutier the Fédération des Bourses du Travail (founded in 1892), which was later amalgamated with the Confédération Générale du Travail (1902), was built on the idea that the emancipation of labour would come through a "general strike" that would paralyze the country and deliver power into the hands of organized workers. The unions would become the directing and administering nuclear cells of production.

The general strike tactic

The syndicalists attracted a number of intellectuals to their ranks, who attempted to provide a philosophical basis for syndicalism and its rejection of the political road to socialism. The most important of their writings, Georges Sorel's *Réflexions sur la violence* (1908; Eng. trans., *Reflections on Violence,* 1916), has continued to exercise considerable influence on the thinking of revolutionary militants, even though Sorel himself soon shifted his allegiance to the extreme right.

Guild socialism. The guild socialist tradition developed in Britain in the years before World War I. Sharing the general socialist hostility to the wage system and production for profit, guild socialists took from the syndicalists their distrust of the state and their emphasis on producers' control. They looked back to the Middle Ages when independent producers, organized in guilds, controlled the conditions of their employment and took pride in creative work. Aiming at self-government in industry, guild socialists urged that industrial organizations, churches, trade unions, cooperative societies, and municipalities be granted autonomy. They argued that every group in society should carry out its particular functions without control from above, and that individuals should have a say in the direction of all those functional units in which they happened to be interested. Cooperation between functional units would replace direction by the state, which would be restricted to providing needed national services such as police protection. The state would be a functional unit among many others, rather than an all-encompassing sovereign.

Although guild socialism owes its origin to several thinkers, it grew into a mature doctrine only when in 1913 it recruited G.D.H. Cole, a brilliant Oxford don, two of whose early books, *The World of Labour* (1913) and *Self-Government in Industry* (1917), contain the best exposition of guild socialist doctrine. The movement never attained wide popular appeal but has continued to be a source of ideas in the British labour movement, if only as a counterpoint to the bureaucratic and centralizing tendencies of Fabianism.

Socialism in the United States. Socialism never became as influential in the United States as it did in Europe. When the Socialist Party was formed in 1901 it claimed a membership of 10,000 that grew to 150,000 in 1912, in which year the party polled a presidential vote of 897,000, or 6 percent of the national total. Although its strongest roots were among recent immigrants from Europe, it also drew its inspiration from the utopian colonies of the 19th century, from the slavery abolitionists, trade unionists, and agrarian reformers, and from isolated socialist groups of the 1880s and 1890s.

The Socialist Labor Party, a predecessor of the Socialist Party, was formed in 1877 but acquired a distinct outlook only when the journalist and polemicist Daniel De Leon joined it in 1890. De Leon attempted to marry a doctrinaire brand of Marxism to a "labourism" nourished in part on French syndicalist doctrine. He and his followers wished to raise the membership of the unions above "paltry routine business" and prepare them for a successful contest with the power of capital, both at the ballot box and in industrial combat.

The U.S. Socialist Party

The Socialist Labor Party remained a sect. But the Socialist Party developed into a mass movement under the leadership of Eugene Debs, a former union official who had been converted to socialism by reading the works of various socialist writers while in jail. The Socialist Party of Debs was neither centralized nor politically homogeneous. In its ranks it harboured reformists and revolutionaries, orthodox Marxists, Christian ministers, municipal reformers, populists who hated the railroads and the trusts, and Jewish garment workers dreaming of fraternity in the sweatshops. It produced no major theoretical works, but it managed in its undoctrinaire way to be an effective voice for the idea of socialism in America. It declined after World War I, its last well-known leader being Norman Thomas.

THE RISE OF RUSSIAN SOCIALISM

The populist tradition. The dominant radical tendency in 19th-century Russia was populism, a doctrine first developed by the author and editor Aleksandr Herzen, who saw in the peasant communes the embryo of a future socialist society and argued that Russian socialism might skip the stage of capitalism and build a cooperative commonwealth based on ancient peasant tradition. Herzen idealized the peasantry. His disciples inspired many students and intellectuals to "go to the people" in order to stir them into revolutionary action.

In the 1860s and 1870s, the more radical populists lost their faith in a peasant revolt and turned instead to terrorism. Small groups of student revolutionaries sought to bring down tsarism through terroristic action; their efforts culminated in the assassination of Alexander II in 1881. Sergey Nechayev's *Revolutionary Catechism,* in the writing of which Bakunin had a hand, stressed that the sole aim of the revolutionary is to destroy "every established object root and branch, [to] annihilate all state traditions, orders and classes in Russia." It is one of the ironies of history that Bakunin helped create in Russia an elitist and terrorist movement composed almost exclusively of alienated intellectuals, while in western Europe he appealed to skilled craftsmen and peasants and appeared to be the heir of Proudhon.

Within the broad stream of populism, terrorism was opposed by an evolutionary socialism that put its faith in peaceful propaganda and the education of the masses. While the elitists pursued their campaign of terror, the gradualists stuck to propaganda among the people.

Marxism in prerevolutionary Russia. The father of Russian Marxism was Georgy Plekhanov, who began his socialist career as a populist and was converted to Marxism when he settled in Geneva in 1880; in 1883 he founded the first Russian Marxist organization, the Osvobozhdenie Truda (Liberation of Labour Group). Plekhanov thought Russian socialism ought to be based primarily on the growing factory proletariat. Rejecting Herzen's idea that Russia was exceptional, he held that the revolution would be European in character and that Russia's place in it would be determined by its own labour movement. In a variety of books and pamphlets in the 1880s and 1890s, Plekhanov attacked the populists and argued that Marx had shown the objective historical necessity of socialism. The laws of social evolution could not be flouted. A bourgeois revolution in Russia was inevitable in the course of industrial development. The organized working class would know how to take advantage of the bourgeois revolution and push it forward.

The Leninists

Against this German brand of Marxism, Vladimir Ilich Ulyanov (1870–1924), later to be known by his party name of Lenin, argued for a more militant approach to revolution. In *What Is To Be Done?* (1902) he formulated his characteristic doctrine. Socialism would be achieved only when professional revolutionaries succeeded in mobilizing and energizing the masses of workers and peasants. Left to themselves, the workers would get no farther than a trade union consciousness. A militant, disciplined, uncompromising organization of revolutionaries was needed to propel the masses into action.

Lenin's followers parted company with the other Russian Marxists at the second congress of the (illegal) Russian Social Democratic Workers Party held in London in 1903. The anti-Leninist position was formulated by the leader of the more orthodox Marxists, L. Martov, when he declared, "In our eyes, the labour party is not limited to an organization of professional revolutionaries. It consists of them, plus the entire combination of the active, leading elements of the proletariat . . . ."

Mensheviks and Bolsheviks

The two factions within the Russian Social Democratic movement at first cooperated and even held joint meetings; the final split came only in 1912. Individual leaders switched from one faction to another (Plekhanov, who originally sided with Lenin, joined his opponents in 1904). Others, such as Leon Trotsky, attempted for a time to stay free from factional alignments. These disputes were fought out in the West, where most of the leaders of both sides lived as émigrés. Within Russia itself, however, Lenin's opponents (the Mensheviks) mainly attracted the better educated and skilled workers, as well as the Jewish intelligentsia, while his Bolsheviks tended to be most successful among the more backward strata of the working class.

After the February Revolution of 1917 toppled the tsarist regime and installed a liberal and vaguely socialistic leadership, the Bolsheviks managed to extend their organization among the urban masses. When Lenin returned from exile in April 1917, he startled his followers by calling for an entirely new strategy. Previously they had believed that their immediate task was to work within the limits of a democratic republic while preparing for future revolutionary opportunities. Lenin argued instead that they must seek power at once. The desire of the masses for an immediate end to the war, the land hunger of the peasantry, the feebleness of the new regime, he urged, made possible what had not been possible in the abortive revolution of 1905: a socialist revolution led by Bolshevik cadres. Moreover, Lenin argued, a Russian revolution would not be isolated for it would soon be followed by a German revolution.

The soviets (workers' and peasants' councils), which had sprung up spontaneously when the tsarist power collapsed, were the main organizational bases from which the Bolsheviks mounted their assault on the established order. Lenin's slogan "All power to the soviets" found a ready response in the major urban centres. In September 1917 the Bolsheviks won elections for the Moscow and St. Petersburg soviets. These now became centres of "dual power" challenging the official government. It was the St. Petersburg soviet that in October 1917 gave Trotsky the military instrument with which he was able to topple the provisional government and install a revolutionary regime headed by Lenin.

Lenin and the Third International. The Bolshevik seizure of power had been undertaken in the belief that the revolution would soon spread to the rest of Europe. Lenin's perspective had always been internationalist. When most of the socialist leaders of the Second International rallied to their national governments in 1914, Lenin denounced them as traitors to the cause and sought to lay the groundwork for a new organization of revolutionary Socialists. After their seizure of power, the Bolsheviks resolved to create a Third International. By the time the delegates had assembled in Moscow in 1919, a revolutionary uprising in Berlin had been crushed and its leaders murdered. The great majority of the German working class was evidently willing to give the Social Democratic leadership of the new German republic a chance. But to the Russian leaders' world revolution still seemed near. Soon after the first congress of the Third International a short-lived soviet republic was proclaimed in Hungary and another in the German state of Bavaria. Communist parties began to be organized in all the major countries of Europe.

When the so-called Communist International (Comintern) met for its second world congress in July 1920, it was no longer a small gathering of individuals or representatives of small sects but a union of delegations from a dozen major Communist parties. The outcome of this meeting was to give the Russian leaders control of the new International, now broken away sharply from the Socialist movement. It adopted 21 conditions for membership in the Comintern, demanding that its adherents reject not only those Socialist leaders who had been "social patriots" in the war but also those who had taken a middle position. It aimed at creating a disciplined and militantly revolutionary world organization patterned after the Russian model, which would accept willingly the direction and unquestioned authority of the Russian leadership.

By 1923 the hoped-for revolutionary tide in Europe had not developed. New uprisings in parts of Germany failed completely in 1923. The Red Army's attempted invasion of Poland had been thrown back. Many Socialists who had for a time joined the Comintern, including the leadership of the Norwegian Labour Party, left-wing Communists in Germany, and Syndicalists in France and Spain, now turned away, rejecting its policy of centralized dictation.

Europe achieved a measure of economic and social stabilization. By the time of Lenin's death in 1924, Moscow was beginning to use the parties over which it still held command as instrumentalities of Russian foreign policy. Although some Comintern leaders like Trotsky still believed that world revolution was on the agenda, their faith was no longer shared by the majority of the Russian leadership.

SOCIALISM BETWEEN THE WARS

Socialists and Communists

The split with the Communists. Communists throughout the world denounced the leaders of the reconstructed Socialist parties as "social traitors" who "objectively" fostered the maintenance of capitalism. They accused them of having repudiated Marxism and betrayed international socialism by collaborating during the war with the bourgeoisie in the defense of their national states. The Socialist leaders retorted by pointing to the dictatorial features of the Soviet state and accusing the Communists of having betrayed the democratic socialist tradition.

The European Socialist movement was irremediably split. In Germany, the Social Democrats united again and succeeded in enrolling the bulk of the working class under their banner; the Communists were reduced to a minority position in the German labour movement. In France, where the Communists at first succeeded in attracting the majority of the Socialist Party, their opponents soon regained ascendancy and the Communists became a minority on the French left. Italian socialism split into Communists and left-wing and right-wing Socialists and thus greatly facilitated Mussolini's march to power. In Great Britain the Communists hardly made a dent in the Labour Party and never became more than a radical sect. European socialism as a whole, as well as socialist movements on other continents, was sharply split between adherents of the Second International and the Communists organized in the Third.

The Comintern followed an erratic course, sometimes veering toward a revolutionary line and sometimes making attempts to collaborate with the more militant strata of the socialists. After the onset of the economic depression in 1929 the Comintern took a sharp leftist turn, expecting the "final crisis" of capitalism to bring proletarian revolution everywhere. It denounced Social Democratic leaders as "social Fascists" and enemies of the working class. In the Prussian Landtag the Communists actually voted with the Nazis to bring down a Social Democratic government, on the theory that the Nazi movement was a passing phenomenon.

At the same time the Socialists gave up in practice, though not always in theory, their commitment to revolutionary doctrine. They became in effect pressure groups trying to extract maximum advantages for the working classes from their respective national regimes. In Germany, in Britain, and in the Scandinavian countries they participated at times in the government. Elsewhere, as in France, they tended to support congenial left-bourgeois regimes. But they lacked, on the whole, a concrete plan of social and economic action, and consequently were ineffective when the world depression unsettled the economies and political regimes of western and central Europe.

Response to the world economic crisis. Nowhere, except in Sweden and Belgium, did the socialists press for comprehensive socialist planning during the depression. Where they were in power they followed orthodox policies of budgetary management and public finance. When they were out of power they contented themselves with a defense of the immediate interests of the workers by demanding more unemployment insurance and opposing reductions in wages.

As the crisis deepened, the Communists gained influence, particularly among the unemployed and those unskilled workers hit most severely by the depression. They did not make deep inroads among other workers.

The rise of Fascism. Hitler's rise in Germany led to the destruction of both the Communists and the Socialists in that country. The Communists had hoped that a Nazi victory would be only temporary, and that afterward they would be called upon to lead the masses of Germany to victory. Their battle cry was, "After the Nazis—We." The Socialists played politics as usual, expecting that the depression would run its "natural" course and that a gradual decline of the Nazi fever would follow. A disunited labour movement proved unable to stay the Nazi march to power. This disaster led both Communists and Socialists to reconsider their previous policies and to revise their strategy and tactics.

Austrian Socialists, threatened with destruction by the reactionary regime of Chancellor Engelbert Dollfuss, resolved to offer armed resistance in February 1934. The Austrian party had long been regarded as a model for both its theoretical contributions and its concrete accomplishments. It enjoyed the nearly total support of the workers; 500,000 of Vienna's 2,000,000 inhabitants were dues-paying members. But the party was almost completely metropolitan and urban. Consequently the bloody battles of February 1934 remained localized in Vienna. The uprising was suppressed after four days, and the party had to go underground.

Socialists in office

Experience in government. *Germany.* The end of World War I had seen a somewhat reluctant Social Democratic Party installed in the seat of German government. Friedrich Ebert, the head of the party, became the first president of the new republic. But the Socialists were split internally. The "majority Socialists," the right wing of the party, wished to proceed in a cautious and pragmatic manner. The "independent Socialists," led by Kautsky and his former antagonist Bernstein, pressed for fundamental structural reforms. The extreme left, led by Rosa Luxemburg and Karl Liebknecht, wished to organize a revolutionary party and founded the Communist Party of Germany. When younger extremists, overruling Luxemburg and Liebknecht, organized a left-wing Putsch early in 1919, they were isolated and easily defeated by the government of the majority Socialists and its allies among right-wing officers. Luxemburg and Liebknecht were assassinated, and the remaining leaders took the group into the Comintern. Another left-wing and Communist putsch in Bavaria a few months later was also unsuccessful. In the early 1920s, the independents reunited with the majority Socialists.

In the first election to the new National Assembly in 1919 the majority Socialists obtained a plurality of the votes cast (39.3 percent), and the independent Socialists won another 8 percent. The Socialist government proclaimed the need for socialization of monopolistic industries and other radical measures. But after the elections of June 1920, a non-Socialist cabinet took office. In subsequent years the cabinets were largely non-Socialist in character, though Socialists participated in some of them. The middle classes were again in the saddle, and when President Ebert died in 1925 the conservative nationalist Hindenburg succeeded him. Throughout the turmoil of the first years of the Weimar Republic, the Social Democrats remained a bulwark of republican legality against both the extreme right and the extreme left. In the *Länder* (states), Prussia in particular, they held positions of governmental power and managed to institute a number of reformist welfare measures. But they failed to gain a controlling voice in national politics.

In the May 1928 elections the Social Democrats emerged as the strongest party in the Reichstag. Although they lacked a majority, their leader Hermann Mueller became chancellor, and their financial expert was named minister of finance. This largely Socialist government, however, proved unable to deal with the economic depression that soon afflicted Germany along with the rest of the world. The government followed an orthodox deflationary policy, pressed for the reduction of unemployment benefits in order to save taxes, and attempted to reduce budget deficits. Unable to stem the tide of depression, it resigned in 1930. This was the last government of the Weimar Republic in which Social Democrats participated. Soon afterward, the Nazis started on their way to power.

Britain. In the general election of 1923 the Labour Party, which had adopted a Socialist program only five years earlier, won a plurality; with the support of the Liberals it formed the first Labour government under Ramsay MacDonald in January 1924. Its tenure proved short. After implementing a few modest reform measures, it was ousted by an electorate which, partly because of manufactured fears of a "Bolshevist menace," turned sharply to the right in the elections of October 1924.

In June 1929 the Labour Party had its second chance. It won 288 out of 615 seats in the House of Commons and, with the support of the Liberals, formed the second Labour government, again under Ramsay MacDonald. But Labour, like the German Social Democrats, proved unable to deal with the depression, particularly with mounting unemployment. It was pledged to far-reaching social reforms that it was not prepared to carry out. The flight of capital from London assumed catastrophic proportions; business circles demanded a balanced budget and lower unemployment benefits. When MacDonald proposed to accede to some of these demands, the trade unions sharply opposed him. He then split the Labour government and formed a national coalition with the Conservatives and the Liberals. For the remainder of the 1930s the Labour Party was out of power.

Italian crises of 1919–22

Italy. In the Italian elections of 1919, the Socialists won 2,000,000 votes out of a total of 5,500,000. Italy seemed on the verge of revolution; large-scale strikes, mass demonstrations, factory occupations, and spontaneous expropriations of landed estates spread throughout the country. In August 1920 a revolutionary situation developed in the industrial north after a breakdown in wage negotiations; 500,000 workers occupied the factories, kept production going, and prepared for armed resistance. The far left called for an extension of the strike, but a divided Socialist leadership hesitated. The discouraged workers retreated. Mussolini's Blackshirts began breaking up working-class meetings. In 1921 the right-wing Socialists proposed that the party form a coalition government with the Liberals, but the left vetoed the idea. Mussolini's terror squads made further inroads in the large industrial centres. A general strike called by the trade unions proved a dismal failure. Soon afterward Mussolini made his March on Rome (October 1922) and was installed as premier. By 1926 parliamentary government had completely ended in Italy. The Socialists were driven underground.

France. None of the French governments from the end of World War I until the middle 1930s included Socialists. Although the Socialist Party was in fact deeply committed to gradualism, it still clung to its prewar policy of not participating in "bourgeois" governments. Only in the mid-1930s, when militant right-wing groups threatened the Third Republic, did the Socialists change their policy. In June 1936 a government took office representing a Popular Front, ranging from the Communists on the left to Radical Socialists in the centre and headed by the Socialist leader Léon Blum. The Communists had at last abandoned their doctrine of "social Fascism" and were now willing to enter coalitions with other parties of the centre and left.

The victory of the Popular Front in June 1936 was accompanied by sit-down strikes in the factories; these helped push the government, headed by Léon Blum, in a radical direction. Collective bargaining rights, never recognized before by French employers, were now protected by law; social security and general working conditions were significantly improved; the 40-hour week was made mandatory. The Blum government attempted to institute a French version of the U.S. New Deal. But after the initial enthusiasm had waned, French employers took courage and pressed the government to return to traditional fiscal and budgetary policies. When in June 1937 his middle-of-the-road partners in the coalition refused his demands for emergency fiscal powers, Blum resigned. The Socialists participated in the next government headed by a Radical Socialist, and Léon Blum later formed another Popular Front government that held office for about a month in 1938. When France went to war against Germany in 1939 the Communist Party, which opposed the war, was banned. After France's collapse in 1940, the Socialist Party was dissolved by the Vichy government.

Sweden. Only in Sweden were Socialists successful in their governmental policies. A Swedish Labour government was formed for the first time in 1932. Unlike the other European Socialist parties, the Swedes broke with orthodox budgetary and financial policies and stressed large-scale intervention by the government in the planning of economic affairs. Extensive public works, financed by borrowing from idle capital resources, helped to reduce unemployment and stimulated the economy; public investment was used methodically to offset the effects of reduced private spending. Unemployment, which had reached 164,000 in 1933, was eliminated by 1938 through a policy of steady economic expansion. The Swedish innovations helped lead the way to the economic policies practiced by almost all Western countries after World War II.

SOCIALISM AFTER WORLD WAR II

New concepts of socialism

The worldwide spread of "socialist" parties. Orthodox Marxists had always assumed that socialism would emerge first in the industrial countries of the world. But a new kind of "socialism" spread rapidly in agrarian societies and backward countries after World War II. In many of these countries Marxism became, despite the intention of its founders, the ideology of industrialization. In the struggle against colonialism the liberation movements, especially the intellectuals and semi-intellectuals who led them, adopted what they conceived to be socialist ideas. It seemed to them that meaningful national independence could be attained only through state direction of the economy. Rapid economic growth, they believed, could be fostered only by restricting consumption and channelling national resources into the building up of productive facilities. In one degree or another the new countries took the Soviet Union as their model for rapid industrialization. All manner of regimes, from totalitarian one-party states to military dictatorships, proclaimed that they were socialist. Only in India and a very few other countries did the ruling party retain the traditional Western socialist vision of social justice, equality, and democracy.

In the meantime, ironically, the Socialists of western Europe were giving up their Marxist views and turning toward the welfare state. During World War II almost all of the Socialist parties had joined governments of national unity. Afterward they sought to become popular parties following the parliamentary road to power and ready to participate in coalition governments with Liberal or Christian Democratic partners. Surrendering the idea that only full state ownership would bring the good society, they aimed at a mixed economy in which public control and a certain amount of planning would bring social benefits for all. This was, in essence, the idea of "the inevitability of gradualism" that the English Fabians and the German revisionists had preached around the turn of the century.

The transformation of western European socialism. *West Germany.* The changed orientation of the postwar German Social Democratic Party was expressed in its Frankfurt declaration of 1951, which made no mention of the class struggle and other traditional Marxist doctrine, stating instead that the party "aims to put economic power in the hands of the people as a whole and to create a community in which free men work together as equals." It advocated public control of the economy but rejected comprehensive state ownership. It accepted planning, but stressed that democratic socialist planning had nothing in common with the Communist and totalitarian kind.

From socialism to welfare

A few years later, in its program of principles adopted in Bad Godesberg in 1959, the party shed the last remnants of Marxism. The name of Marx and the words "class" and "class struggle" are not to be found in that program, which even advocates private property in the means of production. It rejects overall central planning and endorses the idea of a competitive free market. The party stands for "as much competition as possible—as much planning as necessary." A "mixed economy" is seen as the ideal. The party no longer claims to possess a universally valid doctrine, but stands for a pluralistic society in which no party seeks to impose its particular philosophy on society as a whole. Thus to all intents and purposes the Social Democratic Party of Germany, which in 1969 formed a government under the leadership of Willy Brandt, has become a reformist party striving for an extension of the welfare state.

British Labour. The British Labour Party was never committed to Marxism and hence found it easier to adjust to the political realities of the postwar world. In 1945 it won a majority in Parliament for the first time. The government of Prime Minister Clement Attlee, during its six years of power, laid the foundations of the British welfare state. A number of basic industries such as coal, railways, road transport, and steel were nationalized. A comprehensive system of nationalized medical care was established. Social services were extended. Full employment was maintained. Although Labour was voted out of office in 1951, its main achievements remained. The steel industry again reverted to private control, but the Conservatives made no effort to undo the other features of the welfare state.

Hugh Gaitskell, who succeeded Attlee in the leadership of the party, wanted to revamp its program by eliminating earlier pledges that the party seek large-scale nationalization of industry. He was not successful, but in practice the party became married to a reformist course aiming at the extension of the welfare state and of pragmatic planning. When the party returned to power in 1965, its leader, Harold Wilson, prime minister until the elections of 1970, pursued a cautiously reformist policy. Harassed by economic difficulties and forced to pay more attention to the balance of payments than to internal reforms, the Labour government made few policy decisions of a distinctively socialist character.

Trends in French and Italian socialism. The French Socialist Party, reconstituted after World War II, participated in leading positions in the first few postwar French governments. It supported nationalization of parts of French industry, especially public utilities, mining, and much of banking and insurance, as well as wide-ranging measures of public control over the economy and structural reforms in the field of social security. But the party had lost much of its prewar support among the workers to the Communists. The Socialists increasingly became a party of civil servants, middle-class professionals, and other white-collar employees. While they made no attempt to recast their program as the German Social Democrats did, their actual orientation was equally moderate. When they did at length achieve power under the leadership of François Mitterrand in 1981 they undertook to nationalize a number of industrial and financial concerns, but the exigencies of worldwide recession and pressures on the franc kept them to a notably moderate course.

Socialist government in France

The Italian Socialist movement split into a number of parties. The largest organization, the Italian Socialist Party under Pietro Nenni, attempted to revive the pre-Mussolini left-wing socialist tradition. It contended that the interests of the working class could best be served by cooperation with the Communists. Of all the Socialist parties of Europe it stood closest to the prewar Marxist tradition of class conflict and "orthodox" Marxism. After the Hungarian revolution of 1956, however, the party increasingly drew away from collaboration with the Communists, finally entering a centre-left coalition government with the Christian Democrats in 1963. Since then it has in practice become indistinguishable from other western European Socialist parties.

The second largest Socialist party of Italy during the post-World War II period has been the Democratic Socialist Party formed under the leadership of Giuseppe Saragat. It was committed to moderate social reforms, and participated in almost every Italian coalition government after 1947. In 1966 the two major Socialist parties merged, but they separated again in the late 1960s.

The welfare state. All the western European Socialist parties have become committed to the welfare state, though they vary in the extent to which they have formally abandoned their Marxist orientation. Some of their theoreticians still cling to the hope that socialism will eventually go beyond the welfare state toward a society in which class distinctions will have been erased and wealth will be more equitably distributed. But while this may be their dream, it no longer informs their political actions.

African socialism. Socialist ideas were carried to North Africa mainly by French-educated African intellectuals; in addition, many French settlers, especially schoolteachers and civil servants, were Socialists or Communists. The various national liberation movements, especially in Tunisia and Algeria, linked the struggle against colonial domination with socialist ideas. When Algeria became independent its first leader, Ahmed Ben Bella, surrounded himself with French advisers from various Marxist groups.

Collectivization of agriculture and self-management in industry stood high on the agenda of the Algerian national government. When these programs failed, Ben Bella was replaced by Col. Houari Boumedienne, who was pledged to continue "Algerian socialism" but settled in fact for an economy based on state-directed enterprises and private landholdings. The country was, in fact, run by a military dictatorship.

Algerian socialism

In Tunisia a one-party regime was installed after the liberation of 1956; under its leader, Habib Bourguiba, it proceeded to nationalize the major enterprises. The ruling party, the Destour Socialist Party, permitted no rival political organization; it was committed to modernization through planned development of the economy.

Elsewhere in Africa a variety of "African socialisms" sprang up in the 1950s and 1960s. Pres. Léopold Sédar Senghor of Senegal advocated a socialist "humanism" based only partly on Marx. Pres. Sékou Touré of Guinea sought to combine Marxist-Leninist ideas with the communal values of precolonial Africa—to "Africanize" Marxism. Pres. Kwame Nkrumah of Ghana proclaimed "consciencism" as the basis for his regime and stated that "only totalitarian measures can preserve liberty"; he was overthrown in 1966.

In Kenya, Tanzania, and other African countries, the ruling elites proclaimed their adherence to one or another version of "African socialism" while in fact being committed above all to rapid industrialization and modernization. Many African socialist writers stressed the need to build their socialism upon African traditions such as communal land ownership, the egalitarian practices of some tribal societies, and the network of reciprocities and obligations that once existed in tribal societies.

Throughout Africa the commitment to socialism was hardly more than lip-service to an ideal. The pressing need was to move from a subsistence to a market economy, to industrialize, and to organize health services, education, housing, and public administration. Autonomous institutions in which men might strive for concerted political and social goals independent of government control scarcely existed in Africa. Hence the prospect for democratic, as distinct from imposed, socialism was remote.

Arab socialism. The "socialist" movements of the Middle East have been led by European-educated intellectuals belonging to a new middle class of civil servants, army officers, and schoolteachers. Trying to appeal to the Arab people as a whole, and without distinction as to class, they have stood for modernization and for the brotherhood of all Arabs.

The major socialist movement has been the Arab Socialist Party, usually called the Baʿth Party. Founded in Syria, it has rejected tribal or regional loyalties. Baʿth factions have held power in Iraq and Syria but have not promoted specifically socialist measures or made concrete reforms.

When Gamal Abdel Nasser came to power in Egypt in 1952, his group of young army officers of lower middle class origin had little if any interest in socialism. Nasser was led to socialist ideas in his struggle against the domination of foreign business. By the middle of the 1960s Egypt had nationalized all large industrial and financial enterprises, whether domestic or foreign; it had expropriated large-scale landowners; and it had placed all important sectors of the economy under the control of state planners. But the structure of power remained that of a military dictatorship.

Failure of socialism to take root in Asia

Asian socialism. Representatives of a dozen Asian Socialist parties met in Rangoon in January 1953 for the first Asian Socialist Conference. Some of the delegates had international reputations. The governments of several Asian countries—India, Burma, Ceylon, Indonesia, and Singapore—called themselves Socialist. Yet the Socialist parties soon thereafter lost even the semblance of power and influence. In India, where several Socialist organizations competed, the ruling Congress Party was in fact a national party striving to unite within its ranks many divergent political and social tendencies. The Burmese Socialist Party was for many years part of a coalition that ruled the country, but it was outlawed in 1962 when Gen. Ne Win seized power. The Indonesian Socialist Party was abolished by President Sukarno in 1960. Except in Singapore, the postwar Socialist parties in Southeast Asia played no major role in the 1960s.

As the influence of the European-style Socialist parties waned, a variety of authoritarian regimes arose speaking in socialist accents. Indonesia's President Sukarno proclaimed Resopim (Revolution, Indonesian Socialism, and National Guidance) his official ideology. The Burmese military dictatorship proclaimed Burma a socialist state. North Vietnam (and later all of Vietnam) was ruled by a Communist party; in the rest of Indochina, revolutionary movements inspired by Communist ideology battled traditionalist forces supported by the United States. In China, the Communist government of the People's Republic has been in power since 1949.

The Socialist parties of Southeast Asia, after playing a brief role in the struggle for independence, failed to take root in national politics. They were led by European-educated intellectuals, who attempted to emulate European models and were committed to the idea of a democratic road to socialism. But the countries of Southeast Asia were not prepared to follow them; they turned instead to authoritarian regimes that pursued industrial development. Only in Singapore and India were attempts made to combine democracy and socialist planning. In most of Asia, as in Africa, "socialism" has become the ideology of new elites seeking modernization and rapid industrialization.

Japan. Only in Japan, by far the most developed of Asian countries, have traditional socialist organizations become firmly established. The first Socialist Party of Japan, formed in 1901, was soon dissolved and forced to go underground. During and after World War I socialist organizations again sprang up. In 1936 the Social Mass Party elected 18 members to the Parliament and received more than half a million votes. After World War II socialist organizations that had been suppressed since 1940 appeared again. In 1946 the Socialist Party won over 90 seats to become the third strongest party. A year later it gained the largest number of seats in Parliament, and its leader Katayama Tetsu became prime minister in a coalition government. In October 1948, however, the conservatives took office. The Socialists were deeply split between gradualists and revolutionaries. The left wing tended to be sharply anti-American and leaned toward the Soviet Union; the right wing favoured a gradual relaxation of close military and political ties with the United States. The two wings broke apart in the 1950s to form the (left) Socialist Party of Japan and the (right) Democratic Socialist Party. Together they controlled about one-third of the seats in Parliament, but they seemed condemned to a permanent minority status. Japanese socialism still awaits a transformation similar to that undergone by western European Socialism in the postwar era.

Success in Australia, New Zealand, Canada

Other areas and countries. *Australia.* Socialism has deep roots in the British Commonwealth countries of Australia, New Zealand, and Canada. The Australian Labor Party was formed in 1901, when the Australian Commonwealth came into existence. Only three years later its leader, J.C. Watson, became the world's first Labor prime minister. In 1908, and again from May 1910 to June 1913, Labor headed the government. In subsequent years the Labor Party has been frequently in office.

New Zealand. A loose Liberal-Labour alliance dominated New Zealand politics between 1893 and 1906, but the New Zealand Labour Party, as a social-democratic party committed to the socialization of the means of production, did not emerge until 1913. It grew steadily, coming to power in 1935 for the first of several periods of varying duration.

The Australian and New Zealand labour movements have been committed to a gradualist and reformist course since their inception. They are strongly tied to the unions, and though in principle pledged to a Socialist program they are in fact mainly concerned with using governmental control as a means of dealing with immediate problems and expanding social services. The various social security acts that they introduced helped to make Australia and New Zealand modern welfare states, and relatively equalitarian societies, even before World War II.

Canada. Canadian socialism had a slower beginning than its Australian and New Zealand counterparts. Prior to World War I the Canadian Socialist movement was split between two parties, neither of which managed to win seats in the federal Parliament. During the 1920s various Socialist and Labour parties flourished in different parts of Canada, but only rarely did they send a representative to the federal Parliament. Only with the organization of the Co-operative Commonwealth Federation (CCF) in 1932 did the Socialist movement begin to achieve national importance. Basing its campaigns on the need for "social and economic planning on a bold and comprehensive scale," it gained support in most provincial elections and in June 1944 was able to form a government in the province of Saskatchewan, where it remained in power for 20 years. In 1961 progressive union leaders of the labour congress met with the CCF leadership and formed the

New Democratic Party. Whereas the CCF had been largely agrarian in character, the new party had a following in the industrialized parts of the country. Advocating a planned economy, it stood for increased social security, government employment guarantees, large-scale construction of low-rent housing, and the like. Its policies were similar to those of postwar western European socialism.

Latin America. The historical roots of Latin American socialism are fairly old. Several branches of the First International were established in Argentina in the early 1870s. In Chile and Argentina, and to a lesser extent in other Latin American countries, socialists at times played leading roles, but they were hampered by a variety of splits and by the fact that their following consisted mainly of immigrant industrial workers. They failed to make an impact in rural areas. In Chile, however, they participated in various coalition and popular front governments in the 1920s, 1930s, and 1940s. In the elections of 1958, Chilean socialists supported the Popular Action Front (FRAP) candidate, Salvador Allende. He was narrowly defeated, and defeated again in 1964. In 1970, however, he won by a narrow plurality in a three-way presidential election and became head of a government supported by a popular front ranging from Communists to democratic reformers. His government, pledged to the nationalization of foreign-owned industry and to the planned reconstruction of the country, met with increasing economic turmoil and opposition from the middle class; Allende's expulsion by military coup in 1973 left the future of socialism in Chile uncertain. (L.A.C.)

Chilean socialism

Communism

The word communism, a term of ancient origin, originally meant a system of society in which property was owned by the community and all citizens shared in the enjoyment of the common wealth, more or less according to their need. Many small communist communities have existed at one time or another, most of them on a religious basis, generally under the inspiration of a literal interpretation of Scripture. The "utopian" socialists of the 19th century also founded communities, though they replaced the religious emphasis with a rational and philanthropic idealism. Best known among them were Robert Owen, who founded New Harmony in Indiana (1825), and Charles Fourier, whose disciples organized other settlements in the United States such as Brook Farm (1841–47). In 1848 the word communism acquired a new meaning when it was used as identical with socialism by Karl Marx and Friedrich Engels in their famous *Communist Manifesto.* They, and later their followers, used the term to mean a late stage of socialism in which goods would become so abundant that they would be distributed on the basis of need rather than of endeavour. The Bolshevik wing of the Russian Social-Democratic Workers' Party, which took power in Russia in 1917, adopted the name All-Russian Communist Party in 1918, and some of its allied parties in other countries also adopted the term Communist. Consequently the Soviet state and other states governed by Soviet-type parties are commonly referred to as "Communist" and their official doctrines are called "Communism," although in none of these countries has a communist society yet been established. The word communism is also applied to the doctrines of Communist parties operating within states where they are not in power. (For the ideological basis of Communism, see MARXISM.)

THE ORIGINS OF SOVIET COMMUNISM

Communism as it had evolved by 1917 was an amalgam of 19th-century European Marxism, indigenous Russian revolutionary tradition, and the organizational and revolutionary ideas of the Bolshevik leader Lenin. Marxism held that history was propelled by class struggles. Social classes were determined by their relationship to the means of production; feudal society, with its lords and vassals, had been succeeded in western Europe by bourgeois society with its capitalists and workers. But bourgeois society, according to Marxism, contained within itself the seeds of its own destruction: the number of capitalists would diminish, while the ranks of the impoverished proletariat would grow until finally there would be a breakdown and a Socialist revolution in which the overwhelming majority, the proletariat, would dispossess the small minority of capitalist exploiters.

Plekhanov's Marxism

Marxism had been known and studied in Russia for at least 30 years before Lenin took it up at the end of the 19th century. The first intellectual leader of the Russian Marxists was G.V. Plekhanov. Implicit in the teachings of Plekhanov was an acceptance of the fact that Russia had a long way to go before it would reach the stage at which a proletarian revolution could occur, and a preliminary stage would inevitably be a bourgeois democratic regime that would replace the autocratic system of Tsarism.

Plekhanov, like most of the early Russian Marxist leaders, had been reared in the traditional Russian revolutionary movement broadly known as Populism, a basic tenet of which was that the social revolution must be the work of the people themselves, and the task of the revolutionaries was only to prepare them for it. But there were more impatient elements within the movement, and it was under their influence that a group called "People's Will" broke off from the Populist organization "Land and Freedom" in 1879. Both groups were characterized by strict discipline and highly conspiratorial organization; "People's Will," however, refused to share the Populist aversion to political action, and in 1881 some of its members succeeded in assassinating Tsar Alexander II.

Lenin and Russian Populism. During the period of reaction and repression that followed, revolutionary activity virtually came to an end. By the time Lenin emerged into revolutionary life in Kazan at the age of 17, small revolutionary circles were beginning to form again. Lenin was a revolutionary in the Russian tradition for some time before he was converted to Marxism (through the study of the works of Marx) before he was yet 19. From the doctrines of the Populists, notably P.N. Tkachev, he drew the idea of a strictly disciplined, conspiratorial organization of full-time revolutionaries who would work among important sections of the population to win support for the seizure of power when the moment was ripe; this revolutionary organization would take over the state and use it to introduce Socialism. Lenin added two Marxist elements that were totally absent in Populist theory: the notion of the class struggle and the acceptance of the need for Russia to pass through a stage of capitalism.

Lenin's concept of the revolutionary party

Lenin's most distinctive contributions to Communist theory as formulated in *What Is To Be Done?* (1902) and the articles that preceded it were first, that the workers have no revolutionary consciousness and that their spontaneous actions will lead only to "trade union" demands and not to revolution; second, the corollary that revolutionary consciousness must be brought to them from outside by their intellectual leaders; and third, the conviction that the party must consist of full-time, disciplined, centrally directed professionals, capable of acting as one man.

Lenin's tactics led in 1903 to a split in the Russian Social-Democratic Workers' Party. With his left-wing faction, called the Bolsheviks, he strove to build a disciplined party and to outwit and discredit his Social-Democratic opponents. After the collapse of tsarism in February 1917, he pursued a policy of radical opposition to the Socialists and Liberals who had come to power in the provisional government and he eventually succeeded in seizing power in October 1917. Thereafter he eliminated both the opposition of other parties and his critics among the Bolsheviks, so that by the 10th party congress in March 1921 the Bolsheviks (or Communists) had become a monolithic, disciplined party controlling all aspects of Russian life. It was this machine that Stalin inherited when he became general secretary of the party in 1922.

THE THIRD INTERNATIONAL

The victory of the Bolsheviks in Russia gave a new impetus to the more extreme left wings of the Socialist parties in Europe. Lenin's relations with the European Socialist parties had been hostile even before World War I. During the war he had endeavoured to assert his influence over the dissident left wings of the Socialist parties of the belligerent powers, and at two conferences in Switzerland, in 1915 and in 1916, he had rallied these dissident groups to a policy of radical opposition to the war efforts of their governments and to an effort to turn the war into a civil war. He had already decided by 1914 that, after the war, a Third International must be formed to take the place of the Second International of Socialist parties, which had failed to oppose the war despite its strong antiwar tradition. By 1919, when the new Soviet regime in Russia was fighting for its survival, the intervention on the anti-Soviet side by Britain, France, and the U.S. was a powerful and practical argument to be used by Soviet Russia in its appeals for revolution in capitalist countries. It early became clear the Third International would reflect the influence of Soviet Russia and that it was likely to become subordinate to Soviet aims and needs.

Communism's emergence as an international movement

Lenin's 21 conditions. The Third International, or Comintern, had its first congress in 1919. This gathering of a very few parties in Moscow was more symbolic than real; the main structure of the new International was not hammered out until the second congress in July 1920, also in Moscow. Hopes of world revolution ran high; the prestige of the new Soviet state was in the ascendant, and the resolutions adopted at this congress reflected in the fullest possible way Lenin's idea of what a Communist party should be. It was to be the "main instrument for the liberation of the working class," highly centralized and disciplined according to the formula of "democratic centralism" on which the Bolshevik Party had been founded. Twenty-one conditions were laid down by the congress as prerequisites for parties affiliating with the Comintern. These conditions were designed to ensure a complete break with the older Social Democratic parties from which the Communist parties were splitting off. The new parties were required to adopt the name Communist in their title, to urge open and persistent warfare against reformist Social Democracy and the Second International, to maintain a centralized and disciplined party press, to conduct periodic purges of their ranks, and to carry on continuous and systematic propaganda in the army and among the workers and peasants. Each constituent party was to support in every possible way the struggle of "every Soviet republic" against counterrevolution. Decisions of the Comintern and of its executive committee were to be binding on all members, and the breach of any of these conditions was to be ground for expelling individual members from their parties—a provision that in future years was to be interpreted very broadly.

The New Economic Policy. The prestige of Soviet Russia, the rigid discipline imposed by the 21 conditions, and certain other factors ensured the predominance of Russian control and Russian interests over the Comintern. Though the predominance increased during Stalin's time, it was clearly evident while Lenin was still alive. At the third world congress in June and July 1921, the Comintern was confronted by Lenin with his New Economic Policy—a program encouraging small private enterprise, which several months earlier he had put into effect inside Russia. Lenin wanted a temporary halt to the revolutionary upsurge in Europe to give him time to develop stable trade relations with capitalist countries, to whom the Soviet state was preparing to grant trading and industrial concessions. Comintern members were required to support this policy, and the expulsion of the German Communist leader Paul Levi after the failure of a Communist uprising in Germany in March 1921 showed how determined the leaders of the Comintern were to put down inconvenient left-wing "adventures." It was with the requirements of the New Economic Policy in mind that the Comintern executive committee in December 1921 launched the turnaround policy of the United Front and of trade union unity. This policy of rapprochement with Socialists and liberals was likewise designed to gain support for Lenin's policy of consolidation at home by appealing to a broader spectrum of opinion in the capitalist countries.

STALINISM

From Lenin to Stalin

Socialism in one country. Lenin's successor, Joseph Stalin, always claimed to be his faithful follower, and this was to some extent true. Stalin's doctrine that Socialism could be constructed in one country, the Soviet Union, without waiting for revolution to occur in the main capitalist countries (a position he had developed as an integral part of his struggle against Trotsky) was not far removed from the line pursued by Lenin in 1921 when he introduced the New Economic Policy. Both Lenin and Stalin accepted the primary importance of the survival and strengthening of the Soviet state as the main bastion of the future world revolution; both accepted the need for a period of coexistence and trade with the capitalist countries as a means of strengthening socialism in Soviet Russia. Nor did Stalin's later policy of industrialization and collectivization, in theory at least, represent a departure from Lenin's doctrine. Industrialization was central to Lenin's plans, though he did not live to put them into practice. Stalin's view, however, that the construction of socialism led inevitably to an intensification of the class struggle, which in turn required a policy of internal repression and terror, is nowhere to be found in Lenin's writings. On the contrary, Lenin repeatedly emphasized in 1922 and 1923 the necessity of bringing about a reconciliation of the classes and especially of the peasants and workers.

Stalin's internal policy was to have wide repercussions in the Comintern and on Communism generally. From 1924 until 1928 his first concern was to defeat his main rival, Trotsky, and this seems to have been one of the main factors determining his policy at this time. As against the more internationalist and doctrinaire Trotsky, Stalin pursued "socialism in one country" and continued to implement Lenin's New Economic Policy with its limited freedom for business enterprise and peasant individualism. In this he could still claim to be following Lenin's wishes. But Stalin also worked with great skill to ensure his control over the party. By 1927 when Trotsky was expelled from the party, Stalin already controlled both the network of party officials (the *apparat*) and the delegates to congresses and conferences. Debate had been replaced by ritualized unanimity; dissent was permitted only when it served the purposes of the leadership.

When Trotsky was exiled from the country in 1929, he became the focal point for opposition to Stalin among dissident Communists all over the world, although he was to be more a symbol than an active political force. Having defeated Trotsky and his allies, Stalin next switched policies, abandoning the New Economic Policy in favour of rapid industrialization along with the collectivization of agriculture. The collectivization policy ultimately produced a famine, costing the lives of millions of peasants. The reversal of the New Economic Policy and of Lenin's policy necessarily involved eliminating from the political scene Stalin's former allies, headed by Nikolay Bukharin, who wanted to go slower with industrialization and to cultivate support among the peasants. The protracted conflict, first with Trotsky and his ally G.Y. Zinovyev and then with Bukharin, was reflected in the Comintern and in the world Communist movement, which became increasingly subordinated to Stalin's policy concerns inside the Soviet Union.

Stalin and the Comintern. The regimentation of the Comintern and of the parties represented in it began at the fifth world congress in June 1924, immediately after Lenin's death. The elimination of Trotsky and his supporters within the Soviet party was followed by widespread expulsions of the "left" from the other world parties. The control of the Soviet-dominated Comintern apparatus was increasingly asserted over the tightly disciplined governing bodies of the foreign parties, which in turn ruled over their members with the instrument of the purge. Ideologically, this procedure was carried out at first under the screen of the United Front, which called for cooperation with Social Democrats and other moderate leftists. At the sixth world congress in 1928, however, a further switch in policy was dictated by Stalin's internal conflict: the United Front tactic was abandoned, and the Social Democrats now became enemies along with Fascists. The sixth congress also declared the main duty of the international working-class movement to be the support of the U.S.S.R. by every means. The united front tactic was revived in 1935 at the seventh (and last) world congress of the Comintern under the name of the Popular Front, calling for united action by Communists and Socialists together against Fascism.

The Nazi–Soviet pact

Comintern policy changed again in August 1939 when the Soviet Union and Germany concluded a 10-year treaty of nonaggression. This had the effect of freeing Hitler to fight a war against Britain and France. Anti-Fascism was now jettisoned, and the Communist parties were required, up to the moment when Germany invaded the Soviet Union on June 22, 1941, to denounce the allied war against Hitler and to recognize Nazism as "the lesser evil" in comparison with Western imperialism. The Soviet alliance with Germany is usually seen as proof that Stalin was primarily concerned with what he considered to be the interests of the Soviet Union. A secret protocol annexed to the treaty assigned the Baltic states (Latvia, Lithuania, and Estonia), about half of Poland, and Bessarabia to the Soviet sphere of influence. The evidence suggests that Stalin considered the deal with Hitler to be based on mutual interests; the German invasion in 1941 took him by surprise. After the defeat of Hitler, Soviet territorial demands were again advanced.

Stalin's method of rule. The Communist parties of the world were also called on to adopt official Soviet justifications for Stalin's internal purges, which involved the extermination of a large proportion of the Soviet party membership, including most of the leading cadres. The subservience of some Communist parties to official assertions made by the Soviet authorities sometimes earned them the reputation of being little more than agents of the Soviet Union inside their own countries, though this did not necessarily diminish their influence or importance in several countries of Europe or in the United States. They found much support among sympathizers with Marxism, who were prepared to overlook Soviet realities in the service of their ideals or of what they considered to be the historical destiny of mankind—in which they saw Stalinism as merely a transitory stage. The Communists and their parties and their contacts provided a valuable recruiting ground for intelligence agents of all kinds prepared to act against their own countries in the interests of Soviet Russia. The effects of Stalin's internal policy on the Communist parties outside the Soviet Union are of vital importance in understanding the attitude adopted by these parties after 1956, when much of Stalin's policy was officially repudiated.

Stalin's method of rule came, by imitation, to be the standard in all other parties. It hinged primarily upon the dominance of his own personality. He ruled over the country in large measure not through the party, as Lenin had, but through personal agents (like Lavrenty Beria, Andrey Vyshinsky, or Georgy Malenkov) and also through the security police (NKVD). The party as an institution declined under Stalin, and between 1934 and 1952 there was only one party congress, in 1939. The general secretaries of the Communist parties abroad imitated Stalin, and strict hierarchical subordination became the way of party life.

GROWTH OF COMMUNISM DURING AND AFTER WORLD WAR II

The wartime prestige of the U.S.S.R.

The undeclared assault by Hitler on the Soviet Union provoked a wave of sympathy for that country among both the open and secret enemies of Hitler in Europe. The Soviet pact with Hitler, and even the manifest blemishes of Stalin's regime, were forgotten: sympathy with the newly emerged force of resistance to the Nazi scourge far outweighed past memories. Many, it is true, expected the immediate defeat of the Soviet Union. As time went on, however, and the Soviet struggle continued with enormous sacrifice of life and with courage and skill that none could help but applaud, admiration for Soviet military achievements grew even among those who had been most critical and apprehensive of the Soviet political role before the war. The Communists of other countries shared in the prestige won by Soviet military prowess. This was particularly the case in occupied France and Italy where the underground Communist parties played a vital role in the resistance movements. In Yugoslavia, too, the Communist partisan movement led by Tito (Josip Broz) outstripped the nationalist guerrillas in effectiveness and won the material support of Britain.

Russian nationalism. The policy pursued by Stalin accentuated the nationalist side of the war and attempted in every way to play down the Communist element. At home, tsarist history and the rituals of the Eastern Orthodox Church were invoked in efforts to raise patriotic sentiments to the highest possible pitch. Abroad, Communist aims and ideals were replaced by anti-Nazi, liberal-democratic slogans. The dissolution of the Comintern in 1943 was in line with this policy. It had long ceased to be necessary as an instrument of Soviet control over the foreign Communist parties, which was carried on through other channels; but the publicizing of its dissolution added force to the growing persuasion abroad that the Soviet Union had left its revolutionary past behind it and was now a great power with traditional nationalist and security aims. Stalin himself emphasized that the dissolution of the Comintern would "put an end to the lies spread by Hitler that the Soviet Union wished to Bolshevize other countries" and that Communist parties "followed foreign directives." Still another factor promoting the influence of Communism during World War II was the enhanced prestige of Stalin himself and the extent to which his personality influenced the allied leaders Winston Churchill and Franklin D. Roosevelt.

Stalin and eastern Europe. His growing military and political prestige in turn influenced Stalin's policy towards his allies and determined the future course of Communism after victory was won in 1945. Two main lines of Soviet policy can be discerned in the wartime conferences at Tehrān, Yalta, and elsewhere: first, a determination by the Soviet Union that friendly political regimes should be established in the countries on Russia's borders, and second, that the Soviet Union's hard-won status as a great power should be fully recognized in the postwar settlements. These demands were not in themselves unreasonable, considering the enormous price that the Soviet people had paid for victory. In pursuing the creation of a solid Soviet-dominated bloc of Communist states in east-central Europe, Stalin was able to take advantage of the presence of a victorious Soviet army in Poland, Bulgaria, Romania, Hungary, and East Germany. The cases of Yugoslavia and Albania were different, but the regimes that emerged in all these countries were broadly similar forms of Communist party domination based on the Soviet model, even though the ways in which the Communists achieved power varied.

The expansion of Soviet influence

Broadly speaking, three phases could be distinguished. In the first phase there was a genuine coalition of Communist and Socialist parties. This lasted until the spring of 1945 in Romania and Bulgaria, until the spring of 1947 in Hungary, and until February 1948 in Czechoslovakia. Yugoslavia, Albania, Poland, and East Germany never knew this phase: the former two started as "monolithic," while the latter two began their postwar history in the second phase, an alleged coalition in which the Socialist parties were nominally independent and had some share in power but in which their leaders and policies were largely determined by the Communists. In the third phase, the "monolithic" phase, the nominally independent Socialist parties were required to fuse with the Communists, political opposition was largely suppressed, and Socialist leaders went into exile or were dealt with by staged treason trials. In Poland, Bulgaria, and Romania the third phase began in the autumn of 1947; in Hungary, in the spring of 1948. In East Germany the third phase was complete by 1949.

In his policy toward the countries which were destined to form the Soviet bloc, Stalin was aided in part by the inability or unwillingness of the Western allied powers to take steps during the first or second phases described above to prevent the beginning of the third phase and in part by the skillful infiltration of local Communists into key positions. The peasant and Socialist parties, which had substantial support in their countries, were attacked in various ways and demolished as independent political bodies.

Yugoslavia was an exception. There the Communists under the leadership of Tito enjoyed a considerable measure of mass support because of their wartime role as partisan fighters. The People's Democracy they instituted in Yugoslavia was for some years little different in character from that of other Communist-party-dominated states of eastern Europe. An attempt to set up a People's Democracy in Greece failed after three years of civil war, in which the Greek Communists were supported by Yugoslav aid.

The failure in western Europe

In the countries of Europe outside the Soviet bloc, Communist parties proved unable to exploit the prestige that they had acquired during the war. Both in France and in Italy they enjoyed considerable support: in the parliamentary election of 1945 in France the Communists received 26 percent of the vote, and in the general elections to the Constituent Assembly in Italy in June 1946 they received 19 percent. Both parties, however, failed to achieve real national power in the postwar period; their role was confined to fomenting strikes and disorder in the interests of Soviet policy. The detailed story of the Italian and French Communist parties during the period 1945 to 1949 is complex, but, broadly speaking, their attempts at insurrection foundered against the facts of the power of the army and the police and a lack of revolutionary zeal among their worker supporters. On the other hand, their attempts to win power by parliamentary means were frustrated by the distrust that the Socialists felt for them as colleagues in Parliament or in government and by their own evident lack of interest in a viable parliamentary system.

Communism's growth in Asia. Powerful Communist parties emerged after the war in various parts of Asia, in many cases largely as a result of the resistance of the Western powers to growing nationalist movements. Communist-led insurrections, allegedly coordinated by Moscow, broke out in the summer of 1948 in Burma, Malaya, and Indonesia. In Indochina, after the surrender of Japan, the Communists under Ho Chi Minh seized power in the three northern provinces of the country. French colonial policy helped drive the nationalists into the arms of Ho Chi Minh, and by the end of 1946 a guerrilla war had broken out in the country that was to last for nearly three decades before the Communist victory of 1975. In Japan democratic legislation imposed by the United States after its victory permitted the Communists to operate legally. In the succeeding few years they made little progress toward governmental power but won considerable gains in the trade unions and an important measure of influence among university students. In India the Communist Party supported the British war effort after June 1941 and gained ground as a result; it switched to violent insurrection after Indian independence but abandoned this policy in 1950.

Chinese Communism

The most significant factor in the postwar history of Communism in Asia may have been the victory in 1949 of the Chinese Communist Party under the leadership of Mao Tsetung. China, rather than the Soviet Union, seemed destined to play the leading role in Asian Communism. The victory of the Chinese Communists over Chiang Kai-shek and the Kuomintang, like that of Tito's forces in Yugoslavia, owed little if anything to Soviet aid—save that the Russians had handed over to the Chinese Communists the military stores captured from the Japanese during the very short period when the U.S.S.R. was at war with Japan in 1945. Although the Chinese Communist Party had developed under the aegis of the Comintern and acknowledged the doctrinal authority of Lenin and Stalin, its experience had been very different. Its victory had been preceded by long guerrilla warfare. Mao's rise to power had, moreover, been achieved by ignoring Soviet advice as much as by following it. Stalin showed quite clearly from the outset that he intended to keep China in a position of subordination not unlike that which he had successfully marked out for most of eastern Europe—a status the Chinese Communist leaders were not likely to accept. Culturally, economically, and geographically, China was in a strong position to become the model for Communist revolution in Asia and to wrest the leadership of Asian Communism from the Soviet Union. These and other factors were to produce signs of a possible breach between China and the U.S.S.R. within less than 10 years of the proclamation of the Chinese People's Republic on October 1, 1949.

THE WORLD MOVEMENT UP TO STALIN'S DEATH

The Cold War

The wartime alliance had given rise to some hopes that Soviet-Western amity would continue. Stalin's relentless pursuit of security through the domination of neighbouring countries shattered this hope. At home Stalin returned to his prewar tactics: widespread arrests and deportations occurred in the newly incorporated or reincorporated territories of the Soviet Union; the restriction of cultural life was intensified; the straitjacket was reimposed on the party, on the peasants, and on the industrial workers. There is some evidence to suggest that at the time of his death in March 1953 Stalin was planning a new purge on the scale of the 1936–38 purges.

The struggle with the West. Soviet expansion into eastern Europe led to counteractions by the Western powers that Moscow interpreted as part of a master plan to encircle and subjugate the Soviet Union. These included the Truman Doctrine of containment of Soviet expansion proclaimed in March 1947; the offer in June of that year by United States Secretary of State George Marshall to underwrite the economic recovery of Europe; and the North Atlantic Treaty of April 1949, which established a permanent defense force for western Europe, including in its orbit West Germany. Another factor that affected Soviet policy was the monopoly of the atomic bomb enjoyed by the United States from 1945 until 1949. The Soviet Union rejected the Baruch Plan put forward by the U.S. for the international control of atomic weapons and made every effort to produce its own, succeeding in September 1949. The "Cold War" was on.

The defection of Yugoslavia. In September of 1947 a new international organization, the Communist Information Bureau (Cominform), was established. Unlike the old Third International (Comintern), the Cominform was limited in membership to the Communist parties of the Soviet-dominated countries of east-central Europe and to the French and Italian Communist parties. The aim of the Cominform was to consolidate and expand Communist rule in Europe. Plans for the establishment of Communist rule in Czechoslovakia were discussed, and the French and Italian parties were reproved for their failure to win power in their own countries.

The Cominform did not prove a success. Certainly one of its purposes was to hold Yugoslavia more securely within the Communist fold, and for this reason Belgrade was chosen as the seat of the new organization. But within a few months a quarrel broke out between the Soviet and Yugoslav parties, and when the Cominform held its second meeting in June 1948, it was for the purpose of denouncing the Yugoslav Communist Party and expelling it from the organization. The quarrel with Yugoslavia resulted largely from Tito's refusal to submit to domination by the Soviet Union; there was also some suspicion on the Soviet side, possibly well founded, that the Yugoslav party leader hoped to build up a bloc of Communist states in southeastern Europe that would not be totally dependent on the Soviet Union.

The effect of the Soviet-Yugoslav quarrel, which has never completely healed, was momentous. First, it shattered the doctrine that the Communist movement must be monolithic, since a Communist party had challenged Moscow and survived. Second, Yugoslavia, having broken with the U.S.S.R., was in a position to assume a role of considerable influence in the world, especially toward states formed in formerly colonial territories. The Yugoslavs could speak as Communists who, while opposed to the policy of the imperialist powers, were no mere agents of Soviet policy. This position carried a particularly strong appeal in India, but the impact of the Soviet quarrel with Tito was much wider.

A third effect of the Yugoslav defection was a tightening of the Soviet hold over the remaining members of the Communist bloc. In Soviet-dominated lands "Titoism" became synonymous with treason, much as "Trotskyism" had been in the '30s. Purges and public trials ensued throughout eastern Europe. In some cases, like that of Władysław Gomułka in Poland (who was left alive), or Koci Xoxe in Albania, the charge of sympathy with Yugoslavia may have been true; in others, like those of

László Rajk in Hungary or Traicho Kostov in Bulgaria, the offense may have been only an attempt to resist Soviet domination; in the trial of Rudolf Slánský in Czechoslovakia in 1952, a strong anti-Semitic element played a part. Countries of the Communist bloc were seething with anti-Soviet and nationalist feeling by the time Stalin died. Though Stalin's postwar policy was successful in extending the boundaries of Soviet military and political control well into eastern and central Europe, Communism did not win out in France or in Italy, where its chances had appeared strongest. The policy of expansionism and of intransigence founded on suspicion of the United States led to a kind of consolidation of the West against the Soviet Union. In the Far East the Korean War was probably not a success from the Communist point of view. Korea had been divided after the defeat of Japan: in the northern part a Communist government came to power in elections held in November 1946, and in the south a non-Communist government was established. Each claimed to be the legal government of the whole country. Invasion of the south by the north in June 1950 was condemned by the Security Council of the United Nations as aggression, and the Security Council approved military assistance to South Korea under a unified American command. (The absence of the Soviet representative from the Security Council prevented the U.S.S.R. from vetoing this resolution.) The long war, in which China intervened on the side of North Korea, brought heavy burdens and few, if any, advantages, and the conflict between the major powers that it involved led them in the fears of many to the verge of world war. In June 1951 the Soviet Union proposed discussions for an armistice, to which the Western powers agreed. The negotiations were protracted and did not result in an armistice until after Stalin's death in 1953.

THE BREAKUP OF THE WORLD COMMUNIST MONOLITH

Stalin's heirs

The Khrushchev era. Stalin died on March 5, 1953. For a short time, until the beginning of 1955, power was nominally divided between Georgy Malenkov, the chairman of the Council of Ministers, and Nikita Khrushchev, the first secretary of the Communist Party. Almost from the beginning, Khrushchev was the dominant of the two; his victory over his rival was only a matter of time. Malenkov, it would seem, decided quite early that the Soviet Union could not maintain its hold over the Eastern bloc without substantial economic relaxation. The difficulties that always beset the reform of an oppressive regime were soon illustrated in East Germany. Within a week of the announcement by East German leaders that "aberrations" of the past would be rectified and some of the hardships of life alleviated, there was an uprising in the streets of East Berlin; it spread to other parts of East Germany and was quelled only by the use of Soviet armed forces. The blame for this was laid on Lavrenty Beria (the Soviet security chief, shortly to be deposed and executed) and by implication on Malenkov. The new relaxed policy continued, however, in most of the Soviet-bloc countries. Economic reforms were initiated in Hungary, Czechoslovakia, and Poland, but the system of political rule remained unchanged.

Khrushchev, who by the beginning of 1955 had ousted Malenkov, had a comprehensive vision of how the Eastern bloc should be run. He was determined to find a way out of the straitjacket in which Stalin had confined Soviet life; the outcome was to have momentous consequences for Soviet dependencies abroad, which Khrushchev probably did not at the time foresee. His policy toward the Communist satellite countries may be summarized as one of cooperative integration instead of exploitation, with some degree of economic and political autonomy (under Communist Party leadership). A political and military convention between the European Communist states and the U.S.S.R. (the Warsaw Pact) was signed in May 1955. Khrushchev also sought to redesign the Council for Mutual Economic Assistance, the Communist counterpart of western Europe's Common Market, which Stalin had set up in January 1949: he tried (though with indifferent success) to transform the Council for Mutual Economic Assistance into a device for promoting the division of labour, economic specialization, and technical and financial cooperation among the countries of the bloc.

The crises of 1956. In order to demonstrate that Stalin's policy was a thing of the past, Khrushchev made substantial efforts to effect a reconciliation with Tito and the Yugoslav Communists (against the opposition of some of his colleagues, including Vyacheslav Molotov). An agreement with Yugoslavia in June 1956 recognized that "the conditions of Socialist development are different in different countries" and stated that no Socialist country should impose its views on another. This was a momentous change in policy, since it meant that a country could be described as "Socialist" without being obliged to follow all the practices adopted by the Soviet Union or every Soviet turn in foreign relations.

Khrushchev's criticism of Stalin

The reconciliation with Yugoslavia was only one of several important events that made the year 1956 a watershed in the history of Communism. In February, at the 20th congress of the Communist Party, Khrushchev delivered a speech in secret session in which he attacked the period of Stalin's rule in most forthright terms. The speech was not published within the Soviet Union, but its text was widely circulated among Communists both within and outside the Soviet Union and was published by the U.S. State Department. Its effect was enormous. Although the disclosures were neither complete nor entirely new, the fact that Khrushchev had uttered them caused a ferment in the Communist movement that was to prove irreversible. It inaugurated a period of freedom of debate and criticism that had been unknown for a quarter of a century; despite efforts both by Khrushchev and by his successors to keep criticism of the "cult of personality" (the accepted euphemism for Stalin's misdeeds) within bounds, the ferment could not be contained.

The Hungarian Revolution. In the European Communist countries, Khrushchev's disclosures opened the floodgates of pent-up criticism and resentment against the local Stalin-type leaders. In Hungary, Mátyás Rákosi was ousted as party leader in July 1956 and replaced by Ernő Gerő. But Gerő was unable to contain the rising tide of unrest and discontent, which broke out into active fighting late in October, and appealed for Soviet help. The first phase of the Hungarian Revolution ended in victory for the rebels: Imre Nagy became premier and agreed, in response to popular demands, to establish a multiparty system; on November 1 he declared Hungarian neutrality and appealed to the United Nations. On November 4 the Soviet Union, profiting from the lack of response to Nagy from the Western powers, and from the British and French involvement in action against Egypt, invaded Hungary in force and stopped the revolution. In Poland, where the ferment was also reaching dangerous intensity, the Soviet Union accepted a new party leadership headed by the more moderate Władysław Gomułka. There are believed to have been two reasons for this difference in Soviet policy. One was that in Poland the Communist Party remained in control of the situation. The other was that the invasion and subjugation of Poland would have required a military force several times that required in Hungary.

Polycentrism. Inside the Communist states, the suppression of the Hungarian Revolution had a restraining effect. There was, nevertheless, no return to the Stalinist type of domination and exploitation; a slow evolution followed toward a degree of internal autonomy, even in Hungary. The events of 1956 also had profound effects upon Communists outside the Soviet bloc. There were many resignations after the Hungarian Revolution, and those who remained in the fold began to question both Soviet leadership and the nature of a system that had made the ascendancy of Stalin possible. The most trenchant questioning came from the leader of the Italian Communist Party, Palmiro Togliatti, who concluded that the Soviet pattern could no longer be the model for all other countries and called in June 1956 for decentralization of the Communist movement, a view that became known as "polycentrism." "The whole system becomes polycentric, and . . . we cannot speak of a single guide but rather of a progress which is achieved by following paths which are often different." Although the Italian Communist Party, or segments of it, were still prepared to support the Soviet Union at times of crisis, at other times it took positions different from those of the Soviet Union.

Opening of the rift with China

The Sino-Soviet dispute. A gathering of Communist parties in Moscow in November 1957, in which China played a leading role, attempted to reassert a common doctrine while recognizing the need for differences in national practice. At Chinese insistence it also retained the Stalinist emphasis on the leadership of the Soviet Union. For a short time relations between the Soviet Union and China were harmonious: after 1955 Khrushchev had put an end to the humiliating terms that Stalin had imposed on China and inaugurated a policy of substantial economic aid.

The differences between China and the Soviet Union, which were to erupt into an open campaign of mutual abuse by 1962, were discernible to most observers by 1959, when the Soviet Union failed to give immediate political backing to Chinese military action against India and when China, at the same time, showed suspicion of Soviet talks with the United States in pursuit of Khrushchev's policy of "peaceful coexistence." In 1960 the differences widened, though they were still unpublicized. The Soviet Union withdrew its technical advisers from China as a preliminary to what was to prove an almost complete severing of economic relations. A facade of agreement was maintained, and at a conference of Communist parties held in Moscow in 1960 a series of resolutions was put forth to show that unity prevailed as ever in the ranks of the world Communist movement. News of serious disagreements, however, soon leaked out, for the increasing number of dissident groups within the several parties had by now rendered the maintenance of secrecy impossible. In the following year, 1961, the Soviet Union began a public polemic against the Chinese viewpoint. This was disguised as an attack on Albania, since 1959 a client of China and increasingly critical of Khrushchev's foreign policy. By 1962 the quarrel had become open and very bitter. It was conducted as a dispute over doctrine, but the practical issue underlying it was a basic rivalry for leadership of the world revolutionary movement.

The Sino-Soviet dispute had three major effects on this movement. It shattered the pretension that Marxism-Leninism offered a single world view, since at least two radically different ways of interpreting Marxism-Leninism were presented to Communists throughout the world, each backed by a Communist party in power with the prestige of a victorious revolution behind it. Second, it seriously impaired, if it did not destroy, the Soviet claim to be the leader of the world revolutionary movement. Since 1960 nearly all Communist parties have split into pro-Soviet and pro-Chinese portions, though outside Asia the Soviet portion has usually retained predominance. In the important parts of Asia, with the possible exception of India, where the party is divided into several warring factions, China has become the predominant influence upon Communist parties. Third, the mere fact of the dispute tended to create greater flexibility for individual parties within the Communist movement as a whole, even in the case of parties that nominally accepted Soviet leadership. The Romanians, for example, were able to follow a nationalistic course by which they successfully resisted Soviet attempts to integrate the Romanian economy into the bloc pattern. The Romanians also took an independent line in their trade relations with other countries, in refusing to participate in the 1968 invasion of Czechoslovakia, and in their policy toward Israel.

After the fall of Khrushchev in October 1964, his successors made efforts to reunite the world movement. They were only moderately successful. Seventy-five parties met in Moscow in June 1969, but of 14 parties in power five did not attend, and Cuba sent only an observer; Asia and Africa, the main areas of Chinese influence, were very poorly represented. Little unity emerged from the conference; in particular, the efforts of the Soviet Union to secure condemnation of China were unsuccessful. The resolution finally adopted was couched in such general terms as scarcely to conceal that the cracks had been merely pasted over. In the course of the 1970s, the hold of the Soviet Communist party over Communist parties outside the bloc seemed for a time to become weaker, with several parties (notably of France, Spain, and Italy) asserting independence from Moscow and the right to criticize Soviet policy. This movement, nicknamed "Eurocommunism," had lost much of its force by the end of the decade, however.

PROBLEMS OF INTERNAL REFORM

The attempt to modify Stalinism

A continuing problem in the history of Communist countries after the death of Stalin was the reform of their overcentralized political and economic structures. The only country that may be said to have achieved success was Yugoslavia, which had since 1948 asserted and maintained its independence from Soviet interference. After initially collectivizing much of its agriculture, Yugoslavia allowed the collective farms to dissolve. It also established Workers' Councils in the factories and publicized them in its foreign propaganda despite Soviet disapproval. The Yugoslav party program of 1958 contained three points in particular that were diametrically opposed to Soviet theory: that Socialism can be achieved without a revolution, that the Communist Party need not have a monopoly of leadership, and that danger of war arises from the existence of two power blocs in the world and not (as the Soviet Union contended) from the aggressive intentions of the United States. In January 1974, a new constitution was adopted that, apart from making changes in the representational system, provided for a collective presidency consisting of one member from each republic and autonomous province. Tito was elected president for life; after his death in 1980 this office rotated among the several members of the collective presidency.

Suppression of reform in Czechoslovakia. The most dramatic failure of an attempt at reform was in Czechoslovakia. The resignation of the old Stalinist party leader Antonín Novotný and his replacement by Alexander Dubček in January 1968 inaugurated a process of liberalization. The reformers hoped to humanize Communist rule by introducing basic civil freedoms, an independent judiciary, and other democratic institutions. The support of leading economists for this program was particularly significant since it indicated that they realized that the already accepted policy of economic decentralization (which included giving a measure of initiative to individual enterprises) would fail unless accompanied by political changes.

Invasion of Czechoslovakia

While the Czechoslovak Communists had repeatedly declared their intention to remain within the existing system, Moscow, possibly fearing that the developments they had set under way would ultimately endanger the stability of eastern Europe, endeavoured to induce the Czechoslovak party leaders to abandon their course. The Soviet effort failed, possibly because there were no Czechoslovak Communist leaders prepared, with Soviet help, to oust Dubček. Finally a group of Warsaw Pact forces—predominantly Soviet, but with token contributions from the other Warsaw Pact members except Romania—invaded Czechoslovakia on the night of August 20–21, 1968, effectively killing the momentum of the reform movement in Czechoslovakia. A Soviet-controlled security service was installed, and the Dubček leadership was gradually forced out of top posts and eventually expelled from the party. Although the repression was thorough, there was no mass terror.

The Soviet invasion of Czechoslovakia came as a greater shock to many Communists than the invasion of Hungary because it was directed against Communist leaders who strongly asserted their loyalty to Moscow. The motives that prompted Soviet action were probably two: one was the fear that the Soviet defense area created by Stalin after World War II might be endangered if the Dubček regime were allowed to continue; the other was the fear that the entrenched and conservative Communist parties in other European Communist countries, and in the Soviet Union itself, might not be equal to the challenge posed by a reformed Communism in Czechoslovakia.

Khrushchev's reforms. This concern that the power of the Communist party might be diminished may also have acted as a brake on internal reform. The reforms carried out by Khrushchev between 1953 and 1964 had been extensive. The arbitrary powers of the security police were brought under control; there were widespread reviews and rehabilitations (often posthumously) of the sentences of those sent to labour camps under Stalin; and reforms (in 1958) removed the worst anomalies of Soviet criminal law and procedure. The stringent controls over the lives of workers and farmers were relaxed. Discussion and debate were tolerated among writers and intellectuals to a degree that would have been inconceivable under Stalin. The whole system of agricultural management was considerably relaxed, and a system of incentives for the collective farmers was introduced. The limit of reform, as Khrushchev saw it, was the point at which any threat appeared to the party's control over all aspects of life. Under his successor, Leonid Brezhnev, the brake on reform was applied more heavily. Criticism of Stalin decreased. Freedom of opinion was considerably restricted by the introduction of penal provisions against "slandering" the Soviet system: for the first time since Stalin's death there were trials of writers, and the courts ceased to show any inclination to assert their independence as they had under Khrushchev. The numbers of political prisoners steadily increased, although the Brezhnev regime could not be compared to Stalin's. A movement toward economic reform had started under Khrushchev, aiming at some decentralization of economic control through greater freedom for enterprises to plan their own operations and through more influence for market forces. This was continued and officially encouraged after 1964 by Prime Minister Aleksey Kosygin, but it made little headway and was abandoned. The period of the 1970s was one of economic stagnation and conservatism at home, coupled with expansion of military power abroad.

Brezhnev's policy toward reform

COMMUNIST DOCTRINE AFTER STALIN

The errors of "revisionism" and "dogmatism." The most far-reaching innovation in Communist doctrine during the period 1953–70 was the Chinese interpretation of Marxism-Leninism known as Maoism. In the Soviet sphere several profound changes in doctrine took place after the death of Stalin. One change was the rise of ideological dispute for the first time since the early 1920s. The Yugoslav ideas were denounced as "revisionism," a term that harked back to the turn of the century when it had been used to characterize the views of Eduard Bernstein, who had argued that Socialism could be achieved without a revolution. After 1957 the terms "revisionism" and "dogmatism" became an integral part of Communist discourse. They were applied in a variety of meanings. By the Chinese, "revisionism" was used to mean, in effect, Khrushchevism—*i.e.,* the policies Khrushchev had introduced in both domestic and international relations, and which the Chinese opposed. On the Soviet side, "revisionism" became a catch-phrase to designate any political reform that appeared to endanger the dominance of the Communist Party: as defined at the Moscow conference of 1957 (with Chinese approval then) it was applied to all reform movements within the Communist system that denied "the historical necessity of the proletarian revolution," or the "Leninist principles for the construction of the party." The term "dogmatism," in Soviet usage, means a doctrinal conservatism that ignores changing realities, a clinging to received ideas in a way "calculated to alienate the party from the masses." The proper course, in the Soviet view, lies between revisionism and dogmatism: between excessive liberalism (as in Czechoslovakia in 1968), which may threaten the party's power, and excessive conservatism, which can lead to popular revolt (as in Hungary in 1956).

Program of the 22nd congress

Different roads to Socialism. The most important new elements in Soviet doctrine were set out in the party program adopted by the 22nd congress in October 1961; they were also, to some extent, embodied in the declarations of the Moscow conferences of 1957 and 1960. First, there was the concession that there are different roads to Socialism. This may have been no more than a practical recognition of the fact that since the breach with Yugoslavia and the death of Stalin it had no longer been possible for the Soviet Union to impose its own pattern on all Communist states. The invasion of Hungary in 1956, of Czechoslovakia in 1968, and of Afghanistan in 1979 were not, according to Moscow, inconsistent with this doctrine, since in each case the Soviet Union acted out of a duty to assist a fraternal Socialist state in putting down a counterrevolution. In the case of Czechoslovakia, which had not asked for such assistance, a new tenet was added by Brezhnev in November 1968. He contended that when "internal and external" forces hostile to Socialism attempted to restore capitalism in a Socialist country, it became a matter of concern to the whole Socialist community. This tenet was used to justify the action of the Warsaw Pact forces in August 1968 and of the Soviet forces in December 1979.

Peaceful coexistence. The second change in Soviet doctrine was the view that war between the capitalist and Socialist powers was no longer inevitable, as had always been asserted by both Lenin and Stalin. This was a practical recognition of the fact that a war waged with nuclear weapons would be more likely to lead to mutual annihilation than to victory. Khrushchev emphasized the possibility of "peaceful coexistence" between different social systems and the achievement of Socialism by peaceful means. In the 1970s "peaceful coexistence" became known as "détente." This doctrine raised hopes of real peace between Communist and non-Communist states, but the Soviet leaders made it clear that détente would not affect either political warfare against the West or military support for wars of liberation. The massive invasion of Afghanistan by the Soviet Uniȯn in December 1979 left détente seriously impaired.

The third doctrinal change after 1953 was also dictated by practical reality. The Comintern had rigidly applied concepts drawn from Western history to revolutions in Africa and Asia: industrialization, the emergence of a proletariat, and a Socialist revolution carried out under the leadership of a Communist party. This Marxist analysis proved to be totally unrealistic in the case of underdeveloped countries in which the predominant force was nationalism. This was increasingly recognized, after 1956, in Soviet doctrine that declared the proper revolutionary aim in the developing countries to be "national democracy." In Khrushchev's words this meant accepting a "noncapitalist path of development," which would be in the interests "not only of one class but of the broad strata of the people."

Priorities for the future. In the late 20th century the Soviet leadership faced two main problems: a slowing down in the rate of economic growth, to which the party had tied its promises of an improved standard of living, and a ferment of criticism among an intellectual minority, which included an influential component of leading scientists. Two alternatives seemed the most likely in the foreseeable future: either a return to more repressive measures, reminiscent of Stalin, or a reform of the Soviet system in the direction the Yugoslavs had taken. By the beginning of the 1980s, the prospects of reform in the direction of relaxation seemed remote, though the emergence of new leadership following the deaths of Leonid Brezhnev in 1982 and his successor Yury Andropov in 1984 left the question of the future open.

The Polish challenge

In other Communist countries within the Soviet sphere the fate of Czechoslovakia in 1968 seemed likely for some time to act as a deterrent against overbold attempts at reform. However, a return to the kind of domination achieved by Stalin seemed improbable, and the Soviet Union seemed likely to tolerate some degree of autonomy and variety among its fellow members of the Warsaw Pact. The emergence in Poland of Solidarity, a mass trade union movement independent of the Communist Party, briefly challenged the party's power in 1980, but, under Soviet pressure, the government responded in December 1981 with a decree of martial law under which Solidarity and other signs of dissidence were suppressed.

The future of Communism as a world movement necessarily depended upon relations between the Soviet Union and China. Their rivalry and their differences in doctrine seemed too deep for reconciliation. The economic backwardness of China, which would require decades to overcome, was likely to ensure a continuing gap in outlook between the two powers. In the competition for influence over potential revolutionary forces in Asia, Africa, and Latin America the advantages seemed to lie with China, which was closer to them in social and economic structure. On the other hand, the Soviet Union seemed likely to retain its lead over Communist parties in the industrially developed countries; even if it could no longer hope to use them as Stalin had, it could look to them as useful instruments in its continuing struggle against the non-Communist powers in general and the United States in particular. While the threat of Eurocommunism had receded, there seemed little prospect in the early 1980s of a return to the monolithic world Communist movement that Stalin had created.

(L.B.S.)

Anarchism

Anarchism is a term describing a cluster of doctrines and attitudes whose principal uniting feature is the belief that government is both harmful and unnecessary. Derived from a Greek root signifying "without a rule," the terms anarchism, anarchist, and anarchy are used to express both approval and disapproval. In early contexts all these terms were used pejoratively: during the English Civil War of the 17th century the opponents of the radical Levellers referred to them as "Switzerising anarchists," and during the French Revolution the Girondin leader Jacques-Pierre Brissot accused his most extreme rivals, the Enragés, of being the advocates of "anarchy."

> Laws that are not carried into effect, authorities without force and despised, crime unpunished, property attacked, the safety of the individual violated, the morality of the people corrupted, no constitution, no government, no justice, these are the features of anarchy.

These words uttered by the leader of the French Revolutionary moderates in 1793 could serve as a model for the denunciations delivered by all opponents of the anarchists. The latter, for their part, would admit many of Brissot's points. They deny man-made laws, regard property as a means of tyranny, and believe that crime is merely the product of a society based on property and authority. But they would argue that their denial of constitutions and governments leads not to "no justice" but to the real justice inherent in the free development of man's sociality, his natural inclination, when unfettered by laws, to live according to the principles and practice of mutual aid.

ANARCHIST THINKERS

The first man who willingly called himself an Anarchist was the French political writer and pioneer Socialist Pierre-Joseph Proudhon. In 1840, writing his controversial study of the economic bases of society, *Qu'est ce que la propriété?* (*What Is Property?*), Proudhon set out to shock his readers into attention by declaring: "I am an anarchist!" He went on to explain that in his view the real laws of society have nothing to do with authority but stem from the nature of society itself; he foresaw the eventual dissolution of authority and the emergence of a natural social order.

> As man seeks justice in equality, so society seeks order in anarchy. Anarchy—the absence of a sovereign—such is the form of government to which we are every day approximating.

The essential elements of the philosophy to which Proudhon in 1840 gave the name of Anarchism had already been developed by various earlier thinkers. There is a tradition of the rejection of political authority going back to classical antiquity, to the Stoics and the Cynics, and recurring throughout Christian history in dissenting sects such as the medieval Catharists and certain factions of Anabaptists during the Reformation. With such groups—often mistakenly claimed as ancestors by modern anarchist writers—the rejection of political government is merely one aspect of a retreat from the material world into a realm of spiritual grace; it becomes part of the search for individual salvation and as such is hardly compatible with the sociopolitical doctrine of Anarchism that in all its forms consists of (1) a fundamental criticism of the existing power-based order of society, (2) a vision of an alternative libertarian society based on cooperation as opposed to coercion, and (3) a method of proceeding from one order to the other.

The Diggers

English anarchist thought. The first sketch of an anarchist commonwealth in this sense was developed in the years immediately following the English Civil War by Gerrard Winstanley, a dissenting Christian who identified God with reason; he founded the minute Digger movement. In his pamphlet of 1649, *Truth Lifting Up Its Head Above Scandals,* Winstanley laid down what later became basic principles among the anarchists: that power corrupts; that property is incompatible with freedom; that authority and property are between them the begetters of crime; and that only in a society without rulers, where work and its products are shared, can men be free and happy, acting not according to laws imposed from above but according to their consciences. Winstanley was not only the pioneer of anarchist theory but also the forerunner of anarchist activism. He held that only by their own deeds can the people bring an end to social injustice, and in 1649, calling upon the people "to manure and work upon the common lands," he led a band of his followers in occupying a hillside in southern England, where they set about cultivation, established free communism among themselves, and offered passive resistance to the hostile landlords.

The Digger experiment was destroyed by local opposition, and Winstanley himself vanished into such obscurity that the place and date of his death are unknown. But the principles he defended lingered on in the traditions of English Protestant sects and reached their ultimate flowering in the masterpiece of a former dissenting minister, William Godwin, who in 1793 published his *Political Justice*—of which Sir Alexander Gray has said that in it "Godwin sums up, as no one else does, the sum and substance of anarchism, and thus embodies a whole tradition" (*The Socialist Tradition,* 1946, p. 134). This is essentially true, since Godwin not only presents the classic anarchist argument that authority is against nature and that social evils exist because men are not free to act according to reason but also sketches out a decentralized society in which the small autonomous community (the parish) is the essential unit. In his community, democratic political procedures are dispensed with as far as possible, because even the rule of the majority is a form of tyranny, and such procedures as voting dilute the responsibility of the individual; Godwin also condemns "accumulated property" as a source of power over others and envisages a loose economic system in which men will give and take according to their needs. Godwin was a prophet of technological progress; he believed that industrial development would eventually reduce the necessary working time to half an hour a day, provided men lived simply, and that this would facilitate the transition to a society without authority.

Godwin's classic argument

Godwin enjoyed great celebrity in the 1790s and influenced such varied writers as Percy Bysshe Shelley (whose *Queen Mab* and *Prometheus Unbound* are virtually anarchist poems), William Wordsworth, William Hazlitt, and Robert Owen, but he was almost forgotten by the time of his death in 1836. Though his ideas were to have, through Owen, a subterranean influence on the British labour movement, it is only recently that professed Anarchists have recognized his affinities with them. His *Political Justice* had virtually no effect on the quasi-political movement on the continent of Europe during the mid-19th century.

French anarchist thought. It was Pierre-Joseph Proudhon who laid the theoretical foundations of this movement. A brewer's son of peasant stock from the Franche-Comté, he started life (like many later anarchists) as a printer, but by the revolutionary year of 1848 he had already become a polemicist and a radical journalist with two books to his credit, *Qu'est ce que la propriété?* and *Système des contradictions économiques* (*System of Economic Contradictions*). These established him among the leading theoreticians of Socialism, a term that in the early 19th century embraced a wide spectrum of attitudes. In Paris during the 1840s, Proudhon associated with Karl Marx and the Russian Mikhail Bakunin, and, out of the experiences of the Revolution of 1848 (when he served in the Constituent Assembly and voted against the constitution "because it is a constitution"), he developed the libertarian theories that he discussed in later works such as *Du principe fédératif* (*The Federal Principle*) and *De la capacité politique des classes ouvrières* (*The Political Capability of the Working Classes*).

The work of Proudhon

Proudhon was a complex and voluminous writer who remained obstinately independent, refusing to consider himself the founder of either a system or a party. Yet he was justly regarded by Bakunin, Peter Kropotkin, and other leaders of organized Anarchism as their true ancestor, for he had adumbrated their philosophy.

Mutualism, federalism, and direct action were the essential doctrines Proudhon taught. By mutualism he meant the organization of society on an egalitarian basis. He declared that "property is theft," but this did not mean that he advocated communism. He attacked the use of property as a means of exploiting the labour of others, but he regarded "possession"—the right of a worker or group of workers to control the land or tools necessary for production—as an essential bulwark of liberty. He therefore envisaged a society formed of independent peasants and artisans, with factories and utilities run by associations of workers, all united by a system of mutual credit founded on people's banks. In place of the centralized state—the enemy of all Anarchists—Proudhon suggested a federal system of autonomous local communities and industrial associations, bound by contract and mutual interest rather than by laws, with arbitration replacing courts of justice, workers' management replacing bureaucracy, and integrated education replacing academic education. Out of such a network would emerge a natural social unity compared with which the existing order would appear as "nothing but chaos, serving as a basis for endless tyranny."

Proudhon remained all his life an independent polemicist, but in his final, posthumously published work, *De la capacité politique des classes ouvrières,* with its insistence that the liberation of the workers must be the task of the workers themselves organized in industrial associations, he laid the intellectual foundations of a movement that would reject democratic and parliamentary politics in favour of various forms of direct action. Unlike their master, Proudhon's working-class followers of the 1860s did not accept the title Anarchist (though in 1850 an independent revolutionary, Anselme Bellegarrigue, had founded a short-lived magazine entitled *L'Anarchie*); they preferred to call themselves Mutualists, after a working-class secret society bearing the same name to which Proudhon had belonged in Lyons during the 1830s. In 1864, shortly before Proudhon's death, a group of them joined with British trade unionists and European Socialists exiled in London to found the International Workingmen's Association (the First International). The Mutualists became the first opposition within the International to Karl Marx and his followers, who advocated political action and the seizure of the state in order to create a proletarian dictatorship. Marx's most formidable opponents, however, were not the Mutualists but the followers of Mikhail Bakunin, a Russian nobleman turned revolutionary who entered the International in 1868 after a long career as a political conspirator.

The work of Bakunin

Russian anarchist thought. Bakunin had begun as a supporter of nationalist revolutionary movements in various Slav countries. In the 1840s he had come under the influence of Proudhon, and by the 1860s, when he entered the International, he had not only founded his own proto-Anarchist organization, the Social Democratic Alliance, with a considerable following in Italy, Spain, Switzerland, and the Rhône Valley, but had modified the Proudhonian teachings into the doctrine that became known as collectivism. Bakunin accepted Proudhon's federalism and his insistence on the need for working-class direct action, but he argued that the modified property rights Proudhon allowed were impractical and instead suggested that the means of production should be owned collectively, though he still held that each worker should be remunerated only according to the amount of work he actually performed. The second important difference between Bakunin and Proudhon lay in their concepts of revolutionary method. Proudhon believed it was possible to create within existing society the mutualist associations that could replace it; he therefore opposed violent revolutionary action. Bakunin, declaring that "the passion for destruction is also a creative urge," refused to accept the piecemeal approach; a violent revolution, sweeping away all existing institutions, was in his view the necessary prelude to the construction of a free and peaceful society.

Though the individualism and nonviolence implicit in Proudhon's vision have survived in the peripheral currents of the anarchist tradition, it was Bakunin's stress on collectivism and violent revolutionary action that dominated the mainstream from the days of the First International down to the destruction of anarchism as a mass movement at the end of the Spanish Civil War in 1939.

The First International was itself destroyed by the conflict between Marx and Bakunin, a conflict rooted as much in the contradictory personalities of the two leaders as in their rival doctrines—revolution by a disciplined party versus revolution by the spontaneous insurgence of the working class. When the international finally broke apart at the Hague congress in 1872, Bakunin's followers were left in control of the working-class movements in the Latin countries—Spain, Italy, southern France, and French-speaking Switzerland—and these were to remain the principal bases of Anarchism in Europe. In 1873 the Bakuninists set up their own International, which lasted as an active body until 1877; during this period its members finally accepted the title of Anarchist rather than collectivist.

The work of Kropotkin

Bakunin died in 1876. His ideas had been developed in action rather than in writing, for he was the hero of many barricades, prisons, and meetings. His successor as the ideological leader, Peter Kropotkin (who renounced the title of prince when he became a revolutionary in 1876), is more celebrated for his writing than for his actions, though in his early years he led an eventful career as a revolutionary militant, which he described in a fine autobiography, *Memoirs of a Revolutionist* (1899). Under the influence of the French geographer Elisée Reclus (a former disciple of the Utopian Socialist Charles Fourier), Kropotkin developed the variant of anarchist theory known as anarchist communism. Kropotkin and his followers went beyond Bakunin's collectivism, since they argued not only that the means of production should be owned cooperatively but also that there should be complete communism in terms of distribution; this revived the scheme Sir Thomas More had sketched out in his 16th-century *Utopia* of common storehouses from which everyone should be allowed to take whatever he wished on the basis of "From each according to his means, to each according to his needs." In *La Conquête du pain* (*The Conquest of Bread,* 1892) Kropotkin sketched the vision of a revolutionary society organized as a federation of free communist groups. He reinforced the vision by writing *Mutual Aid: A Factor in Evolution* (1902), in which he endeavoured to prove by means of biological and sociological evidence that cooperation is more natural and usual among both animals and men than competition. In his *Fields, Factories and Workshops* (1899) he put forward ideas on the decentralization of industry appropriate to a nongovernmental society.

ANARCHISM AS A MOVEMENT

Anarchism and terrorism

Kropotkin's writings completed the theoretical vision of the Anarchist future, and little new has been added since his time. But this work was of less immediate importance than the emergence among the Italian anarchists of the theory of "propaganda of the deed." In 1876 Errico Malatesta expressed the belief of the Italian Anarchists that "the *insurrectionary deed,* destined to affirm socialist principles by acts, is the most efficacious means of propaganda." The first acts were rural insurrections, intended to arouse the illiterate masses of the Italian countryside. After the insurrections failed, Anarchist activism tended to take the form of individual deeds of protest by terrorists, who would attempt to kill ruling figures in the hope of demonstrating the vulnerability of the structure of authority and inspiring the masses by their self-sacrifice. In this way, between 1890 and 1901, a series of symbolic murders was enacted; the victims included King Umberto I of Italy, the empress Elizabeth of Austria, President Carnot of France, President McKinley of the United States, and Antonio Cánovas del Castillo, the prime minister of Spain. This brief but dramatic series of terrorist acts established the image of the Anarchist as a mindless destroyer; after 1901, however, the Anarchists continued to practice widespread terrorism only in such countries as Spain and Russia, where the general political atmosphere was conducive to violence.

During the 1890s, especially in France, Anarchism was adopted as a philosophy by avant-garde painters and writers. Gustave Courbet had already been a disciple of Proudhon; among those who in the 1890s accepted an Anarchist philosophy were Camille Pissarro, Georges Seurat, Paul Signac, Paul Adam, Octave Mirbeau, Laurent Tailhade, and, at least as a strong sympathizer, Stéphane Mallarmé. At the same time in England, Oscar Wilde declared himself an Anarchist and, under Kropotkin's inspiration, wrote his libertarian essay, "The Soul of Man Under Socialism" (1891).

The artists were attracted by the individualist spirit of Anarchism. By the mid-1890s, however, the more militant Anarchists in France began to realize that an excess of individualism had tended to detach them from the workers they sought to liberate. Anarchists, indeed, have always found it difficult to reconcile the claims of general human solidarity with the demands—equally insistent—of the individual who desires freedom. Some Anarchist thinkers, such as the German Max Stirner, who published *Der Einzige und sein Eigentum* (*The Ego and His Own*) in 1845, have refused to recognize any limitation on the individual's right to do as he will or any obligation to act socially; and even those who accepted Kropotkin's socially oriented doctrines of Anarchist communism have in practice been reluctant to create forms of organization that threatened their freedom of action or seemed likely to harden into institutions.

In consequence, although a number of international Anarchist congresses were held (the most celebrated being those of London in 1881 and of Amsterdam in 1907), no effective worldwide organization was created, even though by the end of the century the Anarchist movement had spread to all continents and was united by informal links of correspondence and friendship between the leading figures. National federations were weak even in countries where there were many Anarchists, such as France and Italy, and the typical unit of organization was the small group dedicated to propaganda by deed or word. Such groups engaged in a wide variety of activities; in the 1890s many of them concentrated on setting up experimental schools and communities that attempted to live out Anarchist principles.

Revolutionary Syndicalism. In France, where individualist trends had been most pronounced and public reaction to terrorist acts had imperilled the very existence of the movement, an effort was made to acquire a mass following. The Anarchists infiltrated the trade unions. They were particularly active in the *bourses du travail* ("labour exchanges"), local groupings of unions, originally set up to find work for their members, that appealed to the Anarchist ideal of decentralization. In 1892 a national confederation of *bourses du travail* was formed, and by 1895 the Anarchists, led by Fernand Pelloutier, Émile Pouget, and Paul Delesalle, had gained effective control and were developing the theory and practice of working-class activism that became known as Anarcho-Syndicalism, or Revolutionary Syndicalism.

Anarchism and the workers

The Anarcho-Syndicalists argued that the traditional function of trade unions—to struggle for better wages and working conditions—was not enough. The unions should become militant organizations dedicated to the destruction of capitalism and the state. They should aim to take over factories and utilities, which would then be operated by the workers. In this way the union or syndicate would have a double function—as an organ of struggle under the present dispensation and as an organ of administration after the revolution. To sustain militancy, an atmosphere of incessant conflict should be induced, and the culmination of this strategy should be the general strike. Many of the Syndicalists believed that such a massive act of noncooperation would bring about what they called "the revolution of folded arms," resulting in the collapse of the state and the capitalist system. But, although partial general strikes, with limited objectives, were undertaken in France and elsewhere with varying success, the millennial general strike aimed at overthrowing the social order in a single blow was never attempted. The Anarcho-Syndicalists acquired great prestige among the workers of France and, later, of Spain and Italy, because of their generally tough-minded attitude at a time when working conditions were bad and employers tended to respond brutally to union activity. After the great French trade-union organization, the Confédération Générale du Travail (CGT), was founded in 1902, their militancy enabled the Anarchists to retain control of the organization until 1908 and to wield considerable influence on its activities until after World War I.

Like Anarchism, Revolutionary Syndicalism proved attractive to certain intellectuals, notably Georges Sorel, whose *Réflexions sur la violence* (1908; Eng. trans., *Reflections on Violence,* 1914 and 1950) was the most important literary work to emerge from the movement. He argued the importance of the general strike as a social myth. The more purist Anarchist theoreticians were disturbed by the monolithic character of Syndicalist organizations, which they feared might create powerful interest structures in a revolutionary society. At the International Anarchist Congress at Amsterdam in 1907, a crucial debate on this issue took place between the young Revolutionary Syndicalist Pierre Monatte and the veteran Anarchist Errico Malatesta. It defined a division of outlook that still lingers in what remains of the historic Anarchist movement, which has always included individualist attitudes too extreme to admit any kind of large-scale organization.

Revolutionary Syndicalism did transform Anarchism, for a time at least, from a tiny minority current into a movement with considerable mass support, even though most members of Syndicalist unions were sympathizers and fellow travellers rather than committed Anarchists. In 1922 the Syndicalists set up their own International with its headquarters in Berlin, taking the historic name of the International Workingmen's Association; it still survives, with headquarters in Stockholm. When it was established the organizations that formed it could still boast considerable followings. The Unione Sindicale Italiana had 500,000 members; the Federación Obrera Regional Argentina, 200,000 members; the Portuguese Confederação General de Trabalho, 150,000 members; and the German Freie Arbeiter Union, 120,000 members. There were smaller organizations in Chile, Uruguay, Denmark, Norway, Holland, Mexico, and Sweden. In Britain the influence of Syndicalism was shown most clearly in the Guild Socialist movement that flourished briefly in the early years of the present century. In the United States, Revolutionary Syndicalist ideas were manifested in the Industrial Workers of the World (IWW), which in the years immediately before and after World War I played a vital part in organizing American miners, loggers, and unskilled workers; but only a small minority of the IWW militants were ever avowed Anarchists.

Syndicalist organizations

Anarchism in Spain. The reconciliation of Anarchism and Syndicalism was most complete and most successful in Spain; for a long period the Anarchist movement in that country remained the most numerous and the most powerful in the world. The first known Spanish Anarchist, Ramón de la Sagra, a disciple of Proudhon, founded the world's first Anarchist journal, *El Porvenir,* in La Coruña in 1845; it was quickly suppressed. Mutualist ideas were later publicized by Pi y Margall, a federalist leader and the translator of many Proudhon books; during the Spanish revolution of 1873, Pi y Margall attempted to establish a decentralist, or "cantonalist," political system on Proudhonian lines. In the end, however, the influence of Bakunin was stronger. In 1868 his Italian disciple, Giuseppe Fanelli, visited Barcelona and Madrid, where he established branches of the International. By 1870 they had 40,000 members, and in 1873 the movement numbered about 60,000, organized mainly in working men's associations. In 1874 the Anarchist movement in Spain was forced underground, a phenomenon recurring often in subsequent decades. It flourished, nevertheless, and Anarchism became the favoured type of radicalism among two very different groups, the factory workers of Barcelona and other Catalan towns and the impoverished peasants who toiled on the absentee-owned estates of Andalusia.

As in France and Italy, the movement in Spain during the 1880s and 1890s was inclined toward insurrection (in Andalusia) and terrorism (in Catalonia). It retained its strength in working-class organizations because the courageous and even ruthless Anarchist militants were often the only leaders who would stand up against the army and the employers, who hired squads of gunmen to engage in guerrilla warfare with the Anarchists in the streets of Barcelona. The workers of Barcelona were finally inspired by the success of the French CGT to set up a Syndicalist organization, Solidaridad Obrera (Workers' Solidarity). Established in 1907, Solidaridad Obrera quickly spread throughout Catalonia, and in 1909, when the Spanish army tried to conscript Catalan reservists to fight against the Riffs in Morocco, it called a general strike. "La Semana Tragica," "the Tragic Week" of largely spontaneous violence that followed (with hundreds dead and 50 churches and monasteries destroyed), ended in violent repression. Tortures of Anarchists in the fortress of Montjuich and the execution of the internationally celebrated advocate of free education Francisco Ferrer led to worldwide protests and the resignation of the conservative government in Madrid. These events also resulted in a congress of Spanish trade unionists at Seville in 1910, which founded the Confederación Nacional del Trabajo (CNT).

Moderates and activists

The CNT, which included the majority of organized Spanish workers, was dominated throughout its existence by the Anarchist militants; these in 1927 founded their own activist organization, the Federación Anarquista Iberica (FAI). While there was recurrent conflict within the CNT between moderates and FAI activists, the atmosphere of violence and urgency in which radical activities were carried on in Spain ensured that the more extreme leaders, such as Garcia Oliver and Buenaventura Durutti, tended to wield the decisive influence. The CNT was a model of Anarchist decentralism and antibureaucratism: its basic organizations were not national unions but *sindicatos únicos,* which brought together the workers of all trades and crafts in a certain locality; the national committee was elected each year from a different locality to ensure that no individual served more than one term; and all delegates were subject to immediate recall by the members. This enormous organization, which claimed 700,000 members in 1919, 1,600,000 in 1936, and more than 2,000,000 during the civil war, employed only one paid secretary. Its day-to-day work was carried on in their spare time by workers chosen by their fellows. This meant that the Spanish Anarchist movement was not dominated by the *déclassé* intellectuals and self-taught printers and shoemakers who were so influential in other countries. One consequence was that the Spanish movement contributed nothing original to the ideological literature of Anarchism; like Bakunin, whom they favoured most among the classic libertarian thinkers, the Spanish Anarchists developed their attitudes in action rather than in writing.

Anarchists in the Spanish Civil War

The CNT and the FAI, which remained clandestine organizations under the dictatorship of Primo de Rivera, emerged into the open with the abdication of King Alfonso XIII in 1931. Their antipolitical philosophy led them to reject the republic as much as the monarchy it had replaced, and between 1931 and the military rebellion led by Francisco Franco in 1936 there were several unsuccessful Anarchist risings. In 1936 the Anarchists, who over the decades had become expert urban guerrillas, were mainly responsible for the defeat of the rebel generals in both Barcelona and Valencia, as well as in country areas of Catalonia and Aragon; and for many early months of the Civil War they were in virtual control of eastern Spain, where they regarded the crisis as an opportunity to carry through the social revolution of which they had long dreamed. Factories and railways in Catalonia were taken over by workers' committees, and in hundreds of villages in Catalonia, Levante, and Andalusia the peasants seized the land and established libertarian communes like those described by Kropotkin in *La Conquête du pain.* The internal use of money was abolished, the land was tilled in common, and village products were sold or exchanged on behalf of the community in general, with each family receiving an equitable share of food and other necessities. An idealistic Spartan fervour characterized these communities, which often consisted of illiterate labourers; intoxicants, tobacco, and sometimes even coffee were renounced; and millenarian enthusiasm took the place of religion, as it has often done in Spain. The reports of critical observers suggest that at least some of these communes were efficiently run and more productive agriculturally than the villages had been previously.

The Spanish Anarchists failed during the Civil War largely because, expert though they were in spontaneous street fighting, they did not have the discipline necessary to carry on sustained warfare; the columns they sent to various fronts were unsuccessful in comparison with the Communist-led International Brigades. In December 1936 four leading Anarchists took posts in the Cabinet of Francisco Largo Caballero, radically compromising their antigovernmental principles. They were unable to halt the trend toward left-wing totalitarianism encouraged by their enemies the Communists, who were numerically far fewer but politically more influential. In May 1937 bitter fighting broke out in Barcelona between Communists and Anarchists. The CNT held its own on this occasion, but its influence quickly waned. The collectivized factories were taken over by the central government, and many agricultural communes were destroyed by Franco's advance into Andalusia and by the hostile action of General Lister's Communist army in Aragon. In January 1939 the Spanish Anarchists were so demoralized by the compromises of the Civil War that they were unable to mount a resistance when Franco's forces marched into Barcelona; the CNT and FAI became phantom organizations in exile.

The competition of Communism

Decline of the Anarchist movement. By this time the movement outside Spain had been either destroyed or greatly diminished as a result of two developments: the Russian Revolution and the rise of right-wing totalitarian regimes. Though the most famous Anarchist leaders, Bakunin and Kropotkin, had been Russian, the Anarchist movement had never been strong in Russia, partly because the more numerous Socialist Revolutionary Party (the Narodniki) had adopted Bakuninist ideas while remaining essentially a constitutional party. After the 1917 revolution the small anarchist groups that emerged in Petrograd (now Leningrad) and Moscow were powerless against the Bolsheviks, and Kropotkin, who returned from exile, found himself without influence. Only in the south did N.I. Makhno, a peasant Anarchist, raise an insurrectionary army that by brilliant guerrilla tactics held a large part of the Ukraine from both the Red and the White armies; but the social experiments developed under Makhno's protection were rudimentary, and when he was driven into exile in 1921 the Anarchist movement became extinct in Russia.

In other countries, the prestige of the Russian revolution enabled the new Communist parties to win much of the support formerly given to the Anarchists, particularly in France, where the CGT passed permanently into Communist control. The large Italian Anarchist movement was destroyed by Benito Mussolini's Fascist government in the 1920s, and the small German Anarchist movement was smashed by the Nazis in the 1930s. In Japan, Anarchism had emerged during the Russo-Japanese war of 1904–05, when the Socialist leader Shusui Kotoku became converted by reading Kropotkin in prison; Kotoku and other Anarchists were executed in 1911 for their involvement in a plot against the Emperor, but, after World War I, new Anarchist organizations appeared, including a Black Federation and a Syndicalist federation. After the Japanese invasion of Manchuria in 1931, the imperial government began to suppress all left-wing groups, and the Anarchist movement was finally destroyed in 1935 after a secret society, the Anarchist Communist Party, had been accused of plotting armed insurrection.

Anarchism in the Americas suffered similar reverses. In the United States, a native and mainly nonviolent tradition developed during the 19th century in the writings of Henry David Thoreau, Josiah Warren, Lysander Spooner, and Benjamin Tucker (editor of *Liberty,* an anarchist journal published in Boston and later in New York City, 1881–1908). Activist Anarchism in the U.S. was mainly sustained by immigrants from Europe, including Johann Most (editor of *Die Freiheit*), Alexander Berkman (who attempted to assassinate steel magnate Henry Clay Frick in 1892), and Emma Goldman, whose *Living My Life* gives a picture of radical activity in the United States at the turn of the century. Anarchism appeared as a dramatic element in American life in 1886, when seven policemen were killed in the Haymarket bombing in Chicago and four Anarchist leaders were executed—unjustly, as later investigations revealed. In 1901 a Polish Anarchist, Leon Czolgosz, assassinated President McKinley. In 1903 the U.S. congress passed a law to bar foreign Anarchists from the country and to deport alien Anarchists found within it. In the repressive mood that followed World War I the Anarchists, like the IWW, were suppressed; Alexander Berkman, Emma Goldman, and many others were imprisoned and deported.

In Latin America strong Anarchist elements were involved in the Mexican Revolution. The Syndicalist teachings of Ricardo Flores Magon influenced the peasant revolutionism of Emiliano Zapata. After the deaths of Zapata in 1919 and Flores Magon in 1922, the revolutionary image in Mexico, as elsewhere, was taken over by the Communists. In Argentina and Uruguay considerable Anarcho-Syndicalist movements existed early in the 20th century, but they too were greatly reduced by the end of the 1930s through intermittent repression and the competition of Communism.

CONTEMPORARY CURRENTS

Wider influence of anarchist ideas

In the 1970s the theories of Anarchism aroused more interest and sympathy than at any time since the Russian revolution. The same cannot be said of the Anarchist movement itself. After World War II, Anarchist groups and federations re-emerged in almost all countries where they had formerly flourished—the notable exceptions being Spain and the Soviet Union—but these organizations wielded little influence compared to that of the broader movement inspired by libertarian ideas.

Such a development is not inappropriate, since Anarchists have never stressed the need for organizational continuity, and the cluster of social and moral ideas that are identifiable as Anarchism has always spread beyond any clearly definable movement. The Russian writer Count Leo Tolstoy refused to call himself an Anarchist because those who used the title at that time in Russia were terrorists; nevertheless, he had developed out of his rational Christianity a form of pacifist radicalism that rejected the state and all forms of government, called for the simplification of life in the name of moral regeneration, and sought to replace property by free communism. An impressive example of the breadth of Anarchist influence is Mahatma Gandhi, who based his strategy of nonviolent civil disobedience in South Africa and India on the teachings of nonviolent Anarchists such as Tolstoy and Thoreau and who remembered his reading of Kropotkin when he devised for India the plan of a decentralized society based on autonomous village communes. Gandhi's village India has not come into being, but the movement known as Sardovaya, led by Vinoba Bhave and Jaya Prakash Narayan, has been working toward it through *gramdan*—community ownership of land. By 1969 a fifth of the villages of India had declared for *gramdan,* and, while this remained largely a matter of unrealized gestures, it represented perhaps the most extensive commitment to basic Anarchist ideas in the contemporary world.

Anarchism's appeal to intellectuals

In the West the appeal of Anarchism has been strongest, at least since 1917, among the intellectuals. Kropotkin's arguments for decentralization have wielded their strongest influence among writers on social planning such as Patrick Geddes, Lewis Mumford, and Paul Goodman—himself a declared Anarchist. Anarchist treatises from Godwin onward have contributed to the progressive movement in education, and the most influential of the many books written by the libertarian critic Sir Herbert Read was his *Education Through Art* (1943). But in the 1960s and 1970s Anarchism became popular among rebellious students and the left in general because of the values it opposed to those of the increasingly technological cultures of western Europe, North America, and Japan.

Here the mediating thinker was undoubtedly Aldous Huxley, who anticipated many elements of the "counter culture" of the 1960s and 1970s. In *Brave New World* (1932) Huxley had presented a warning vision of the kind of mindless, materialistic existence a society dominated by technology might produce. In his "Foreword" to the 1946 edition, he stated his belief that only by radical decentralization and simplification and by a politics that was "Kropotkinesque and co-operative" could the dangers implicit in modern society be avoided. In his later writings he explicitly accepted the validity of the Anarchist critique of existing society.

The emergence of Anarchist ideas in a wider frame of reference began in the American civil rights movements in the 1950s, with their recognition of the need to resist injustice through other than legal channels in the name of a morality different from that recognized by constituted authority. By the end of the 1950s the new radicalism in the United States, Europe, and Japan had begun to take a course separating it from the narrower issues of the civil rights movement; it shifted toward a general criticism of the elitist structure and materialist goals of modern industrial societies—Communist as well as capitalist. Within this movement there was a limited revival of traditional Anarchism—exemplified in the sudden popularity of the American writer Paul Goodman and the emergence in Britain of a sophisticated monthly, *Anarchy,* that applied Anarchist ideas to the whole spectrum of modern life.

Anarchism and contemporary rebellion

In all of this, however, Anarchism was only one strand in what can be described as a climate of rebellion rather than an ideology. Anarchist ideas were mingled with strains of Leninism, early Marxism, unorthodox psychology, and often with elements of mysticism, neo-Buddhism, and Tolstoyan Christianity. None of the leaders of the student rebellions in the United States and Germany or the militants of Zengakuren in Japan could be called in any literal and complete sense an Anarchist, although they had read Bakunin as well as Marx and "Che" Guevara. In Paris, the leaders of the Anarchist Federation admitted that they had no influence at all on the strikes and street fighting in 1968 that seemed to threaten the very existence of the French state.

To all these movements, which rejected the old parties of the left as strongly as they did the existing political structure, the appeal of Anarchism was strong and understandable. The Anarchist outlook, in its insistence on spontaneity, on theoretical flexibility, on simplicity of life, on love and anger as complementary and necessary components in both social and individual action, appeals to those who reject the impersonality of institutions and the calculations of political parties. The Anarchist rejection of the state, the insistence on decentralism and local autonomy, found strong echoes among those who talked of participatory democracy. The recurrence of the theme of workers' control of industry in so many manifestos of the 1960s, notably during the Paris insurrection of 1968, showed the enduring relevance of Anarcho-Syndicalist ideas.

The Anarchist insistence on direct action had an almost universal appeal to the radicals of the 1960s and 1970s, who advocated extraparliamentary action and confrontation. Some student groups in France, the United States, and Japan accepted Bakunin's pan-destructionism, holding that existing society was so corrupt that it had to be swept completely away. They were fond of quoting a sentence of Bakunin:

> There will be a qualitative transformation, a new living, life-giving revelation, a new heaven and a new earth, a young and mighty world in which all our present dissonances will be resolved into a harmonious whole.

This kind of secular apocalypticism, which envisaged total violence as the paradoxical way to total reconciliation, linked many of the radicals with the Anarchists who took to terrorism in the 1880s and 1890s. Now, as then, violence proceeded from theory to action, as the upsurge of terrorist bombings and urban-guerrilla activity demonstrated at the end of the 1960s.

Community reform

Anarchism as an ideal. But there were others who believed with Tolstoy and Gandhi that a moral change must come first and who sought revolution by evolution. Paul Goodman, with his proposals for urban reform as a step toward a freer society, was one of them. Some turned to the creation of models of a different kind of society in the form of utopian communities. This has been a recurrent feature of radical movements, appearing in the United States during the mid-19th century, in Britain as a form of constructive pacifist protest during World War II, and in the United States in the 1970s as a way of manifesting one's rejection of normal life-styles.

Experience suggests that utopian experiments and radical actions are not likely to achieve that wholly nongovernment society of which libertarians have dreamed. The true value of their vision was stated by the Anarchist poet Herbert Read in his last book, *The Cult of Sincerity* (1968):

> My understanding of the history of culture has convinced me that the ideal society is a point on a receding horizon. We move steadily towards it but can never reach it. Nevertheless we must engage with passion in the immediate strife . . .

It is precisely as an ideal, as a touchstone to judge the existing world, that the Anarchist vision is useful. It corresponds to a recurrent and necessary strain in human thought—the revulsion against regimentation, against large organizations, against complexity and luxury. The insights of the Anarchist are likely to find their maximum usefulness in urban and rural planning, in the development of community relations based on full participation, in integrated education, and perhaps most of all in encouraging what Read calls a "process of individuation, accomplished by general education and personal discipline." Anarchism is a moral and social doctrine before it is a political one; it stands as a permanent reminder of the perils of national and corporate giantism and of the virtues of local interests and loyalties. It teaches the vigilance by which man may be able to avoid such bleak utopias as those of Aldous Huxley and George Orwell. (G.W.)

Fascism

Fascism comprises a political attitude and a mass movement that tended to dominate political life in central, southern, and eastern-central Europe between 1919 and 1944. Common to all fascist movements was an emphasis on the nation (race or state) as the centre and regulator of all history and life, and on the indisputable authority of the leader behind whom the people were expected to form an unbreakable unity. The word fascism itself was first used in 1919 by Benito Mussolini in Italy; in the following years the influence of fascism made itself felt in countries as far away as Japan, Argentina, Brazil, and the Union of South Africa, its specific aspects varying according to the country's political traditions, its social structure, and the personality of the leader. The Italian word *fascio* (derived from the Latin *fasces,* a bundle of rods with an ax in it) symbolized both aspects: the power of many united and obeying one will and the authority of the state, which was the supreme source of law and order and all national life.

THE PHILOSOPHICAL BASES OF FASCISM

The Spartan model

Fascism rejected the main philosophical trends of the 18th and 19th centuries, the "spirit" of the American and French revolutions with their emphasis on individual liberty and on the equality of men and races. The message of the Enlightenment had served to enhance the dignity of the individual and had emphasized openness in a secularized society. In contrast, fascism extolled the supreme sovereignty of the nation as an absolute. It demanded the revival of the spirit of the ancient *polis* (city-state), above all of Sparta with its discipline and total devotion to duty, and of the complete coordination of all intellectual and political thought and activities against modern individualism and scientific skepticism. The Italian slogan "to believe, to obey, to combat" was fascism's antithesis to "liberty, equality, fraternity," and to the prophetic and Christian messages of peace. The combination of an unquestioning faith and of a virile combativeness was to transform the nation into a permanently mobilized armed force to conquer, maintain, and expand power.

In its beginnings fascism was not a doctrine and had no clearly elaborated program. It was a technique for gaining and retaining power by violence, and with astonishing flexibility it subordinated all questions of program to this one aim. From the beginning it was dominated by a definite attitude of mind that exalted the fighting spirit, military discipline, ruthlessness, and action and rejected all ethical motives as weakening the resoluteness of will. Stressing the irrational and instincts and activism, fascism insisted that the strong will always prevail over the weak, the more resolute over the irresolute. Ultimately everything depended upon the decisions of the leader, decisions to be blindly obeyed and immediately executed. Fascism returned to an authoritarian order, based upon the subordination of the individual and the inequality of caste and rank.

The reliance on power. Power is, of course, an element present in all political life. The first major writer to abandon the moral and normative approach to politics in favour of pure power was the Florentine man of letters Niccolò Machiavelli (1469–1527). A man of the Renaissance, he looked to the people of pre-Christian antiquity as the original possessors of *virtú,* the civic virtue necessary to the modern ruler; he believed that Christianity was, unfortunately, "true," but that its stress on meekness and humility would damage political man, weakening and at the same time fanaticizing him. Machiavelli's methodology involved the empirical observation of human nature and behaviour, which he believed to be changeless. His deep feelings about the degradation and corruption of Italy at his time led him to put his hope into the daring and the violence of a great man who would exercise power ruthlessly but with prudence. Power, Machiavelli apparently believed, legitimized the state, if rationally applied, as *raison d'état,* by a man able to manipulate the people and use the army for his own purposes. In his quest for a "new prince" and a new principle of policy he knew that he was opening "a road as yet untrodden by man." The road led to the absolute sovereign state.

The emphasis on sovereign-state power. In the bitter and protracted religious conflict of 16th-century France, the French jurist Jean Bodin (1530–96) stressed the importance of the sovereign, but by no means unlimited, power of the state in effective government. During the constitutional crisis of 17th-century England the philosopher Thomas Hobbes (1588–1679) saw sovereign power as more absolute, unlimited by the subjects who have authorized it and responsible only to God. For Machiavelli the state was a work of art, created by the skill of the prince whom Machiavelli wished to teach the rules of conduct; for Bodin and Hobbes the state was a rational contrivance to lift central authority above religious and civil disputes. The peace treaty of Westphalia in 1648, in an attempt to end over a century of religious warfare, gave the secular sovereign, generally a hereditary absolutist monarch, the right to determine the religious beliefs of his subjects. The maintenance of law and order became the highest guiding principle, but even at this stage the state had not yet become the object of awe or emotional veneration.

The state in German Romanticism

The state became such only after the French Revolution, and above all in the emotional teaching of German Romantic philosophers, such as Johann Gottlieb Fichte (1762–1814) and Georg Wilhelm Friedrich Hegel (1770–1831). For these men the national collectivity assumed, morally and politically, an absolute rank. Fichte's utopian "closed state" was authoritarian, anti-individualistic, and economically self-sufficient.

But Fichte did not endow the state with the sacral aura that Hegel gave it. For a century German historiography was to accept the Hegelian concept of the absolute-power state that acts in its own self-interest without consideration for humanitarian principle or for the rights of individuals or of other states. Hegel's followers, like those of Fichte, overlooked the complexity and ambiguity of his philosophy and concentrated on his exaltation of the state as an end in itself, as the "actuality of the ethical idea," as "absolutely rational," and as the source of all "concrete freedom." Only as a subject of the state (specifically, for Hegel, the Prussian monarchy) does the individual gain objective reality and an ethical life. The state's unconditional sovereignty reveals its nature above all in war.

The integrally autarkic state: the Dreyfus Affair. In the first part of the 19th century Hegel was, then, the philosopher of a militaristic monarchy opposed to the growing strength of 19th-century middle-class liberalism. At the end of the century another antiliberal, anti-individualistic, and authoritarian movement made its influence felt, this time in the middle-class democracy of the French Republic. Its leader was Charles Maurras (1868–1952), who repudiated the disunity and verbosity of parliamentarianism. He admired the Roman Catholic Church (but not Christianity, which in his opinion glorified universalism, social unrest, and pacifism), but for the virtues of hierarchic discipline and traditionalist order. The supreme norm of political life was to Maurras the absolute primacy of France and this meant action in France's interest instead of hesitation, parliamentary discussion, and consideration of world opinion.

Maurras and Barrès: the state can do no wrong

In 1894 a military tribunal convicted Capt. Alfred Dreyfus, the only Jewish officer in the French general staff, of pro-German espionage. Dreyfus insisted on his innocence, and his case became the centre of a bitter dispute involving questions of the precedence of national interest over objective justice. If Dreyfus were innocent, then the French army command on which France depended was dishonoured and stood accused of intrigue and worse. Antiliberal forces saw in the dispute an opportunity to overthrow the parliamentary republic, in which many high army officers and the church saw something fundamentally un-French, the work of Freemasons and Protestants, Anglo-Saxons and Jews. Whereas Maurras, in his struggle against democracy, pleaded for a return to the ancient monarchy, his compatriot, Maurice Barrès (1862–1923), saw the remedy in a leader who, being in close touch with the masses, would be the embodiment and the effective will of their thought and feeling. Against the rational cosmopolitanism of the "uprooted" intellectuals (a term that he made the title of his most famous novel in 1897), Barrès extolled the close community of a nation with its deep roots in past generations and in the ancestral soil (the "blood" and "soil"—*Blut* and *Boden* of Hitler). To Barrès the individual was merely a link in the chain of generations, inevitably determined by the blood of common forefathers.

Vitalism and elitism. The beginning of the 20th century felt the influence of Friedrich Nietzsche's (1844–1900) contradictory and often misunderstood work. Not a forerunner of fascism, Nietzsche despised German nationalism, anti-Semitism, and the authority of the state. He was, in fact, an extreme individualist, in revolt against all uncritical obedience by believers or followers and against the traditional values of church and fatherland. But he also despised the common man and democracy; he believed in great personalities and their exclusive rights. He found his time lacking in greatness and heroism, and he glorified the courage of warriors, though he meant first of all warriors of the mind, strong enough to overcome their own pettiness and their acceptance of faith or belief that came to them from second or third hand.

Spengler and Sorel: the struggle of peoples

Oswald Spengler (1880–1936) shared Nietzsche's feeling of the decadence of Western civilization, brought about by Christianity and democracy, and his faith in the need for an aristocratic elite. Writing during World War I, Spengler insisted that all history is struggle among nations and that each nation's future will be decided by its power relationships to other peoples. Each people must be "in condition" for inevitable struggle and must trust its leaders; what is significant is not the victory of truth but the triumph of the will-to-power.

The French radical antiliberal Socialist Georges Sorel in his revolutionary syndicalism emphasized the dynamism and new vitality of a heroic proletariat against an effete bourgeoisie. In his *Reflections on Violence* (1908) Sorel claimed that the working-class movement needed irrational myths to carry out its role in history. This idea influenced many Socialists in Latin countries, especially in north-central Italy, at the time Benito Mussolini was growing up. Violence, Sorel declared, was "sublime" when it was exercised by a movement with a historical mission. In Sorel, radical socialist theory of the left fused with a radical conservatism of the right in common rejection of bourgeois "mediocrity."

Among Italian pre-1914 social philosophers, other more conservative forerunners preached an elitist doctrine of vitality and the competitive power struggle. They turned from Count Cavour's liberal faith in parliamentarianism, which had established the unified Italian kingdom of 1861, to a quest for new elites and new rationales. Among them was Gaetano Mosca (1858–1941). Mosca's *Elimenti di scienza politica* (1896; Eng. trans. *The Ruling Class,* 1939) owed much to the Austrian professor of public law Ludwig Gumplowicz (1838–1909), whose fundamental work *Der Rassenkampf* ("Racial Struggle") in 1883 established the "group" as the fundamental unit of sociology, which he interpreted as the science of the interaction of groups. Material need was to Mosca the prime motive of human conduct; conquest and the satisfaction of the conqueror's need by the labour of the conquered, the fundamental essence of history.

Mosca and Vilfredo Pareto (1848–1923) argued that there had always been a ruling class of a few men who held power over the majority; that society is, thus, hierarchically organized, though the elites may change and, in fact, the change of elites is as much the essence of history as are wars between ethnic groups. The new elites carry with them their own values, expressed in social myths that can neither be proved nor disproved and that serve as a call and inspiration to action. The collective psychology of Scipio Sighele (1868–1913) and Gustave Le Bon (1841–1931) hypothesized that crowds obey collective subconscious emotions rather than the rational individuality of their single members, and, therefore, that crowds are highly susceptible to manipulation by leadership that can arouse in them heroism or savagery of which the individual alone would be incapable. Endlessly reiterated statements rather than rational thought influence public morality.

THE CONDITIONS FOR THE EMERGENCE OF FASCISM: POLITICAL PREREQUISITES

The troubled state of Europe in the years before 1914 was greatly intensified by the pressures that World War I put on societies that were not yet socially and politically modernized. This was true in varying degrees of Germany, Italy, and Japan, all of whom had entered the war in the expectation of great gains in territory and status, and of acquiring full equality with the older societies of the West. The deep moral depression and confusion which the defeat of 1918 produced in Germany was due to the apparently inexplicable difference between expectations and final failure.

National frustration

The discrepancy between Germany's advanced industry and her semifeudal-state structure, with its traditional authoritarian basis, had weakened her war effort. The failure on the battlefield led to a deeply emotional nationalism which ascribed the shortcomings not to Germany's backward political structure but to enemy plots and domestic "enemies." Thus the authoritarian militarist elite was not discredited by the defeat of 1918; on the contrary, the fear of Bolshevism brought support for the defense of the traditional structure.

Though Italy was to be among the victors in World War I, the relative backwardness of its political, social, and economic structure in 1914 put an immense strain on all aspects of Italian society and life. The failure of expected gains from the war to materialize led to a weakening of the country's insecure liberal foundations. Those who, like Benito Mussolini, had agitated for Italy's entrance into the war in 1915 tried to direct the discontent and fear of the population against the victorious democracies who, unlike Italy, had emerged strengthened from the war. Social unrest frightened the propertied classes, the major landowners, and the church. Instead of carrying through long overdue reforms, they sought for a strong man who could sway part of the masses, war veterans, and lower middle class and turn them against the threat of Bolshevism. Fascism was thus regarded as a bulwark against the modernization of Italy. Though this was an underestimation of the syndicalist and radical aspects of Mussolini's original position, he was able to achieve, more than Hitler did later, an accommodation with the old ruling class, the monarchy, the army, and the church.

Fascist movements, wherever they have arisen, have frequently been inspired by national feelings of disappointment and by the assumed need to close ranks in order to reach often fantastic goals (*e.g.,* the revival of Roman glory). Japan's fascist movement was linked to its attempt in World War I to establish a protectorate over China, which was frustrated, largely, as Japan felt, by the United States. Similarly, the strength of the fascists in Hungary owed much to the bitter national resentment at the loss of its non-Magyar subjects to new or enlarged nation-states created at the end of the war. These new states were not politically strong, and, as in Spain or Latin America, traditional right-wing conservatism, backed by the church and the pre-1914 oligarchy, found itself in conflict with the dynamism of fascism and its contempt for traditional ideas. Later, both were to enter into uneasy alliances in their fear of Bolshevism.

SOCIAL AND ECONOMIC CONDITIONS THAT ENCOURAGE THE DEVELOPMENT OF FASCISM

Politically and socially the modern, industrial middle-class societies that developed in Britain, Scandinavia, Switzerland, the Low Countries, and France in the 19th century showed a great power of resistance to fascism, which was, on the whole, confined to fringe movements. Even in Germany, with her bitter resentments accumulated from her failure in World War I, which Hitler masterfully manipulated and fused with older resentments, fascism would probably not have come to power had it not been for the inflation crisis of 1923 and the widespread unemployment of the early 1930s. Finally, the rise of fascism in Germany owed much to the weakness of civilian democracy in that country. Germany had originated as a nation-state in 1871, thanks almost entirely to the military-authoritarian tradition of Prussia and the victories achieved by its army without the aid of any other power. The new *Reich* was proclaimed at the gates of Paris, the capital of the defeated enemy, and the German middle classes and German scholarship all willingly accepted the traditional values of the efficient ruling class, though by 1900 these values were insufficient to support an expanding modern industrial society.

Nationalism and economic crisis

Neither a capable semimilitary bureaucracy nor a scientific technology existed in Italy, Spain, Portugal, Greece, or Romania, where low urban and rural productivity during the early 20th century constantly widened the gap between these countries and the modern West and provided a basis for the growth of fascism.

In Italy, the farthest advanced of these countries, it was estimated in 1900 that half the population could neither write nor read. Though the north of Italy made great progress in the first 15 years of the present century, the fact is that even in 1914 the per capita income of the country, measured in standard gold units, was only 105 compared with 237 for Britain and 182 for France. Southern Italy's peasant population lived, according to one account, "in conditions of utmost destitution, illiteracy was the rule. Afflicted with gross dietary deficiency, accustomed to deprivation, and mulled by ceaseless toil, whole regions were innocent of the most elementary education and were consequently not equipped to participate in the challenge" offered by modern civilization. Archaic traditions and authoritarian religion preserved in Italy and Spain, in Romania, Greece, and Hungary a social system that was outside the modern world.

Fascism was an effort to employ anti-individualism and authoritarianism in modernizing economically backward societies. Arturo Labriola (1873–1959), an early syndicalist, spoke of Italy as a colony of "plutocratic Europe." The leader of an aggressive Italian nationalism, Enrico Corradini (1865–1931), influenced by Nietzsche and Maurras, saw the future as a conflict not between workers and capitalists but between proletarian and plutocratic nations. It was in that sense that fascism may have influenced the new African nations as they tried to organize themselves in the 1950s and 1960s. Despite its reactionary view of man, fascism regarded itself as representing youth against senility, the wave of the future against the effete heritage of the 19th century, biological vitality against the craving for peace and comfort.

THE ITALIAN EXPERIENCE

It has been shown that fascist movements arose where traditions and social structures favoured them. Unlike democracy, fascism demands a charismatic leader, a "new prince" as Machiavelli called him, who can gather all the prefascist emotional and social strains into one persuasive philosophy and will appeal to the masses. In Italy Benito Mussolini (1883–1945) was such a leader. From 1902 to 1914 Mussolini wrote many articles in the spirit of a radical Marxist. In this Marxist period of his life he became in 1912 editor of Italy's leading socialist daily, *Avanti!,* in Milan, a position of great influence for such a young man. He was then an extremist in his Marxist views, stressing revolutionary idealism and militant antimilitarism but deprecating dogmatism.

Dynamics of Italian Fascism

Whereas Marx's conception of history was based on 19th-century humanism and rationalism, Mussolini's was influenced by an early 20th-century emphasis on vitality and biological vital force, on a synthesis of Nietzsche and Sorel with Marx. This inclination to a heroic and active life-force made him, together with Gabriele D'Annunzio (1862–1938), the poet and glorifier of action, the chief propagandist of Italy's entrance into the war on the side of the Allies. It was over this issue that he broke with the Socialist Party, resigned from his editorship, and founded his own paper, *Il Popolo d'Italia.*

Mussolini's philosophy. Mussolini's philosophy, which developed slowly as his struggle for power and for a powerful state progressed, was officially presented in his article on the *Dottrina del fascismo* ("Doctrine of Fascism") in the *Enciclopedia Italiana* (1932). It reveals the pragmatic beginnings of the movement with complete frankness: "Our program is simple. We wish to govern Italy. They ask us for programs, but there are already too many. It is not programs that are wanting for the salvation of Italy but men and will power." By 1932 he had found a traditional philosophical garb for his vitalistic doctrine in the neo-Hegelian idealism of Giovanni Gentile (1875–1944), which saw the state as the source of all ethics and all individual life.

For Mussolini, all theoretical considerations were subservient to the "inexorable dynamics" of the factual situation. It is, he insisted, the role of the leader to master this dynamic process: he knows that the "iron logic of nature" will make the strong prevail over the weak. In contrast to Marxism, which asserts a rational logic of history that it claims will bring about the final triumph of the weak in an act of universal salvation, there is no fulfillment of history in Fascism. Instead, all history is incessant struggle, and the struggle itself is welcomed for its own ethical value. For

> war alone brings up to their highest tension all human energies and puts the stamp of nobility upon the peoples who have the courage to meet it. Fascism carries this antipacifist struggle into the lives of individuals. It is education for combat. . . . War is to the man what maternity is to the woman. I do not believe in perpetual peace; not only do I not believe in it but I find it depressing and a negation of all the fundamental virtues of man.

Triumph and decline of Italian Fascism. From 1922, when he first seized control of the Italian government in Rome, until 1927, Mussolini progressively consolidated his dictatorship, and the Fascist state took form. State and party became monolithic instruments in the hands of Mussolini, the *capo di governo* (head of the government) of the state, the *duce* (leader) of the party. The *legge fascistissime* (most Fascist laws), drafted by the leading Fascist jurist Alfredo Rocco, turned Parliament into a party congress, practically fused legislative and executive power, and made the Grand Council of Fascism an instrument of the duce, who alone could summon it and determine its agenda.

With this success Mussolini began to look farther into the future. Until 1930 he had emphasized the long overdue modernization of a backward country by a strong and efficient government and a new dedicated elite that could communicate a sense of vitality, virility, and energy to the whole people. But on October 25, 1932, Mussolini assured a Milan audience of the world leadership of fascist Italy. "Today, with a fully tranquil conscience, I say to you, that the twentieth century will be a century of fascism, the century of Italian power, the century during which Italy will become for the third time the leader of mankind." In 1934 Mussolini claimed that Fascism, an Italian movement in 1922, had "since 1929 become not merely an Italian phenomenon but a world phenomenon." To achieve his ends, Mussolini demanded the transformation of Italy into a *nazione militarista* and *guerriera.* In this goal he succeeded even less than in that of Italy's modernization.

The ambition of Italian Fascism

In his earlier stages Mussolini regarded Fascism as a development within Western civilization and looked with distrust generally upon Germany and specifically upon Hitler's National Socialism, which he recognized as "one hundred percent racism: Against everything and everyone: Yesterday against Christian civilization, today against Latin civilization, tomorrow, who knows, against the civilization of the whole world." But his imperialism and his overestimation of the power of Fascism drove him into the arms of National-Socialist Germany.

The conquest of Ethiopia (1935) opposed Italy to the West and the League of Nations. In the following year, dazzled by German success, Mussolini began to speak of a Rome-Berlin Axis and celebrated it during his visit in Berlin in 1936 and Hitler's return visit in May 1938. Hitler's sincere admiration for him as "the leading statesman in the world, to whom none may even remotely compare himself" increased Mussolini's self-delusion. The intervention for Fascism in Spain's Civil War and the Munich accord with England in September 1938 seemed a crowning success in Mussolini's policy; in reality they masked, with Italy's acceptance of German anti-Semitic laws, the fall of Italy to the position of Germany's satellite. As such, but also out of desire for glory and booty, Italy in June 1940 entered the war started by Germany in September 1939. The war revealed Italy's backwardness and inefficiency, until three years of warfare brought the fall of Mussolini and his party.

THE GERMAN EXPERIENCE

Hitler found a much better prepared soil in the antidemocratic traditions of the German *Reich* than Mussolini did in the tradition of the Italian Risorgimento, the 19th-century movement for Italian unity. Like Maurras and unlike Mussolini, Hitler had never been a Socialist or a man of the left. He was entirely unknown even in 1919 when as an agitator for the *Reichswehr* (army) he attacked the Western and German democrats, moderate Socialism, and Russian Bolshevism. He was, unlike Mussolini, the man of the one idea, which in him assumed demonic or maniac dimensions, but he was in tune with older currents of German thought and emotionalism.

Hitler's racial myths

Hitler's "Mein Kampf" and German anti-Semitism. Hitler grew up as a subject of the multi-ethnic Catholic Habsburg monarchy. He shared with the Austrian Georg von Schönerer (1842–1921) a virulent hatred for the Habsburgs, the "inferior" Slavs, for Catholicism as un-German, and above all for Jews and Judaism as non-Aryan. Hitler's propaganda methods were influenced by another Austrian, Karl Lueger (1844–1910) who, though a supporter of the dynasty and church, became a popular mayor of Vienna by opposing capitalistic liberalism and Marxian Socialism and appealing to the emotions of a lower middle class that felt threatened by both. In the restless years after 1918 Hitler took up this appeal. His anti-Semitism went far beyond Lueger's, becoming an obsession with him; the Jewish problem was no longer political, religious, or economic but the all-explaining theme of history. Hitler regarded the Nordic Aryans as the only creative race on earth, the only source of human greatness and progress. He believed that its end would mean the end of all civilization. Since he saw the German *Reich* as the highest expression of Aryanism, he proclaimed that it was necessary, not only for Germany but for the salvation of mankind, to secure the victorious survival of Germany by maintaining the purity of German "blood" against contamination by inferior races.

The total rejection of all miscegenation was one of the two fundamental pillars in Hitler's two-volume *Mein Kampf,* which he wrote in 1924 and 1926, and which became the bible of the new faith. The other was the absolute necessity of conquering a vast land base in eastern Europe, which was to become German by the ejection and enslavement of the "inferior" though white Slavic peoples. Germans would settle the immense and fertile plains and thus create a geopolitically unassailable *Reich.* The existing (and any future) Slav leadership class was to be exterminated to secure German domination. To these two fundamental ideas Hitler remained faithful from the very beginning of his agitation until his death. They had no parallel in Mussolini's writings, which glorified war and its spirit but presented no plan of a great aggressive war or any racial fanaticism. When Hitler dictated his last will shortly before his suicide, he repeated once more his fundamental interpretation of history: "Above all, I demand of the nation's leaders and followers scrupulous adherence to the race laws and to ruthless resistance against the world poisoners of all peoples, international Jewry," which he identified with the despised liberalism of Western capitalism and the socialism of Russian Marxism. His last words in 1945 expressed the same ideas that had guided him in 1919.

Hitler's program

Hitler's persuasion of the Germans. From the beginning, Hitler served the fascist movement by his understanding of the potential of the spoken word and the psychology of the masses. His appeal to the Germans as the most exalted race in the world counteracted the disillusionment and inferiority complex of a people believing itself surrounded by a hostile world. Hitler wrote in *Mein Kampf* that all propaganda must hold its intellectual level at the capacity of the least intelligent of those at whom it is directed and that its truth is less important than its success. "The slighter its scientific ballast, and the more exclusively it considers mass emotions, the more complete will be its success." In 1937, moreover, at the ninth party congress in Nürnberg, Hitler declared that "Germany has experienced the greatest revolution in the national and racial hygiene which was undertaken for the first time on an organized basis in the country. The consequences of this German race policy will be more decisive for the future of our people than the effects of any other laws. For they are creating the new man."

Many Germans believed in the reality and the superiority of this new man. Thus the racial interpretation of history and the fascist contempt for democracy lured Germany into war against Communism and democracy at the same time. By 1942 Germany had challenged the whole world and seemed at that point to have a good chance of emerging victorious from this total ideological war. Three years later it collapsed.

OTHER FASCIST MOVEMENTS IN EUROPE

The defeat of German fascism sealed the future of many other fascist movements that had come to power or grown in importance in many European countries partly with Germany's help or protection. In some of them, radical revolutionary fascism, eager for the modernization of the country, found itself in conflict with the authoritarian, semifeudal structure with which it often made common cause against Western liberalism, represented by a generally weak domestic middle class, and against an often exaggerated threat of Bolshevism from the outside. Yet reactionary authoritarianism and fascism fused in different ways in different countries, and with Italy's and Germany's defeat, the merely authoritarian reactionary regimes survived more easily than did the outright fascist ones.

The Spanish Falange

In most European countries there were a number of competing small fascist parties with no strong leader. Some of these came to power by National Socialist military success. In other cases (Britain, Switzerland, Sweden, or Denmark) the liberal parliamentary forces proved to be strong enough to keep the fascist movements within narrow bounds, and in others reactionary elements were able to use fascist movements as their support. The only one of these movements that could claim world attention on the international scene was the originally very radical Falange Española under the leadership of José Antonio Primo de Rivera (1903–36). The Spanish republican regime was established in April 1931, and Rivera was elected a deputy of the right in 1933. But in the next year he broke with it and united the Falange with the Juntas de Ofensiva Nacional Sindicalistas (Committees of Nationalist Syndicalist Offensive), "a movement steeped in true Spanish frenzy, launched by the young and dedicated to combatting . . . the irresponsible hypocrisy of the bourgeoisie" (Eugen Weber, *Varieties of Fascism,* D. Van Nostrand Co., Inc., Princeton, New Jersey, 1964, p. 117). The Falange was ultranationalist and eager for a thorough reform of Spain's antiquated social order. But in the election of 1936, won by the popular front of leftist moderate and radical parties, the Falange was unable to elect even a single deputy.

When civil war broke out in Spain in 1936, the republican government outlawed the Falange, which sided with General Francisco Franco; and in 1937 Franco united it with the military formations of the deeply reactionary Catholic Carlists, the Requetés, and made it an instrument of his personal leadership. But whereas in Italy and Germany fascism had absorbed the state, in Spain the victorious conservative oligarchy absorbed fascism.

In a similar way the authoritarian and traditionalist oligarchy in Antônio de Oliveira Salazar's Portugal kept fascist movements within very narrow limits, while using some of Mussolini's conservative slogans, as did the clerical semi-fascism in Austria under the two chancellors Engelbert Dollfuss (assassinated by the National Socialists in 1934) and Kurt von Schuschnigg and the Slovakian independence movement under Father Josef Tiso.

Fascism in the Balkans

Radical fascist movements developed in some Balkan countries—most prominently in Croatia, Hungary, and Romania. They shared with the conservatives the bitter hostility to Marxism and the Soviet Union, but they were obsessed by an extremist spirit of terroristic violence in a strange union with religious fanaticism. As a result of German victory the Croatian Ustaše, a party under the leadership of Ante Pavelić (1889–1959), turned Croatia into a state on the model of the most extremist National Socialist Party formation, the SS, or Schutzstaffel, into which only the most dedicated and racially pure Germans were admitted. The Ustaše persecuted and killed hundreds of thousands of Orthodox Serbs and Jews. Catholic monks and other priests are alleged to have taken an active part in this struggle for the "purity" of the Croatian land and faith, and the Ustaše envisaged the revival of the Great Croatian kingdom as it had existed under Peter Krešimir (1058–74) and Zvonimir (1076–89). The dream was destroyed first by Fascist Italy's occupation of Croatian Dalmatia and of Slovenia and finally by the collapse of Italy and Germany.

Fascism ruled in Croatia only four years; an even more violent fascism disturbed the political life of Romania for almost 25 years. Independent of the rise of Italian or German fascism, the fascism of Corneliu Zelea Codreanu was rooted in older traditions of the Romanian Orthodox Church and the Romanian peasantry. As a student at the University of Iași, Codreanu organized fellow students, sons of poor priests and peasants, into a National Christian Anti-Semitic League, which indulged in murder and a strange fanatical morality. In 1927 the movement was reorganized as a Legion of the Archangel Michael, characterized not by a flag but by an icon and by its members' dedication to a frugal ascetic life. The Iron Guard, as the Legion called its armed branch, represented an attempt at a fundamental reform of Romanian life. But the attempt to build a new life was made on the basis of wild blood sacrifices, one of whose victims in 1938 was Codreanu himself. In 1940, after the abdication of the King, a civil war within a coalition of the conservative Army under Gen. Ion Antonescu and the Legion under Horia Sima, Codreanu's successor, ended with the Army crushing the Legion in January 1941. Antonescu's Romania actively participated in Germany's war against Russia, but the Romanian fascist movement remained, after orgies of death, crushed, and its surviving leaders found refuge in National Socialist Germany.

In a similar way the fate of fascism and of reactionary circles was intertwined in Hungary. The Hungarian government under the regency of Miklós Horthy (1868–1957), the last commanding admiral of the Austro-Hungarian Navy, dreamed of the restoration of the former nine-centuries-old Great Hungarian realm. The suppression of the short-lived communist regime in Budapest in 1919 combined with the fascist hatred of Bolshevism to produce what was, for that time, an unprecedented "white" terror. But Horthy himself was a moderate conservative, and even when Hungarian extreme nationalism led him to form a profascist government under Gyula Gömbös in October 1932, Horthy followed a moderate course. Gömbös died suddenly in October 1936. In September 1940 the Hungarian National Socialist Party merged with the Arrow Cross Party and found its leader in Ferenc Szálasi, who had been a capable general staff officer in the army but had developed a fanatical racial faith in Hungarism that in some ways recalled Hitler and Codreanu. After Hungary's entry into the war, the Germans in 1944 occupied Hungary and interned Horthy. Szálasi was finally the head of a state, which he planned as the "Corporatist Order of the Working Nation." But the war was soon lost after a bloody winter of massacres of Jews and political opponents. Szálasi was executed; Horthy found refuge in Portugal.

Fascism elsewhere in Europe

Fascism outside Italy and Germany suffered from the lack of charismatic leaders who could embody a nationalist myth and yet be adroit political tacticians. In Nordic Germanic Europe, which, according to racial doctrine, should have shown a strong penchant for National Socialism, except for Norway few sympathized with German fascism—much fewer than among the eastern and southern Europeans. In Norway the great writer Knut Hamsun (1859–1952) and Vidkun Abraham Quisling (1887–1945), former minister of war in the farmers' party administration and son of a rural clergyman, took a pro-Nazi stand based on their opposition to Western "plutocracy" and their fascination with peasant rootedness, the opposition of the "natural" countryside to the "corruption" of urban civilization. Quisling founded, in 1933, the Nasjonal Samling, with the cross of St. Olaf as a symbol and the restoration of the greatness of Viking Norway as a goal. Under the German occupation Quisling became prime minister, but his following remained insignificant in his native land.

Stronger than in northern Europe was the fascist movement in Belgium. This was partly due to the fact that the Belgian nation was divided into a dominant French-speaking minority and a Flemish-speaking majority that struggled for complete equality. In 1931 Flemish extremists formed the Verbond van Dietsche Nationalsolidaristen, the Union of Dutch National Solidarists, who dreamed first of a Great Netherlands, later of a Great Burgundy, both times avoiding a close cooperation with German National Socialism. A more important fascist group originated in French-speaking Belgium and found in Lêon Degrelle an energetic demagogue and leader who strove for an agreement with the Flemish and showed understanding for modern social developments. Degrelle had grown up as a Catholic and in his young years had been influenced by Charles Maurras. He called his organization Rex and virulently attacked the parliamentary system. After a rapid growth his party lost its influence and Degrelle himself was defeated in a by-election in 1937 by a typical representative of the Belgian upper middle class with the approval of the Catholic primate. Degrelle's Flemish adherents joined the radical wing of the Flemish nationalists, the Vlaamsche National Verbond, and Degrelle formed a Walloon legion for the German SS and fought in the war against Russia. The confused character of the Rexists was symbolized by its red flag with a crown and a cross on it. In Belgium, as in Switzerland and The Netherlands, the parliamentary system of the conservative middle class was too strong to allow the success of a fascist party, and any threat of communism was too remote and improbable. In these countries terroristic violence so characteristic of the regimes of Mussolini, Hitler, Codreanu, or Szálasi remained unknown.

Of a very different nature was the short-lived fascist movement in England, where Sir Oswald Mosley, a member of the aristocracy, formerly a member of the Conservative Party, and later a member of the Labour government, founded in 1932 under the impact of the economic crisis the British Union of Fascists. The Union's plan was to replace the "old gang" of politicians by a new young elite, in tune with the new time. The paramilitary form of his party meetings and its "defence force" ran so deeply counter to Britain's centuries-old civilian tradition that in spite of Mosley's undeniable rhetorical and intellectual gifts the English masses resisted him, and the strong British Conservative government of the period offered little foundation for arousing fear of a domestic "Bolshevik" threat. The Conservative government was intelligent enough to forbid, after some street fighting in 1936, all paramilitary uniforms. By 1938 Mosley, perhaps the most intelligent and rational of all fascist leaders, ceased to be a figure of public importance.

The picture offered by France in the critical years was more complex. But there, too, fascist movements were diffuse and short-lived. One of the followers of Charles Maurras, George Valois, disappointed with Maurras's lack of active interest in social reforms, founded in 1925 the Faisceau, composed largely of war veterans; but only the crisis of the 1930s produced a number of groups, all nationalist, anticapitalist, and anti-Marxist. The largest of them, the Croix de Feu (Fiery Cross) of Col. François de La Rocque, rejected dictatorship and most of the fascist rhetoric. Characteristic for French fascism were the "proletarian" anti-Communists, led by former Socialists and Communists like Marcel Déat, the founder of the Rassemblement National Populaire, and Jacques Doriot, the former Communist mayor of Saint-Denis, a Parisian "red" suburb, who founded the Parti Populaire Français. More than in other countries, these French fascist groups were joined by some intellectuals who were disenchanted with the apathy of French life and yearned for revolution and action. The most prominent of them, Pierre Drieu La Rochelle (1893–1945), summed up the nihilism that is fundamental to all fascism when he wrote in 1934: "Liberty is exhausted. Man must seek new strength in his black basic nature. I say it, an intellectual, the eternal libertarian." (From *Socialisme Fasciste,* Gallimard, Paris, 1934, p. 102.)

FASCIST MOVEMENTS OUTSIDE EUROPE

Reaction and fascism in Japan

Like nationalism, socialism, and Communism, fascism was a European movement. But in the 1930s, when economic crisis seemed to reveal weaknesses of the liberal tradition, in general, fascism spread to Asia and Latin America, adapting itself to the social conditions of the countries there. Only in two cases did it become significant and assume an official role. The first instance was Japan, where fascism resembled German National Socialism in its reassertion of ancient models of life. On February 26, 1936, the army took control of Japan and supported a national or tribal mysticism that bore a close resemblance to that of Germany. Young "patriotic" officers tried to assassinate a number of leading Japanese statesmen, aristocrats, and high officers who seemed to them to represent the influence of foreign "dangerous thought," of the West, of liberalism, and of individualism, which threatened the traditional military spirit and absolute dedication to the cult of the emperor. In a number of cases the young officers were able to kill their victims. "The massacre was immensely popular in the army. The army acted as though the revolt was the work of the whole body and had succeeded. In its new orders the army said that it could not tolerate liberalism, that internationalism and individualism must be banished, and nationalism and the Japanese principle be promoted." (A. Morgan Young, *Imperial Japan 1926–1938,* William Morrow and Co., Inc., New York, 1941, p. 34.)

Different from other fascist countries, Japan saw the embodiment of its national ideal not in a popular leader but in the emperor. The national destiny was to be fulfilled by observing the duty to the throne and thus attaining the highest pinnacle of morality. The imperial will was to fix standards of justice and injustice, of right and wrong. Philosophy was regarded as good only when it was in conformity with the imperial will. Except for Germany, Japan was then the only nation that thought itself strong enough to extend its national ideal over the five continents. In the late 1930s Japanese professors and writers, as did their German colleagues, demanded the rejection of rational and universal ethics and the return to the ancient tribal gods in order to make the nation the most perfect instrument for its mission of conquest. The war in China was "not presented as one of conquest and exploitation but rather as a holy crusade to rid the land of unjust rulers [Chiang Kai-shek, red communists] and inaugurate there a regime of peace, righteousness and prosperity." (Harley F. MacNair, *The Real Conflict Between China and Japan,* The University of Chicago Press, Chicago, 1938, p. 193.) Later Japan spoke of a coprosperity sphere of a new Japan-led Asia.

Perón's fascistic populism

Less ambitious were the aspirations of fascism in Argentina, where the movement resembled that of Italy rather than of Germany. The initiative in Argentina came, however, not from former socialists and syndicalists but from officers, the Grupo de Oficiales Unidos (Group of United Officers), who seized power in 1943. The initials of the group's name stood for *gobierno* (authoritarian government), *orden* (order), and *unidad* (national unity). The officers believed that as a result of its relative wealth and its ethnically pure Caucasian population, Argentina was to assume the leadership in the struggle against "Yankee imperialism" and in the modernization of the continent. Among the officers, Juan Perón assumed a leading role and turned the movement into one of national socialism. With the help of his wife, Eva María (1919–52), he sought the support of the poorer masses, the *descamisados,* or shirtless ones. His attack on the traditional ruling oligarchy brought him wide popular support, so that he was twice elected president of Argentina, in 1946 and 1951. Perón organized the workers into a Confederación General del Trabajo, which was devoted to him, and claimed to have replaced "plutodemocracy." He created a mass party standing for "justicialism," a middle way between Communism and capitalism. But his own inconsistent policies and the accusation of widespread graft provoked a revolt of the armed forces in September 1955, as a result of which Perón left the country and settled in Spain.

By the latter years of the 20th century fascism seemed to have lost the attractions it had exercised in the 1920s and 1930s. After the fall of fascism in Germany and Japan both nations experienced a wave of great prosperity, which weakened, especially in the young generation, the formerly strong appeal of a militant nationalism. The conquest of a "living space," which Germany, Japan and Italy before World War II thought indispensable for the growth of their national economy, not only revealed itself as unachievable even to a broad anticommunist coalition, such as the three leading fascist powers established in 1937, but as a superfluous fancy, for the loss of empire brought misery neither to Britain nor Germany, The Netherlands or Italy, France or Japan. An unexpected industrial and agrarian productivity raised the living standards of the masses in the democracies, which could now concentrate upon necessary domestic reforms and abandon the lure of military glory. By the late 20th century fascism seemed a trend characteristic of the recent and yet faraway past. Except for small marginal movements, most elements were seeking to secure what fascism had denied, the cooperation of peoples of various civilizations and ideologies and the condemnation of war.

(H.K.)

Nationalism

Nationalism may be defined as a state of mind in which the individual feels that everyone owes his supreme secular loyalty to the nation-state. Nationalism is a modern movement. Throughout history men have been attached to their native soil, to the traditions of their parents, and to established territorial authorities; but it was not until the end of the 18th century that nationalism began to be a generally recognized sentiment molding public and private life and one of the great, if not the greatest, single determining factors of modern history. Because of its dynamic vitality and its all-pervading character, nationalism is often thought to be very old; sometimes it is mistakenly regarded as a permanent factor in political behaviour. Actually, the American and French revolutions may be regarded as its first powerful manifestations. After penetrating the new countries of Latin America it spread in the early 19th century to central Europe and from there, toward the middle of the century, to eastern and southeastern Europe. At the beginning of the 20th century nationalism flowered in the ancient lands of Asia and Africa. Thus the 19th century has been called the age of nationalism in Europe, while the 20th century has witnessed the rise and struggle of powerful national movements throughout Asia and Africa.

Identification of state and people

Nationalism, translated into world politics, implies the identification of the state or nation with the people—or at least the desirability of determining the extent of the state according to ethnographic principles. In the age of nationalism, but only in the age of nationalism, the principle was generally recognized that each nationality should form a state—its state—and that the state should include all members of that nationality. Formerly states, or territories under one administration, were not delineated by nationality. Men did not give their loyalty to the nation-state but to other, different forms of political organization: the city-state, the feudal fief and its lord, the dynastic state, the religious group, or the sect. The nation-state was nonexistent during the greater part of history, and for a very long time it was not even regarded as an ideal. In the first 15 centuries of the Christian Era, the ideal was the universal world-state, not loyalty to any separate political entity. The Roman Empire had set the great example, which survived not only in the Holy Roman Empire of the Middle Ages but also in the concept of the *res publica christiana* ("Christian republic" or community) and in its later secularized form of a united world civilization.

As political allegiance, before the age of nationalism, was not determined by nationality, so civilization was not thought of as nationally determined. During the Middle Ages civilization was looked upon as determined religiously; for all the different nationalities of Christendom as well as for those of Islām there was but one civilization—Christian or Muslim—and but one language of culture—Latin (or Greek) or Arabic (or Persian). Later, in the periods of the Renaissance and of Classicism, it was the ancient Greek and Roman civilizations that became a universal norm, valid for all peoples and all times. Still later, French civilization was accepted throughout Europe as the valid civilization for educated people of all nationalities. It was only at the end of the 18th century that, for the first time, civilization was considered to be determined by nationality. It was then that the principle was put forward that a man could be educated only in his own mother tongue, not in languages of other civilizations and other times, whether they were classical languages or the literary creations of other peoples who had reached a high degree of civilization.

Cultural nationalism

From the end of the 18th century on, the nationalization of education and public life went hand in hand with the nationalization of states and political loyalties. Poets and scholars began to emphasize cultural nationalism first. They reformed the mother tongue, elevated it to the rank of a literary language, and delved deep into the national past. Thus they prepared the foundations for the political claims for national statehood soon to be raised by the people in whom they had kindled the spirit.

Before the 18th century there had been evidences of national feeling among certain groups at certain periods, especially in times of stress and conflict. The rise of national feeling to major political importance was encouraged by a number of complex developments: the creation of large, centralized states ruled by absolute monarchs who destroyed the old feudal allegiances; the secularization of life and of education, which fostered the vernacular languages and weakened the ties of church and sect; the growth of commerce, which demanded larger territorial units to allow scope for the dynamic spirit of the rising middle classes and their capitalistic enterprise. This large, unified territorial state, with its political and economic centralization, became imbued in the 18th century with a new spirit—an emotional fervour similar to that of religious movements in earlier periods. Under the influence of the new theories of the sovereignty of the people and the rights of man, the people replaced the king as the centre of the nation. No longer was the king the nation or the state; the state had become the people's state, a national state, a fatherland. State became identified with nation, as civilization became identified with national civilization.

That development ran counter to the conceptions that had dominated political thought for the preceding 2,000 years. Hitherto man had commonly stressed the general and the universal and had regarded unity as the desirable goal. Nationalism stressed the particular and parochial, the differences, and the national individualities. Those tendencies became more pronounced as nationalism developed. Its less attractive characteristics were not at first apparent. In the 17th and 18th centuries the common standards of Western civilization, the regard for the universally human, the faith in reason (one and the same everywhere) as well as in common sense, the survival of Christian and Stoic traditions—all of these were still too strong to allow nationalism to develop fully and to disrupt society. Thus nationalism in its beginning was thought to be compatible with cosmopolitan convictions and with a general love of mankind, especially in western Europe and North America.

EUROPEAN NATIONALISM

The first full manifestation of modern nationalism occurred in 17th-century England, in the Puritan revolution. England had become the leading nation in scientific spirit, in commercial enterprise, in political thought and activity. Swelled by an immense confidence in the new age, the English people felt upon their shoulders the mission of history, a sense that they were at a great turning point from which a new true reformation and a new liberty would start. In the English revolution an optimistic humanism merged with Calvinist ethics; the influence of the Old Testament gave form to the new nationalism by identifying the English people with ancient Israel.

English Puritanism and nationalism

The new message, carried by the new people not only for England but for all mankind, was expressed in the writings of John Milton, in whose famous vision the idea of liberty was seen spreading from Britain, "celebrated for endless ages as a soil most genial to the growth of liberty" to all the corners of the earth.

> Surrounded by congregated multitudes, I now imagine that . . . I behold the nations of the earth recovering that liberty which they so long had lost; and that the people of this island are . . . disseminating the blessings of civilization and freedom among cities, kingdoms and nations.

English nationalism then was thus much nearer to its religious matrix than later nationalisms that rose after secularization had made greater progress. The nationalism of the 18th century shared with it, however, its enthusiasm for liberty, its humanitarian character, its emphasis upon the individual and his rights and upon the human community as above all national divisions. The rise of English nationalism coincided with the rise of the English trading middle classes. It found its final expression in John Locke's political philosophy, and it was in that form that it influenced American and French nationalism in the following century.

American nationalism was a typical product of the 18th century. British settlers in North America were influenced partly by the traditions of the Puritan revolution and the ideas of Locke and partly by the new rational interpretation given to English liberty by contemporary French philosophers. American settlers became a nation engaged in a fight for liberty and individual rights. They based that fight on current political thought, especially as expressed by Thomas Jefferson and Thomas Paine. It was a liberal and humanitarian nationalism that regarded America as in the vanguard of mankind on its march to greater liberty, equality, and happiness for all. The ideas of the 18th century found their first political realization in the Declaration of Independence and in the birth of the American nation. Their deep influence was felt in the French Revolution.

Jean-Jacques Rousseau had prepared the soil for the growth of French nationalism by his stress on popular sovereignty and the general cooperation of all in forming the national will, and also by his regard for the common people as the true depository of civilization.

The nationalism of the French Revolution was more than that: it was the triumphant expression of a rational faith in common humanity and liberal progress. The famous slogan "liberty, equality, fraternity" and the Declaration of the Rights of Man and of the Citizen were thought valid not only for the French people but for all peoples. Individual liberty, human equality, fraternity of all peoples: these were the common cornerstones of all liberal and democratic nationalism. Under their inspiration new rituals were developed that partly took the place of the old religious feast days, rites, and ceremonies: festivals and flags, music and poetry, national holidays and patriotic sermons. In the most varied forms, nationalism permeated all manifestations of life. As in America, the rise of French nationalism produced a new phenomenon in the art of warfare: the nation in arms. In America and in France, citizen armies, untrained but filled with a new fervour, proved superior to highly trained professional armies that fought without the incentive of nationalism. The revolutionary French nationalism stressed free individual decision in the formation of nations. Nations were constituted by an act of self-determination of their members. The plebiscite became the instrument whereby the will of the nation was expressed. In America as well as in revolutionary France, nationalism meant the adherence to a universal progressive idea, looking toward a common future of freedom and equality, not toward a past characterized by authoritarianism and inequality.

Napoleon's armies spread the spirit of nationalism throughout Europe and even into the Near East, while at the same time, across the Atlantic, it aroused the Latin Americans. But Napoleon's yoke of conquest turned the nationalism of the Europeans against France. In Germany the struggle was led by writers and intellectuals, who rejected all the principles upon which the American and the French revolutions had been based as well as the liberal and humanitarian aspects of nationalism.

German nationalism began to stress instinct against reason; the power of historical tradition against rational attempts at progress and a more just order; the historical differences between nations rather than their common aspirations. The French Revolution, liberalism, and equality were regarded as a brief aberration, against which the eternal foundations of societal order would prevail.

The 1848 revolutionary wave

That German interpretation was shown to be false by the developments of the 19th century. Liberal nationalism reasserted itself and affected more and more people: the rising middle class and the new proletariat. The revolutionary wave of 1848, the year of "the spring of the peoples," seemed to realize the hopes of nationalists such as Giuseppe Mazzini, who had devoted his life to the unification of the Italian nation by democratic means and to the brotherhood of all free nations. Though his immediate hopes were disappointed, the 12 years from 1859 to 1871 brought the unification of Italy and Romania, both with the help of Napoleon III, and of Germany; at the same time the 1860s saw great progress in liberalism, even in Russia and Spain. The victorious trend of liberal nationalism, however, was reversed in Germany by Bismarck. He unified Germany on a conservative and authoritarian basis and defeated German liberalism. The German annexation of Alsace-Lorraine against the will of the inhabitants was contrary to the idea of nationalism as based upon the free will of man. The people of Alsace-Lorraine were held to be German by objective factors, by race, independent of their will or of their allegiance to any nationality of their choice.

In the second half of the 19th century, nationalism disintegrated the supranational states of the Habsburgs and the Ottoman sultans, both of which were based upon prenational loyalties. In Russia, the penetration of nationalism produced two opposing schools of thought. Some nationalists proposed a westernized Russia, associated with the progressive, liberal forces of the rest of Europe. Others stressed the distinctive character of Russia and Russianism, its independent and different destiny based upon its autocratic and orthodox past. These Slavophiles, similar to and influenced by German romantic thinkers, saw Russia as a future saviour of a West undermined by liberalism and the heritage of the American and French revolutions.

One of the consequences of World War I was the triumph of nationalism in central and eastern Europe. From the ruins of the Habsburg and Romanov empires emerged the new nation-states of Austria, Hungary, Czechoslovakia, Poland, Yugoslavia and Romania. Those states in turn, however, were to be strained and ravaged by their own internal nationality conflicts and by nationalistic disputes over territory with their neighbours.

Russian nationalism was in part suppressed after Lenin's victory in 1917, when the Bolsheviks took over the old empire of the tsars. But the Bolsheviks also claimed the leadership of the world Communist movement, which was to become an instrument of the national policies of the Russians. During World War II Stalin appealed to nationalism and patriotism in rallying the Russians against foreign invaders. After the war he found nationalism one of the strongest obstacles to the expansion of Soviet power in eastern Europe. National communism, as it was called, became a divisive force in the Soviet bloc. In 1948 Tito, the Communist leader of Yugoslavia, was denounced by Moscow as a nationalist and a renegade; nationalism was a strong factor in the rebellious movements in Poland and Hungary in the fall of 1956; and subsequently its influence was also felt in Romania and Czechoslovakia and again in Poland in 1980.

ASIAN AND AFRICAN NATIONALISM

Nationalism began to appear in Asia and Africa after World War I. It produced such leaders as Kemal Atatürk in Turkey, Sa'd Pasha Zaghūl in Egypt, Ibn Sa'ūd in the Arabian peninsula, Mahatma Gandhi in India, and Sun Yat-sen in China. Atatürk succeeded in replacing the medieval structure of the Islāmic monarchy with a revitalized and modernized secular republic in 1923. Demands for Arab unity were frustrated in Africa and Asia by British imperialism and in Africa by French imperialism. Yet Britain may have shown a gift for accommodation with the new forces by helping to create an independent Egypt (1922; completely, 1936) and Iraq (1932) and displayed a similar spirit in India, where the Indian National Congress, founded in 1885 to promote a liberal nationalism inspired by the British model, became more radical after 1918. Japan, influenced by Germany, used modern industrial techniques in the service of a more authoritarian nationalism.

The new nations

The progress of nationalism in Asia and Africa is reflected in the histories of the League of Nations after World War I and of the United Nations after World War II. The Treaty of Versailles, which provided for the constitution of the League of Nations, also reduced the empires of the defeated Central Powers, mainly Germany and Turkey. The league distributed Germany's African colonies as mandates to Great Britain, France, Belgium, and South Africa, and its Pacific possessions to Japan, Australia, and New Zealand under various classifications according to their expectations of achieving independence. Among the League's original members, there were only five Asian countries (China, India, Japan, Thailand, and Iran) and two African countries (Liberia and South Africa), and it added only three Asian countries (Afghanistan, Iraq, and Turkey) and two African countries (Egypt and Ethiopia) before it was dissolved in 1946. Of the mandated territories under the League's control, only Iraq, Lebanon, and Syria achieved independence during its lifetime.

Of the original 51 members of the United Nations in 1945, eight were Asian (China, India, Iraq, Iran, Lebanon, Saudi Arabia, Syria, and Turkey) and four were African (the same as in the League). By 1980, 35 years after its founding, the United Nations had added more than 100 member nations, most of them Asian and African. Whereas Asian and African nations had never totalled even one-third of the membership in the League, they came to represent more than one-half of the membership of the United Nations. Of these new Asian and African nations, several had been created, entirely or in part, from mandated territories.

After World War II, India, Pakistan, Ceylon (Sri Lanka), Burma, and Malaya (Malaysia) in Asia, and Ghana in Africa achieved independence peacefully from the British Commonwealth, as did the Philippines from the United States. Other territories had to fight hard for their independence in bitter colonial wars, as in French Indochina (Vietnam, Laos, Cambodia) and French North Africa (Tunisia, Algeria). Communism recruited supporters from within the ranks of the new nationalist movements in Asia and Africa, first by helping them in their struggles against Western capitalist powers, and later, after independence was achieved, by competing with Western capitalism in extending financial and technical aid. Chinese nationalism under Chiang Kai-shek during World War II was diminished with the takeover of the Chinese Communists. But Chinese Communism soon began to drift away from supranational Communism, as the European Communist countries had earlier. By the late 1960s Russian and Chinese mutual recriminations revealed a Chinese nationalism in which Mao Tse-tung had risen to share the place of honour with Lenin. As Chinese Communism turned further and further inward, its influence on new Asian and African nations waned.

Political and religious differences

Ambitions among new Asian and African nations clashed. The complex politics of the United Nations illustrated the problems of the new nationalism. The struggle with Dutch colonialism that brought the establishment of Indonesia continued with the UN mediation of the dispute over West Irian (Irian Jaya). In the Suez crisis of 1956, UN forces intervened between those of Egypt and Israel. Continuing troubles in the Middle East, beginning with the establishment of Israel and including inter-Arab state disputes brought on by the establishment of the United Arab Republic, concerned the UN. Other crises involving the UN included: the India-Pakistan dispute over Jammu and Kashmir; the Korean partition and subsequent war; the four-year intervention in the Congo; the struggle of Greece and Turkey over newly independent Cyprus; and Indonesian and Philippine objection to the inclusion of Sarawak and Sabah (North Borneo) in newly formed Malaysia.

Many new nations, all sharing the same pride in independence, faced difficulties. As a result of inadequate preparation for self-rule, the first five years of independence in the Congo passed with no semblance of a stable government. The problem of widely different peoples and languages was exemplified in Nigeria, where an uncounted population included an uncounted number of tribes (at least 150, with three major divisions) that used an uncounted number of languages (more than 100 language and dialect clusters). The question of whether the predominantly Muslim state of Jammu and Kashmir should go with Muslim Pakistan or Hindu India lasted for more than 20 years after the India Independence Act became effective in 1949. Desperate economic competition caused trouble, as in Israel where the much-needed waters of the Jordan River kept it in constant dispute with its water-hungry Arab neighbours.

In Europe the spirit of nationalism appeared to wane after World War II with the establishment of international economic, military, and political organizations such as NATO, the European Coal and Steel Community, Euratom, and the Common Market. But the policies pursued by France under Pres. Charles de Gaulle and the problem of a divided Germany showed that the appeal of the nation-state was still very much alive. (H.K./Ed.)

Liberalism

Liberalism does not lend itself to easy definition. A major difficulty is that, with some exceptions, liberals have shunned dogma, preferring generally a pragmatic to a doctrinaire approach to social problems. Another, which has been a prolific source of misunderstanding, has been liberals' own frequently opposing views concerning the scope of government. The confusion thus engendered is sometimes compounded by a tendency to identify liberalism exclusively with its 18th- and 19th-century variant, or with the program of this or that liberal party, in a formulation that has on occasion led many to announce the "decline" or "end" of liberalism and to compose obituaries that have been quite misleading. Through the centuries liberalism has changed drastically in content, but it has maintained a constant form. Those who note the first and neglect the second understandably find the term confusing and its application inconsistent.

HISTORICAL BACKGROUND

Liberalism is the culmination of a development that goes back to the Hebrew prophets, the teachings of the pre-Socratic philosophers, and the Sermon on the Mount, from all of which there emerged a sense of the importance of human individuality, a liberation of the individual from complete subservience to the group, and a relaxation of the tight hold of custom, law, and authority.

Throughout much of his history, man as an individual has been submerged in his group. His emancipation as an individual can be understood as a unique achievement of Western culture, perhaps its very hallmark. If this be so, then the emergence of liberalism was, in an important sense, inseparable from Western man's quest for freedom; for liberalism, in the broadest sense, seeks to protect the individual from arbitrary external restraints that prevent the full realization of his potentialities.

Mercantilism and the rise of the middle class

Medieval society did not provide a soil in which the first seeds of liberalism might easily germinate. The Middle Ages produced a society of status in which the rights and responsibilities of the individual were determined by his place in a stratified, hierarchically ordered system. Such a closed, authoritarian order, however grandiose in outline and noble in aspiration, was bound to place great stress upon acquiescence and conformity. As new needs and interests, generated by the slow commercialization and urbanization of Europe, gained strength, the medieval system was modified to accommodate the ambitions of national rulers and the requirements of an expanding industry and commerce. The ensuing policies and arrangements came to be known as mercantilism, a policy of state intervention that, in theory at least, might be extended to regulation of the most minute details of economic life (*cf.* Eli Heckscher, *Mercantilism,* 1935). However, as such intervention came more and more to serve established interests and to inhibit enterprise, it was challenged by the members of the newly emerging middle class. The challenge took the form of revolt, first against the Universal Church, and later against mercantilist states, presided over by absolute monarchs. The former manifested itself in the Protestant Reformation and the quest of Calvinists and Calvinist sects for freedom of conscience; the latter in the great revolutions that rocked England and France in the 17th and 18th centuries, notably the Glorious Revolution of 1688, the French Revolution a century later, and the successful revolt of England's American colonies. Classical liberalism as an articulated creed is a product of those great collisions.

The fortunes of liberalism differ with the historical conditions in each country—with the strength of the crown, the élan of the aristocracy, the pace of industrialization, and the circumstances of national unification. Thus, by contrast with England, the character of liberalism in France reflected the decadence of its nobility and the absolutism of the Bourbons. The failure of liberalism in Germany in the 19th century was attributable in great part to the dominant role of a militarized and Lutheran Prussia and the reactionary influence of Austria. The advent of liberalism in Italy was delayed by the armies of Austria and of Louis-Napoleon and the opposition of the Vatican. Whatever the variations, the liberal impact on authoritarianism reverberated throughout Europe and its dominions, voiced by a Kossuth in Hungary, a Mazzini in Italy, a Thorbecke in The Netherlands, a Bolívar in South America. For a moment even Russia, in 1905 and again between March and November in 1917, heard the echo and, had there been a sufficiently numerous middle class to listen, the course of modern history would no doubt have been changed.

ECONOMIC AND POLITICAL FOUNDATIONS

The authors of the liberal creed differ widely, even in the countries in which liberalism was cradled. But their agreements sufficiently exceed their differences to permit their being included in the same tradition—a tradition whose main manifestations are both economic and political. The fact that the classical liberals, perhaps more perceptively than their successors, regarded these as only abstractly separable is indicated in the very title of their science—political economy.

"Private vices, public benefits"

The economics of the "free" market. On the economic side 18th- and 19th-century liberalism based itself on the sovereignty of the market and the "natural harmony of interests." On this view, if individuals are left free to pursue their self-interest in an exchange economy based upon a division of labour, the welfare of the group as a whole will necessarily be enhanced. Classical liberal economists describe a self-adjusting market mechanism free from all teleological influences. While moral goals are invoked and ethical criteria presupposed in passing ultimate judgment on the system, they play no part in determining the sequence of events within it. The one propelling force is the selfishness of the individual, which becomes harnessed to the public good because in an exchange economy he must serve others in order to serve himself. It is only in a *free* market, however, that this happy consequence can ensue; any other arrangement must lead to regimentation, exploitation, and economic stagnation. The most celebrated formulation of this doctrine is to be found in Adam Smith's *Wealth of Nations:*

> He generally, indeed, neither intends to promote the public interest, nor knows how much he is promoting it.... by directing... industry in such a manner as its produce may be of the greatest value, he intends only his own gain, and he is in this, as in many other cases, led by an invisible hand to promote an end which was no part of his intention. Nor is it always the worse for the society that it was no part of it. By pursuing his own interest he frequently promotes that of the society more effectually than when he really intends to promote it. I have never known much good done by those who affected to trade for the public good. It is an affectation, indeed, not very common among merchants, and very few words need be employed in dissuading them from it.

The same thought had been expressed more tersely by Bernard de Mandeville (*c.* 1670–1733) in the phrase, "private vices publick benefits." A poet (Pope) said it best:

Thus God and Nature link'd the gen'ral frame,
And bade Self-love and Social be the same.

Every economic system must be concerned with at least two basic problems: first, arrangement must somehow be made for determining *what* shall be produced—*i.e.,* how "scarce means" shall be allocated—and, second, there must be some way of apportioning what is produced. In a controlled economy this is accomplished by a planning agency acting at the behest of the government. In the economy envisaged by the classical economists of the 18th and 19th centuries, and those conservative economists who today may be called neoclassicists, this is accomplished in the free market through the price mechanism. In such a market the theoretically free choices of individual buyers and sellers determine how the resources of society (labour, goods, capital) shall be employed. These choices manifest themselves in bids and offers that in their totality determine the price at which a commodity will sell. Theoretically, when the demand for a commodity is great, prices will rise, making it profitable for producers to increase the supply; as supply approximates demand, prices will tend to fall until producers divert productive resources to other uses. In this way the closest possible coincidence is said to be achieved between what is wanted and what is produced. Moreover, in the distribution of the wealth thereby produced, the system is asserted to assure a reward in proportion to merit. The assumption is that in a freely competitive economy in which no one is barred by status from engaging in economic activity, the income received from such activity is a fair measure of its value to society.

Man—the rational economic animal

Presupposed in the foregoing is a conception of man as an economic animal rationally engaged in minimizing costs and maximizing gains (wages, profit, interest). Egoistic and hedonistic assumptions about human nature, which were taken for granted, led easily to an emphasis on man as a forward-looking and end-seeking creature. "When matters of such importance as pain and pleasure are at stake... who is there that does not calculate," Bentham asked. In Pope's descriptive couplet, although

Self-love, the spring of motion, acts the soul;
Reason's comparing balance rules the whole.

If, as the Enlightenment liberals assumed, "reason's comparing balance" rarely fails man in any of his activities (Bentham went so far as to construct an entire theory of penology based on the assumption that even would-be lawbreakers carefully balance the pleasure to be derived from their contemplated crime against the pain of punishment), it could be affirmed *a fortiori* that they would also meticulously balance utilities against costs at the marketplace. Since rational men best know their own interest, it must follow that interference by agencies of government could only diminish that "greatest happiness of the greatest number" that the followers of Bentham claimed to desire.

Assumed also by the classical creed is a conception of the consumer as sovereign, decreeing by his purchases how the resources of society shall be allocated. No one has celebrated this coronation of the consumer more eloquently than Ludwig von Mises in his *Omnipotent Government:*

> Within the market society the working of the price mechanism makes the consumers supreme. They determine through the prices they pay and through the amount of their purchases both the quantity and quality of production. They determine directly the prices of consumers' goods, and thereby indirectly the price of all material factors of production and the wages of all hands employed.
>
> ... In that endless rotating mechanism [*i.e.,* a market society] the entrepreneurs and capitalists are the servants of the consumers. The consumers are the masters, to whose whims the entrepreneurs and capitalists must adjust their investments and methods of production. The market chooses the entrepreneurs and the capitalists and removes them as soon as they prove failures. The market is a democracy in which every penny gives a right to vote and where voting is repeated every day.

The justification of private property

Implied by the logic of this economic creed is a functional justification of private property, often buttressed, to be sure, by a doctrine of natural right to shield manifestly functionless claims to property. John Locke's *Two Treatises of Government* (1690) pointed the way by defining property as "whatsoever... [man] hath mixed his labour with...." Adam Smith found that "The property which every man has in his own labour... is the original foundation of all property...." And Bentham noted that "It is this right that has overcome the natural aversion to labour...." Since acquisitiveness and indolence were regarded as inborn and ineradicable human traits, security of property had to be preserved if incentive was not to be destroyed and the production of goods discouraged. Both French revolutionaries and English gentry could be rallied to such a defense of property; and the American Constitution as well as the French Declaration of the Rights of Man and of the Citizen, both of them liberal testaments, charge government with basic responsibility for its protection.

The role of government. On the political side the guiding principle of historical liberalism has been an undeviating insistence on limiting the power of government. Inspired by the need to remove the state from destructive interference with the economic life of the community, the principle degenerated, under the influence of men like Herbert Spencer, into a doctrinaire form of antistatism. Spencer would even forbid government, either local or national, to assume responsibility for the paving, lighting, and sanitation of cities. Even the less doctrinaire Bentham's sole advice to the state was to "Be quiet," and Edmund Burke—who in this context must be reckoned a liberal—declared that "It is in the power of the state to prevent much evil; it can do very little positive good...." Thomas Paine, eloquent pamphleteer for the American Revolution (of which Burke approved), wrote in his "Common Sense" that "Government, even in its best state, is but a necessary evil...," and generations of Americans almost ritualistically repeated Thomas Jefferson's advice that that government is best that governs least.

The prevailing view was perhaps best spelled out in Adam Smith's enumeration of government functions. Smith assigned the "sovereign" three tasks: to protect the group from outside violence; to protect individual members of society from the injustices or oppressions of their fellow citizens; and, finally, to erect and maintain "those public institutions and those public works, which, though they may be in the highest degree advantageous to a great society, are however of such a nature, that the profit could never repay the expense to any individual, or small number of individuals...." An agenda more faithful to Adam Smith's intent would provide a central place for the enforcement of contracts, without which voluntary individual arrangements could hardly replace government fiat in the conduct of economic affairs.

In general, liberals believed that government must not do for the individual what he is able to do for himself. Legislation like Britain's Ten Hours Act (1847), which limited the labour of women and children to 58 hours a week, was denounced by the English jurist A.V. Dicey as late as 1905 as socialistic. Criticizing the Adulteration of Food or Drink Act and the Sale of Food and Drugs Act of 1899, he contended that such laws safeguard individuals from mistakes "which often may be avoided by a man's own care and sagacity" and therefore "rest upon the idea... that the State is a better judge than a man himself of his own interest..." (*Law and Public Opinion in England,* 2nd ed.; 1914; pp. 237–238; 263–264). Such views were even more prevalent in the United States.

Separation of powers. The institutional devices by means of which liberals sought to limit government to the exercise of its proper functions were numerous: federalism (though not in France and Britain), bicameralism, the separation of powers. The last of these, that is, the distribution of power among such functionally differentiated agencies of government as the legislature, the executive, and the judiciary, and the system of checks and balances by which it was accomplished, was given its classic embodiment in the U.S. Constitution and its political justification in that brilliant apology for the Constitution *The Federalist* (1787–88), by Hamilton, Madison, and Jay. Montesquieu had already formulated such a doctrine in his famous *De l'esprit des loix* (1748; Eng. trans., *The Spirit of Laws,* 1750), receiving the idea from Locke, who had not developed it fully.

The Federalist

Such a separation of powers could have been achieved, of course, through a "mixed constitution," that is, by having a monarch, a hereditary chamber, and an elected assembly share in power with some appropriate differentiation of function. The Greek historian Polybius hailed such a mixed government as the glory of the Roman constitution, and Blackstone (misled perhaps by Montesquieu) in his famous *Commentaries on the Laws of England* (1765–69) extolled its embodiment in the constitution of England, as Burke was to do later with greater eloquence. But it was precisely despotic kings and functionless aristocrats (more functionless in France than in England) who thwarted the interests and ambitions of the middle class, which turned, therefore, to the principle of majoritarianism.

Liberal conflicts of interest. The greatest check on government is, of course, the threat of dismissal by its constituents. But in determining the crucial question of who the constituents are to be, classical liberalism fell victim to an ambivalence, torn between the great emancipating tendencies generated by the revolutions with which it was associated and middle-class fears that democracy would undermine private property. Most 18th- and 19th-century liberal spokesmen feared popular sovereignty, and for a long time suffrage was limited to property owners. In Britain even the important Reform Act of 1867 did not completely abolish property qualifications. France, for her part, bore revolutionaries but suckled reactionaries: although the Revolution of 1789 proclaimed the ideal of universal manhood suffrage and the Revolution of 1830 reaffirmed it, there were no more than some 200,000 qualified voters in a population of about 30,000,000 during the reign of Louis-Philippe, "citizen king" installed by the ascendant bourgeoisie in 1830. And, in the United States, Jefferson's brave language in the Declaration of Independence notwithstanding, it was not until 1860 that universal white male suffrage prevailed.

Although by the time of the Revolution he had changed his opinion, Benjamin Franklin spoke for the Whig liberalism of the Founding Fathers when he observed that "as to those who have not landed property the allowing them to vote is an impropriety." John Adams in his famous *Defense of the Constitutions of Government of the United States of America* (1787) was more explicit, finding that, if the majority were to control all branches of government, "Debts would be abolished first; taxes laid heavy on the rich, and not at all on others; and at last a downright equal division of everything be demanded and voted." Thomas Babington Macaulay spoke for English Whigs when he found universal suffrage "incompatible, not with this or that form of government, but with all forms of government." French statesmen like François Guizot (1787–1874) and Adolphe Thiers (1797–1877) expressed similar sentiments, and on the rest of the continent universal suffrage was for the most part a remote ideal until the turn of this century.

The fear of majority tyranny

Objections to democracy were not limited to misgivings about the fate of private property. Many liberals genuinely feared the potentialities for tyranny latent in democracy. If the will of the majority is to be supreme, everyone will be at its mercy. Benjamin Constant (1767–1830), noted French diplomat, expressed a general concern when he observed that from the point of view of the individual it makes no difference whether he is tyrannized by a single despot or by the totality of individuals composing his society; he is oppressed just the same. Indeed the latter could well be worse: tyrannicide might rescue him from an individual oppressor and, in any event, others would commiserate with him; whereas if oppressed by a large majority, he would have no comparable recourse or comfort.

Despite such misgivings, the voices of Thomas Paine and Jefferson, of Rousseau and de Tocqueville, and of the English philosophical radicals led by Bentham and James and John Stuart Mill finally prevailed among the liberal philosophers. But the democratic principle that they espoused had first to be reconciled with the liberal requirement that power be limited. The problem was to accomplish this in a manner consistent with the democratic ideal, that is, without doing violence to the majoritarian principle and in conformity with Bentham's dictum that each person is to count as one and no one as more than one. How, since hereditary elites were discredited, could the power of a majority be checked without giving greater voice to property owners or some other kind of "natural" elite?

The liberal formula. The essence of the liberal solution is twofold. It lies partly in making the decision of any given majority subject to the concurrence of other majorities distributed over a period of time. The majority that elects a president of the United States, for example, is different from the majority that two years before elected one-third of the Senate and two years later elects another third. Likewise two-thirds of the senators are elected by a majority other than the majority by which members of the House of Representatives are elected. These groups, in turn, are checked by the Constitution, which was approved and later amended by majorities no longer alive. While an act of Britain's Parliament is forthwith a part of the British constitution, Parliament before acting on a highly controversial issue of great concern to the country will seek a mandate from the people, that is to say, a majority other than the majority that elected it. Thus, the "people" to be consulted in a liberal democracy, if they are not quite that partnership envisaged by Burke "... between those who are living, those who are dead, and those who are to be born," are not either, except in a revolutionary crisis, a "momentary aggregation" whose sovereign will may be ascertained by a plebiscite. Burke's characterization of democracy as "the vulgar practice of the hour" may well apply to those spontaneous assemblies of the people to which Rousseau would assign plenary power; it can be denied that it applies to the mainstream version of liberal democracy that is essentially *constitutional* democracy, that is to say, democracy in which the power of a current majority is checked by the verdicts of majorities that preceded it.

The second part of the solution is related more directly to the initial inspiration of liberalism, its basic commitment to the autonomy and integrity of the individual, which the limitation of power is, after all, intended to preserve. In the liberal understanding, the individual is not only a citizen who shares a social compact with his fellows; he is a person and, as such, he possesses rights that the state may not invade. Even as a citizen, he must, if majoritarianism is to be meaningful, possess such rights. Majorities do not form in a vacuum. Unless a majority verdict is some miraculous coincidence or spontaneous merger of individual judgments—which it is not—it can come about only if individuals are free to some extent to formulate and express and exchange their views. This involves, beyond the right to speak and write freely, the freedom to associate and organize and, above all, freedom from fear of reprisal. But the individual has rights apart from his role as citizen. These are rights that secure his personal safety and hence his protection from arbitrary arrest and punishment. And beyond these are those rights that preserve large areas of *privacy*. In a liberal democracy there are affairs that do not concern the state, even if the state's interest in them reflects the will of an overwhelming majority. Such affairs may range from the worship of God, to works of art, to how parents raise their children. And for the liberals of the 18th and 19th centuries they included, above all, most of those activities in which individuals engage in production and trade.

Classical liberals differed in their interpretation of individual rights—whether to understand them as "natural" (Locke) and therefore "inalienable" and "indefeasible," to justify them as functional (Bentham) or traditional (Burke). But eloquent and persuasive declarations affirming such rights were embodied in the English Bill of Rights of 1689, the American Declaration of Independence and Constitution, the French Declaration of the Rights of Man and of the Citizen of 1789, and the basic documents of nations throughout the world that later used these as their models. Freedom thereby became more than the right to make a fractional contribution in an intermittent mandate to government; it designated the right of men to live their own lives.

The problems of the dispossessed

Rhetoric and reality rarely coincide. Liberals were not usually troubled over the small numbers by whom, in fact, such freedom was enjoyed, and the masses were still too inarticulate to remind them. In the dawn of the Industrial Revolution, the blessings of liberty must have seemed remote to millions of brutally exploited factory workers huddled together in the slums of their dreary cities. Liberals have been accused of concentrating exclusively on the property right to the neglect of such abuses, and, in general, few liberals paused to ponder how the freedom they cherished would be used, and it did not occur to them, as it has to a later generation of libertarians, that such freedom brings with it a heavy burden of choice and decision. Finally, in their concern for the individual, liberals depicted him too simply as an isolated monad, and society as a collection of such monads, failing to reckon with the myriad social relationships through which individuals achieve identity.

ACHIEVEMENT AND FAILURE

In historical perspective it can be seen that the complex of forces, of which classical liberalism was the rationalization, wrought great changes. The feudal system was destroyed. Capitalism replaced the static society of the Middle Ages. A functionless aristocracy was deprived of its privileges. Tyrants were challenged and curbed. The middle class was left free to employ its energies in expanding the means of production and vastly increasing the wealth of society. Representative government came into its own. As they set about limiting the sovereign power, liberals converted the ideal of constitutional government into a reality, and they developed a doctrine of individual rights, including the right to worship freely, the right of a free press, of free speech, and of free assembly, which lies at the heart of modern democracy. J.S. Mill's essay *On Liberty* (1859) is justly celebrated as one of the great testimonials to civil liberties.

However, vast economic changes, first in Great Britain and later in the United States, led increasingly at the turn of this century to disenchantment with the principal economic basis for liberalism—the ideal of a market economy. (This ideal had never evoked comparable loyalties on the continent of Europe.) The advent on a large scale of absentee ownership made it increasingly difficult to invoke a functional justification for many forms of private property. Multitudes whose real bargaining power fell far short of what it was in theory—especially those seeking work—did not actually enjoy the "free choices" postulated by the economists. Economic man's "pecuniary sagacity" (Veblen) fell far short of the idealizations of textbook writers; and modern psychology suggests that he is at least as much a creature of impulse, habit, and custom as he is a rational calculator with his eye on the main chance. The often impenetrable complexity of goods offered on the market in an economy where transactions are no longer simple affairs of horse-trading, not to mention the mendacity of much advertising, seemed to make the consumer more a subject than a sovereign. Reality failed to approximate the state of free competition envisaged by the classical economists. Orthodox liberal economists have accordingly referred with increasing frequency to "frictions" and "exceptions" and have employed qualifications ("in the long run"; "other things being equal") to save their generalizations. The result has been an abstract science that a new generation of liberals and multitudes of socialists came to regard as having only limited relevance to the real world.

Concentration of wealth and irrational economies

Worst of all, according to its critics, the profit system concentrated vast wealth in relatively few hands, with several decisively adverse consequences. First, great masses of people failed to benefit from the wealth flowing from mines and mills and lived under impoverished conditions that became increasingly anomalous in an affluent society. Second, since those who alone could consume the output of a vastly expanded productive system lacked the purchasing power, the system, after other markets were glutted, endemically failed to realize its productive potentialities and recurrently came to a near halt in periods of stagnation that have come to be called "depressions." Finally, those who owned the means of production, or their managers, were endowed with vast power that could be used to overwhelm the individual as surely as the power of a 17th-century despot, thanks to the device of incorporation on a scale far beyond the simple requirements of efficiency and economy in production. In short, the near atrophy of government achieved in the 19th century had left a vacuum that private interests readily filled. Businesses were effectively organized and some of them used their power to influence and control government, to manipulate an inchoate electorate, to limit competition, and to obstruct substantive social reform. Some of the same forces that had once released the productive energies of Western society now restrained them; some of the very energies that had demolished the power of despots now nourished a new despotism. Such, at any rate, was the verdict of 20th-century liberals and such were the conditions that led them to oppose private collectivism by supporting a positive role for government and encouraging the formation of power centres outside business and government.

CONTEMPORARY LIBERALISM

The contemporary variant of liberalism is even more amorphous than the classical. There are no "fathers" such as John Locke or Adam Smith. In Germany and the Scandinavian countries, where socialists, even though they trace their ancestry to Marx, are overwhelmingly revisionist, it is difficult to distinguish between their program and programs elsewhere designated as liberal. British mainstream socialism never embraced Marx, and the Labour Party has at times been difficult to distinguish from the Liberal Party, which largely accounts for the decline of the latter. U.S. liberal social legislation since the crash of 1929 has been notably formless. But even so, the outlines of contemporary liberalism are fairly discernible.

Critique of the market. Cognizant of the real achievements of the profit system, liberals do not seek its abolition, only its modification and control. They find no fixed line laid up in heaven eternally dividing the private and the public sectors of the economy; the determination, they contend, must be by reference to what works. The spectre of regimentation in completely planned economies and the dangers of bureaucracy even in mixed economies deter them from jettisoning the market and substituting an omnicompetent state. On the other hand—and this is a basic difference between classical (or neoclassical) and contemporary liberalism—most liberals now believe that the dispensations of the market, as it has in fact operated, must be supplemented and corrected in substantive ways. They hold that the rewards dispensed by the market are too crude a measure of the contribution many or most people make to society, and that the needs of those who lack opportunity or are physically handicapped are ignored. They contend that enormous social costs incurred in production are not reflected in market prices, and that resources are used wastefully. Not least, liberals charge that the market biasses the allocation of human and physical resources in the direction of satisfying superficial wants (for oversized motor cars in annual models, changing fashions in attire, and unnecessary gadgets), while basic needs (for schools, housing, rapid public transit, sewage treatment plants) go unmet. Finally, although liberals believe that prices, wages, and profits should continue to be subject to negotiation among the interested parties and responsive to conventional market pressures, they insist that price–wage–profit decisions *affecting the economy as a whole* must be reconciled with public policy.

The liberal program. To achieve a juster distribution of rewards, liberals have relied on two major strategies. First, they have promoted the organization of workers and consumers to improve their power to bargain with employer-producers. Such a redistribution of power has had political as well as economic consequences, making possible a party system in which at least one party is responsive to the interests of wage earners and consumers. Second, enlisting the political support of the economically deprived, liberals have evolved the so-called welfare state, with its panoply of social services "from the cradle to the grave."

Foundations of the welfare state

Social legislation, beginning with free public education and workmen's accident insurance, now includes support for all who are physically and mentally handicapped, programs of old-age, unemployment, and health insurance, minimum-wage laws, and—for the most part still in the blueprint stage—guaranteed annual incomes. Such legislation is most comprehensive in the Scandinavian countries and, among countries with mature economies, possibly least comprehensive in the United States, where social legislation at the federal level was virtually ignored until passage of the Social Security Act of 1935.

Liberals have been less successful in correcting what they regard as an irrational schedule of priorities by using the taxing powers of government to obtain greater allocations for the public sector of the economy. They have been least successful, at any rate in the United States, in persuading the business community and labour unions to accept government participation in pricing and wage decisions.

Initially, the quest for these objectives was viewed as requiring a redistribution of wealth to be achieved by steeply graduated income and death (inheritance) taxes—a program likely, if it were to become confiscatory, to evoke the unlimited resistance of high-income groups. Increasingly, as modern technology promised mature economies miracles of abundance, attention shifted to the institutional failures that kept such economies from fully realizing their productive potentialities, especially during periods of mass unemployment and depression. The culmination of this effort was the White Paper on Employment Policy produced by Britain's wartime coalition government and, in the United States, the Employment Act of 1946, which goes well beyond the British, declaring it to be "the continuing policy and responsibility of the Federal Government to use all practicable means . . . to promote maximum employment, production, and purchasing power." Thereafter the old rhetoric about "sharing the wealth" gave way to concentration on growth rates as liberals, inspired by J.M. Keynes' *General Theory of Employment, Interest and Money,* urged the use of fiscal policy—*i.e.,* use of the government's power to borrow, tax, and spend—not merely to counter contractions of the business cycle but to encourage expansion of the economy. Here, clearly, was a program less disruptive of class harmony and the basic consensus essential to a democracy than the old Robin Hood method of "taking from the rich and giving to the poor."

Industrialism reconsidered

In the 1960s the emphasis of liberals again shifted as it became more and more evident that expanding production is not an unmixed blessing. The litany is now a familiar one. The same industry that produces our wealth pollutes our rivers, lakes, and atmosphere. Its gleaming artifacts become mountains of junk and it menaces the ecological balance of nature. It concentrates millions in drab cities that, in the United States at least, are being evacuated by those who can afford to escape to the suburbs. The result is a daily movement of traffic, choking the highways, poisoning the air, jangling the nerves. The inventory could be extended indefinitely. Much of the "wealth" produced by industry hardly justifies such consequences.

Some criticism goes deeper. It is urged that, quite apart from its disastrous impact on the environment, modern technology deracinates man, depriving him of his sense of identity, chaining him to a treadmill, trapping him in a depersonalized, regimented, over-organized society. The voice of the romanticist, denouncing preoccupation with production as a bourgeois aberration or an archaic Protestant ethic and urging a return to the relaxed, simple life, is once again heard. That voice is often found persuasive, it so happens, by a new generation of youthful intellectuals and vagabonds.

Most liberals contend that environmental disruption is not a necessary consequence of expanding industry but of a failure to subordinate the quest for quick profits to the requirements of intelligent planning. The "fallacy of romantic regression," as the American psychologist Kenneth Keniston has called it, is also rejected. Nevertheless, there is increased respect for Freud's reminder in *Das Unbehagen in der Kultur* (Eng. trans., *Civilization and Its Discontents,* 1930) that "power over nature is not the *only* precondition of human happiness," and more and more concern is expressed for the "quality" of life. However, Freud went on to observe that, if happiness requires more than power over nature, this "is no ground for inferring that technical progress is worthless from the standpoint of happiness." Mainstream liberals have agreed.

THE FUTURE OF LIBERALISM

The traditional coalition on which liberals in the United States have relied is in disrepair. Ethnic minorities have become confused and divided by the separatist creed of their more militant leaders; turmoil has left the academic community traumatized; and labour unions lack a good deal of their old dynamism. Even the once engagé artist has been lost to aestheticism and esotericism. The increasing polarization between right and left finds the liberals attacked on both flanks, but the appearance of beleaguerment may well be deceptive. Much of the liberal program has in fact become public policy. When, in the 1930s, Pres. Franklin D. Roosevelt established the National Resources Planning Board, it was called un-American by conservatives, its budget was slashed, and it was finally abolished. Today, although most conservatives might resist liberal proposals to establish a council on national *goals,* they would no doubt support a national agency to plan the use of natural resources. Despite appearances, Western society could be on the eve of less, rather than more, polarization.

All of this suggests a new source of strength for liberalism epitomized in what has been called the "end of ideology." Whether the movement of Western capitalist nations toward mixed economies and the Soviet Union's flirtation with the profit motive is indicative of such a trend need not be explored here. But, clearly, both socialists and conservatives in western Europe (and the British Commonwealth) are much less prone to apply pre-established dogma to new problems. And, except for the far right and left, there is a growing tendency in the United States to reject doctrinaire approaches to social problems. Ideology may not be dead, but it appears to be in decline. If the result is a pragmatic, experimental temper in which old principles are adjusted to new needs and dissent is not identified with disloyalty, this is very much in the liberal mood.

Changing policies, enduring values

The expansion of government power and responsibility sought by liberals today is clearly opposed to the contraction of government power and responsibility sought by liberals yesterday. The content of liberalism varies with varying conditions: liberals may one day challenge and another day cherish the church; in one age they may seek less government intervention in economic affairs, in another age more; they have been hospitable to the interests and ambitions of the business community, under changed circumstances they may be hostile; for decades they have preached the virtues of labour unions, they may one day consider their vices. But in every case the inspiration is the same: a hostility to concentrations of power that threaten the freedom of the individual and prevent him from realizing his potentialities; a willingness to reexamine and reconstruct social institutions in the light of new needs. This willingness, tempered by aversion to sudden, cataclysmic change, is what sets the liberal off from the radical, who often ignores its hazards. Also the very eagerness constantly to entertain and encourage useful change distinguishes the liberal from the conservative. If the content of liberalism varies, the above listed characteristics constitute its distinctive and enduring form.

The 18th-century French philosopher Condorcet could write that "human perfectibility is in reality indefinite" and that "the progress of this perfectibility, henceforth independent of any power that might wish to stop it, has no other limit than the duration of the globe upon which nature has placed us." Another French philosopher of the Enlightenment, Helvétius, wrote that "to be happy and to be powerful is only a matter of perfecting the science of education." In his *Autobiography,* John Stuart Mill, reminiscing about the elder Mill, writes:

> So complete was my father's reliance on the influence of reason over the minds of mankind, whenever it is allowed to reach them, that he felt as if all would be gained if the whole population were taught to read, if all sorts of opinions were allowed to be addressed to them by word and in writing, and if by means of the suffrage they could nominate a legislature to give effect to the opinions they adopted. He thought that when the legislature no longer represented a class interest, it would aim at the general interest, honestly and with adequate wisdom . . .

Commenting on the faith of the mid-19th century, Henry Adams noted: "Education was divine, and man needed only a correct knowledge of facts to reach perfection . . ."

Sobered by the tragic events of this century and informed by new psychological insights into man's biasses and perversities, a chastened and far more sophisticated liberalism no longer shares the naive confidence of most classical liberals in man's rationality and perfectibility and in the inevitability of progress. Most of today's liberals are more likely to heed those who warn that human nature is ineradicably flawed than heed those who hope that man may be persuaded to apply the scientific method to the solution of social problems and thereby find contentment. There is a strong suspicion that if man had no serious social problems he would invent them. Nevertheless, the strong commitment of liberals to social reform suggests a persistent optimism and a belief that man can control his fate and build a better world. To this extent the Heavenly City of the Philosophers still beckons them. (H.K.G.)

Conservatism

Continuity and stability

The term conservatism, although it has had different implications in varying historical and geographical contexts, is best reserved to denote a preference for institutions and practices that have evolved historically and that are thus manifestations of continuity and stability. Political thought, from its beginnings, contains many strains that can be retrospectively labelled conservative, but it was not until the late 18th century that conservatism began to develop as a political attitude and movement reacting against the French Revolution of 1789. The noun seems to have been first used after 1815 by French Bourbon restorationists such as François-René, vicomte de Chateaubriand. It was used to describe the British Tory Party in 1830 by John Wilson Croker, the editor of *The Quarterly Review;* and John Calhoun, a formulator of conservative minority rights against majority dictatorship in the United States, also used the term in the 1830s. The generally acknowledged originator of modern, articulated conservatism (although he never employed the term) was the British parliamentarian and political writer Edmund Burke in his essay *Reflections on the Revolution in France* (1790). Pro-parliamentarian opponents of the French Revolution, such as Burke, believed that the violent, untraditional, and uprooting methods of the Revolution outweighed and corrupted its liberating ideals. More authoritarian opponents, such as the polemicist and diplomat Joseph de Maistre, also rejected the ideals themselves. The general revulsion against the course of events in France provided conservatives with an opportunity for restoring the pre-Revolutionary traditions, and a sudden flowering of more than one brand of conservative philosophy followed.

CONSERVATIVE ATTITUDES

Because Burke's case against radicalism and revolution has also influenced liberals, there is often no sharp distinction between liberals and conservatives in action. In philosophy, however, conservatism has maintained certain sharply nonliberal assumptions about human nature.

Whether intentionally or unconsciously, whether literally or metaphorically, for example, conservatives tend to assume in politics the Christian doctrine of man's innate original sin, and herein lies a key distinction between conservatives and liberals. Men are not born naturally free or good (conservatives assume) but are naturally prone to anarchy, evil, and mutual destruction. What the 18th-century French philosopher Jean-Jacques Rousseau denounced as the "chains" that hinder man's "natural goodness," are for Burkeans the props that make man good. These "chains" (society's traditional restrictions on the ego) fit man into a rooted, durable framework, without which ethical behaviour and responsible use of liberty are impossible.

Conservatism as a state of mind

The conservative temperament may be, but need not be, identical with conservative politics or right-wing economics; it may sometimes accompany left-wing politics or economics. Regardless of a conservative's politics or economics, however, it can be said that two characteristics of the conservative temperament are: a distrust of human nature, of rootlessness, of untested innovations; and a corresponding trust in unbroken historical continuity and in traditional frameworks within which human affairs may be conducted. Such a framework may be religious or cultural or may be given no abstract or institutional expression at all. In relation to the latter aspect, many authorities on conservatism—a minority in France and a majority in England—consider conservatism an inarticulate state of mind and not at all an ideology. Liberalism argues; conservatism simply *is.* When conservatism becomes ideologized, logical, and self-conscious, then it resembles the liberal rationalism that it opposes. According to this British approach, logical deductive reasoning is too doctrinaire, too 18th century. Whereas the liberal and rationalist mind consciously articulates abstract blueprints, the conservative mind unconsciously incarnates concrete traditions. And, because conservatism embodies rather than argues, its best insights are almost never developed into sustained theoretical works equal to those of liberalism and radicalism.

Conservatism is often associated with some traditional and established form of religion. After 1789, the appeal of religion redoubled for those craving security in an age of chaos. The Roman Catholic Church, because its roots are in the monarchic Middle Ages, has appealed to more conservatives than any other religion. Himself a Church of England Protestant, Burke praised Catholicism as "the most effectual barrier" against radicalism. But conservatism has had no dearth of Protestant and strongly anticlerical adherents also.

Conservatives typically view society as a single organism and condemn as "rationalist blueprints" the attempts of progressives to plan society in advance from pure reason instead of letting it evolve naturally and unconsciously, flowering from the deep roots of tradition. They dismiss a liberal society as "atomistic," meaning composed of disrupted elements held together merely mechanically. A society, they argue, has to be rendered whole by religion, idealism, shared historical experiences, commitment to its long-standing political institutions, and by the emotions of reverence, cooperation, and loyalty; a society, they believe, can, to the contrary, be rendered atomistic by materialism, class war, excessive laissez-faire economics, greedy profiteering, overanalytical intellectuality, subversion of shared institutions, insistence on rights above duties, and by the emotions of skepticism and cynicism. Except for the German Romantic school, conservatives do not carry their conceptions of the organic wholeness of society to the extreme at which the individual becomes nothing, society everything, for they recognize that, at that extreme, one no longer has conservatism but totalitarian statism.

VARIETIES OF CONSERVATISM

The Burkean foundations. Burke did more than any other thinker to turn the intellectual tide from a rationalist contempt for the past to a traditionalist reverence for it. An Irishman, he loved England, including its established Anglican Church and its nobility, with an outsider's passion. In 1765 he became private secretary to Charles Watson-Wentworth, 2nd marquess of Rockingham, the head of the less liberal wing of the Whig Party. Against the untraditional tyranny of George III, Burke defended the American Revolution of 1776, which he viewed as being in defense of traditional liberties, but attacked the radical French Revolution of 1789 as tyranny by mobs and deracinated theorizers. At a time (1790) when the French Revolution still seemed a bloodless utopia, he predicted its later phase of terror and dictatorship, not by any lucky blind guess but by an analysis of its devaluation of tradition and inherited values.

Burke's conception of the social contract

Indeed, the core of Burke's thought and of conservatism is fear of rootlessness. Rousseau's *Social Contract* of 1762 had favoured a contract merely among the living, to arrange government for their mutual benefit. Burke, instead, argued:

> Society is indeed a contract... [but] as the ends of such a partnership cannot be obtained in many generations, it becomes a partnership not only between those who are living, but between those who are dead, and those who are to be born.... Changing the state as often as there are floating fancies,... no one generation could link with the other. Men would be little better than the flies of a summer.

Burke's veneration of the past may be contrasted with the rationalist hostility of Karl Marx, the most influential social critic of modern times: "The legacy of the dead generations weighs like a nightmare upon the brains of the living." But for Burke the contract is with "the future" as well as with the past, and he thus urges improvement, as long as it is evolutionary: "A disposition to preserve and an ability to improve, taken together, would be my standard of a statesman."

Burke was defending not conservatism in the abstract but, rather, one concrete instance of it, the unwritten British constitution. His arguments, however, were not always consistent. Sometimes he justified that constitution by "natural rights"; more often by "prescriptive right." Natural rights meant a universal code external to any given constitution; prescriptive right, a local code authoritative (prescriptive) by virtue of its age and its links with the past, which are prima facie evidence of its value. Sometimes he argued that natural rights preceded the constitution and gave it "latent wisdom." But, when arguing against French rationalists, who would justify their own revolutionary constitution by natural rights, he argued instead, and more typically:

> Our constitution is a prescriptive constitution... [whose] sole authority is that it has existed time out of mind... without any reference whatever to any other more general or prior right.

Burke shocked his century by his brutal frankness in defending "illusions" and "prejudices" as socially necessary. In doing so, however, he was, in fact, being not so much a cynic as one of the few old-fashioned Christians among 18th-century intellectuals. He was an old-fashioned Christian in the sense of believing man innately depraved, innately steeped in original sin, and incapable of bettering himself by his feeble reason. So defined, man could be tamed only by following an ethically trained elite and by education in "prejudices," such as family, religion, and aristocracy. He called landed aristocrats "the great oaks" and "proper chieftains," provided they tempered their rule by a spirit of timely reform from above and remained within the constitutional framework. He defended the Church of England for its political as well as its religious function, "To keep moral, civil, and political bonds, together binding human understanding."

Coleridge and Wordsworth. After Burke, the English poets Samuel Taylor Coleridge and William Wordsworth were significant figures in the formulation and expression of conservative sentiment. They began, however, as utopian liberals supporting the French Revolution. Wordsworth spoke for a whole generation of European intellectuals with his famous salute to the new dawn in France: "Bliss was it in that dawn to be alive, but to be young was very heaven." Disillusionment followed, and Coleridge and Wordsworth reacted against liberalism and rationalism and turned to traditional monarchy and the Church of England.

In 1798 Wordsworth and Coleridge jointly published their book of poems, *Lyrical Ballads,* marking the revolt of the human heart against abstract 18th-century rationalists and thereby helping to create a new philosophical climate. Conservatism was permanently influenced by Coleridge's prose works: *Lay Sermons,* 1816–17; *Biographia Literaria,* 1817; *Philosophical Lectures,* 1818–19; *Aids to Reflection in the Formation of a Manly Character, on the Several Grounds of Prudence, Morality, and Religion,* 1825; and his various *Letters* and *Specimens of Table Talk.* His public lectures exercised an indirect influence by molding the minds of university students who later became national leaders.

Coleridge's views on social classes

According to Coleridge, society divided its functions among different "class orders." Each class had its valuable function, but this did not necessarily include the right to vote and rule. That right was best left to an ethically trained aristocracy, functioning within the strict lawful limits of Parliament. All classes, Coleridge argued, must cooperate harmoniously within the organic unity of the constitution. His greatest influence on practical politics was through his disciple Benjamin Disraeli, later to be Conservative prime minister, and his disciple's disciple, Sir Winston Churchill. Coleridge considered businessmen often subversive, not conservative; they allegedly gnawed at the foundations of Christian monarchy by substituting a newfangled, un-Christian religion known as economic profit. Thus Coleridge, defining "shopkeepers" as "the least patriotic and the least conservative" class, fought against the Whig Reform Bill of 1832, which made "hucksters" the dominant voting group.

Maistre and Latin conservatism. It would convey an unbalanced picture of conservatism to present only the moderate and British brand founded by Burke and to omit the more extreme and Latin brand founded by Maistre (died 1821). Whereas Burkean conservatism is evolutionary, the conservatism of Maistre is counterrevolutionary. Both favour tradition against the innovations of 1789, but their traditions differ: the former fights against 1789 for the sake of traditional liberties, the latter for the sake of traditional authority. The former is not authoritarian but constitutionalist—and often parliamentary—whereas the latter, in its stress on the authority of some traditional elite, is often justifiably called not conservative but reactionary. To call it totalitarian, however, would be to go much too far, for its authority does not try to be "total," in the sense of taking over the total personality, the total culture, but is restricted to politics—and sometimes also religion. The distinction between the authoritarian and the totalitarian separates even the most reactionary conservative from the totalitarian Nazis and Communists.

After the breakdown of the French Revolution, Maistre became the most influential philosophical spokesman for the *ancien régime.* Against the slogan "liberty, equality, fraternity," he seemed almost personally to embody the slogan "throne and altar." His program consisted of a restoration of hereditary monarchy, but a more religious and less frivolous monarchy than before. He was an international refugee after the French, during the Revolution, invaded his native Savoy—then a French-speaking province of the Italian-speaking monarchy of Piedmont-Sardinia. He became for 14 years Sardinian ambassador to Russia, where his restorationist faith was strengthened by the example of the absolute monarchy still functioning there.

Maistre on the role of monarchy

Both restorationist and evolutionary conservatives defended monarchy as a social cement needed to hold society together, to keep it "organic," not "atomistic." But, while the Maistre school (key source of conservative thought in Spain and Italy as well as France) defends monarchy as absolute, the evolutionary British school defends it merely as being "pragmatic"; that is, useful. Maistre and many continental monarchists carried their belief in the monarchy to the extreme of demanding "love" even for an "unjust" ruler, earthly or heavenly:

> We find ourselves in a realm whose sovereign has proclaimed his laws. . . . Some . . . appear hard and even unjust . . . What should be done? Leave the realm, perhaps? Impossible: the realm is everywhere. . . . Since we start with the supposition that the master exists and that we must serve him absolutely, is it not better to serve him, whatever his nature, with love than without it?

This chain of authoritarian reasoning reached its climax in a logical if inhuman paradox: "The more terrible God appears to us . . . the more our prayers must become ardent. . . ." Cruel as these arguments sound, the motive of the personally mild Maistre was humane: revolts against cruel authority would inflict even crueler sufferings on mankind. He drew from the French Revolution the lesson that submission to traditional authority, though admittedly a bitter pill, was Europe's cure for a still more bitter chaos.

Maistre's politics were a theological drama in which "order" (his key concept) was angelic, "chaos" diabolic, and "revolution" original sin. Seduced by the glittering *Social Contract* of Rousseau, giddy and inexperienced nations might lust after democracy or a plebeian Bonapartist dictatorship. But they would come to a perfectly dreadful end, which would serve them right for provoking the wages of sin: "Because she [Europe] is guilty, she suffers" (1810). From suffering, Maistre argued, Europe would learn that the purest order is a fatherly Christian monarchy. Even kings must avoid rocking the boat of order with liberal "innovations": Europe must "suspect" the word "reform." In *Du Pape* (1817; "Concerning the Pope"), he analyzed "order" further: its hierarchical pyramid logically required one supreme apex. That apex must be no earthly monarch, of which there were so many, but the union of earthly and spiritual power in the papacy.

The vast extent of the instability following the French Revolution surprised even its supporters, and the problem of how to restabilize society emerged as one of some practical importance. According to Maistre's *Soirées de Saint-Pétersbourg* (left unfinished 1821; "Evening Conversations in St. Petersburg"), the solution was more faith and more police. That combination he summed up in his own frank formula: "the pope and the executioner." The pope was the positive bulwark of order: he gave faith. The executioner was the negative bulwark: he suppressed disorder. Himself an intellectual, Maistre indicted intellectuals as "rebellious" and "insolent" fomenters of disorder.

Maistre, this very secular exalter of clericalism, resembled not the Church Fathers but the very rationalists he attacked. He arrived at his glorification of unreason and of divine authority not by mystic intuition—not even by unthinking acceptance of traditional authority—but by using his own mind independently, rationally, and with steps of deductive logic. Though Maistre would never have admitted it, he might be characterized as the last abstract rationalist of the whole Voltairean Age of Reason. Even more than the rationalist Voltaire and as much as the rationalist Jacobins, Maistre believed in pure and absolute ideas, although his idea was absolute authority rather than absolute reason. In Maistre the destructive deductive logic of the 18th century was carried so far that it destroyed even itself—pure reason committing suicide for the sake of pure order.

This division into Burke and Maistre wings does not mean both were equal in importance or influence. No work of Maistre or any other anti-Jacobin has approached the influence of Burke's classic essay. Burke, above all, was the first to formulate the rebuttal to the French Revolution; his arguments were borrowed, sometimes word for word, by all later conservatives, including the restorationists. Maistre's rigid hierarchical conservatism is in the latter part of the 20th century dying out, whereas Burke's more flexible brand is stronger than ever, permeating all parties of the West, emphatically including democratic Socialists with their increasing stress, in Great Britain and Germany, on what a Fabian Socialist has called, in good Burkean language, "the inevitability of gradualness."

Later French conservatism

French conservatism after Maistre presents a diversified range of views, from the thought of Charles Maurras, the far-right editor of *L'Action Française* who seemed more fascist than conservative and became a Nazi collaborator, to the anti-authoritarian Alexis de Tocqueville, author of *Democracy in America* (1835–40) and the most Burkean French critic of the Revolution and of plebiscitarian mass democracy. To some extent, however, Tocqueville, an evolutionary parliamentarian, can also be regarded as a liberal thinker. In between Maurras and Tocqueville come the great anti-Jacobin Hippolyte-Adolphe Taine; the philosophical novelist Maurice Barrès, more a nationalist than anything else but conservative in his stress on organic roots; and Louis-François Veuillot, the editor after 1843 of the newspaper *L'Univers Réligieux* and a clerical restorationist who ably readapted Maistre to the industrial modern world. An influential right-wing extremist, less clerical and more statist than Maistre and Veuillot, was Louis-Jacques-Maurice de Bonald, the apologist for Napoleon's empire and then for the Bourbon Restoration.

Metternich and the Concert of Europe. The problems posed by the widespread social unrest of the Revolutionary and Napoleonic periods and their aftermath, and the insecurity of governments in the face of demands for constitutions and liberal reforms, provoked a reaction of more immediate and far-reaching consequence than the writings of conservative theorists. During the period 1815–48, Prince Metternich, a major influence in Austria and in Europe generally, devoted his energies to erecting an anti-revolutionary chain of international alliances throughout Europe in order to protect the multinational empire that he administered.

Metternich viewed the liberal revolutions of the 1820s and 1830s in Italy, Spain, and Germany as being unhistorical and unrealistic. Liberals were trying to transplant from England free institutions, which had no historic roots on the Continent. He retorted with Burkean arguments about the need for old roots and orderly organic development. Hence, his sarcastic comments on the liberal revolutions in Naples and elsewhere:

> A people who can neither read nor write, whose last word is the dagger—fine material for constitutional principles! . . . The English constitution is the work of centuries. . . . There is no universal recipe for constitutions.

Though his repressive Carlsbad Decrees of 1819 infringed inexcusably on basic liberties, his attitude was not always so negative. Just before his fall in 1848, he was at last winning acceptance from the archdukes of his sincere, thoughtful, and practical plan (postponed too long by the reactionary emperor Francis I) to convoke delegates from all the provincial estates to a representative body in Vienna.

Metternich was a dominating figure at the Congress of Vienna, the international peace conference of 1815 after the Napoleonic Wars. The Vienna peace was based on certain conservative principles shared by the Austrian delegate

Conservative objectives following the Napoleonic Wars

Metternich, the British delegate Robert Castlereagh, the French delegate Talleyrand, and the formerly liberal Russian tsar Alexander I. These principles were conservatism, in reaction against Revolutionary France; traditionalism, in reaction against 25 years of rapid change; legitimism (the principle of hereditary monarchy as the only lawful rule); and restoration (the principle of restoring the kings ousted after 1789).

The European great powers also aimed at the enforcement of peace by subsequent conferences between kings, and those subsequent conferences gave rise to a period of international cooperation known as the Concert of Europe. As liberal democrats correctly pointed out, the weakness of that first successful attempt at a "United Nations" was its narrowly aristocratic base. But it did achieve the positive function—and important precedent—of peacefully arbitrating several disputes. The debit of the conservative Concert of Europe was its bigoted suppression of democratic social progress.

Goethe's spiritual conservatism. Johann Wolfgang von Goethe was Germany's greatest dramatist, poet, and personality. In his youthful "storm and stress" period of the 1770s, Goethe went through a phase of revolt and of nationalism. In his old age, however, he became Germany's greatest cultural influence for classical balance and for antinationalist cosmopolitanism, influencing many outside Germany, including, in England, Coleridge. In 1815 Goethe and Metternich both took pride in being "good Europeans," not German nationalists. After a friendly personal conversation with Metternich, Goethe wrote that Metternich "inspires with the assurance that reason, reconciliation, and human understanding will lead us out of present chaos." Later, in 1830, Goethe urged a mature synthesis between a conservative framework and liberal goals:

> The genuine liberal tries to achieve as much good as he can with the available means to which he is limited; but he would not use fire and sword to annihilate the often inevitable wrongs. Making progress at a judicious pace, he strives to remove society's deficiencies gradually without at the same time destroying an equal amount of good by violent measures. In this ever-imperfect world he contents himself with what is good until time and circumstances favor his attaining something better.

His rhymed credo "Nature and Art" (1802) expressed his conservative and classic stress on voluntary submission to law: "Only in self-restriction does the master reveal himself. And only law can give us liberty." His political drama *Die natürliche Tochter* (1803; *The Natural Daughter*) reflected his hostility to the French Revolution, radicalism, and mass movements. Much quoted by classicists, such as the United States' Irving Babbitt, was Goethe's definition: "The classical I call the healthy and the romantic the diseased." Yet his *Faust* drama (*Part I* published 1808, *Part II* 1832) retained the liberal-minded stress of his younger days on constant change, "constant striving," as salvation. His most unique achievement consisted of his being, so to speak, self-invented. By sheer strength of character, he remolded his naturally revolutionary and romantic temperament into what the world accepted as a conservative and classicist temperament.

Savigny and Ranke

Perhaps Germany's most mature conservative thought came from her great historians. Friedrich Karl von Savigny (died 1861) and Leopold von Ranke (died 1886) were outstanding as pupils of Burke in their reverence for history as organic growth. Savigny stressed that custom, operating over centuries, creates its own framework. On custom, Savigny founded an entire science of historical jurisprudence, denying the abstract, liberal "rights of man." Similarly, Ranke saw every society in terms of its own unique evolution. He opposed the universal generalizations of the 18th-century Enlightenment; every people, he wrote, "is related directly to God" in its own concrete way.

Tsarist and Dostoyevskyan conservatism. Whereas Western conservatism arose from reactions to the French Revolution, Russian tsarist conservatism had different and older origins. The practice of the absolute Tatar khans and the theory of Byzantine caesarism combined to produce an un-Western elephantiasis of autocracy. Nevertheless, two antiliberal traditionalists of Russia made such an impact on the West—the first by politics, the second by art—that their mention is indispensable: Konstantin Pobedonostsev and Fyodor Dostoyevsky. The former was the tutor and chief ideologist of two tsars (Alexander III and, until the Revolution of 1905, Nicholas II). His book *Reflections of a Russian Statesman* (1898) denounced free press, trial by jury, parliamentary government, secular education, skepticism toward the divine mission of tsars, and, above all, intellectuals.

Dostoyevsky's disillusionment with his youthful radicalism resembled Coleridge's in its psychological as well as literary consequences. Both turned to an organic, religious, and monarchic society, to which they paid more homage via literature than via politics. Dostoyevsky attacked Socialism, liberalism, materialism, and atheism. He preached Greek Orthodox tsarism, Slavic traditionalism, and the redemption of mankind by "Holy Russia." His novel *The Possessed* (1871–72) pictured the idealistic ends of Socialists as corrupted by their terroristic means, and he boasted somewhat fawningly to Alexander III about the book's effectiveness against radicals. His novel *The Brothers Karamazov* (1880) contrasted a dry Western rationalism with a more deeply moving Russian mysticism. To the end he retained from his young Socialist days his characteristic compassion for what he called "the insulted and injured"; only now he expressed this in the more spiritual creed of Christian love. What influences many modern readers so compellingly is not his political but his cultural conservatism, exalting vision beyond external material progress.

American conservatism. The American Revolution owed many of its ideals to Burke's interpretation of the British heritage of 1688, the heritage of mature self-government. Burke favoured the Revolution as defending the traditional rights of freeborn Englishmen against newfangled royal usurpations. In that sense, one might describe it not as the Revolution but as the "Conservation" of 1776.

In *The Rights of the British Colonies Asserted and Proved* (1764) the American spokesman James Otis typically argued that the demand for no taxation without representation was an old British tradition. America, he said, was conserving "the British Constitution, the most free one on earth." "We claim nothing," added George Mason of Virginia, "but the liberty and privileges of Englishmen." Almost all other revolutions, colonial or otherwise, have been radical in the sense of demanding new or increased liberties and a new order. In contrast, the American demand of July 6, 1775 (*Declaration of the Causes & Necessity of Taking Up Arms*), was for conserving old liberties and the old order: "in defence of the freedom that is our birth right and which we ever enjoyed until the late violation of it." Such words promulgated no democracy, no abstract "Rights of Man"; rather, they promulgated what Burke called "prescriptive right. . . . considering our liberties in the light of an inheritance." Despite important exceptions, which should not be minimized, it was not until the election of the more truly "revolutionary" Andrew Jackson (1828) that the democratic doctrines of the pamphleteer Thomas Paine gained solid roots in the United States, dividing the nation between conservative and progressive traditions. Paine was the man whom the Burkean John Adams (president 1797–1801) came to loathe most—for eternally sloganizing about apriorist utopias. A leading historian, Daniel Boorstin, has observed in *The Genius of American Politics* (1953):

> The ablest defender of the Revolution—in fact, the greatest political theorist of the American Revolution—was also the great theorist of British conservatism, Edmund Burke. . . . Ours was one of the few conservative colonial rebellions of modern times.

Conservative doctrines of Hamilton and Madison

The spirit of the United States was partly molded by two masterpieces of Burkean conservatism, both published in 1787–88: *The Federalist,* by Alexander Hamilton, James Madison, and John Jay, and *Defence of the Constitutions of Government of the United States of America,* by John Adams. The achievements attributed by historians to the *Federalist* papers exceed those of any other series of newspaper articles in history, for they helped forge national unity during a separatist crisis. In the context of Shays's Rebellion of 1786 against the judiciary, they saved government by law from government by mob and established minority rights against majority dictatorship. They based American liberty on the Burkean principle of historical roots, prescriptive right, and judicial precedent instead of on vague grand rhetoric about democratic utopias and the masses. Similar in thought and richer in historical background was the *Defence* by Adams, one of the most penetrating analyses of self-government ever written.

The U.S. Constitution was drawn up in Philadelphia by the U.S. Constitutional Convention of 1787. The objectives of many liberal democrats were: easy amendment; facilities for mass pressure and rapid change; unchecked popular sovereignty; universal manhood suffrage; a single parliamentary body; and the basing of liberty on a long list of universal a priori abstractions, such as Burke later criticized in the French Declaration of the Rights of Man and of the Citizen. But in the Constitution of 1787 the Federalists foiled each of these objectives. They made amendments slow and difficult, greatly reduced the number of voters by property restrictions, created a congress of two parliamentary bodies, and based liberty primarily, though not entirely, on the concrete, inherited precedents of British tradition. Except for the House of Representatives (a sop to democrats), the main cogs of government—president, Senate, justices—were not to be chosen directly by the people but, respectively, by the electoral college, state legislatures, and appointment, and not until 1913 did an amendment eliminate this intentionally undemocratic election of senators. The judicial branch (Supreme Court) continues to be a nonelective, nonremovable elite not responsible to democratic majorities. Yet it can veto as unconstitutional measures passed by a democratic majority of the two elective, removable branches of Congress.

The American Founding Fathers adopted a conservative constitution in reaction against current mob excesses and against the democratic–utopian rhetoric of the earlier Declaration of Independence (drawn up by Thomas Jefferson) with its grand abstractions about "life, liberty, and the pursuit of happiness." Yet the Constitution was the Burkean, not the reactionary brand of conservatism. Thus it defeated not only the liberal objectives but also the more extreme conservative ones, including a hereditary, titled aristocracy and Hamilton's notion of a president for life with absolute veto power.

John Calhoun

The United States' only consistently conservative party was the Federalist Party of John Adams and Alexander Hamilton. Hamilton was perhaps too much the reckless commercial adventurer to be classified under conservative or any other principles, but Adams remains the closest New World equivalent to Burke. After the death of the Federalist Party in the early 1800s, two mutually hostile kinds of political conservatism emerged: that of the urban New England Brahmins and that of the Southern semifeudal landowners. The latter received their most persuasive defense in the famous *A Disquisition on Government and Discourse on the Constitution and Government of the United States* of Calhoun, the closest New World equivalent to Maistre. This more extreme, very regional Calhoun conservatism is still influential in much of the American south, typically cutting across Democrat or Republican party lines, and is still alien to New England conservatism.

Modern U.S. political parties, being pragmatic alliances of geographic patronage groups rather than matters of doctrine, cannot realistically be classified under "isms." It is nearer to reality to look for conservatism, instead, in the indirect diffusion—cutting across all party lines—of the above described restraining principles of the Constitution.

CONSERVATISM IN THE 20TH CENTURY

The 19th and, particularly, the 20th century (that is, the period since the 18th-century Enlightenment) have in many ways been antithetical to conservatism, both as a political philosophy and as a program of particular parties identified with conservative interests. As described above, the consciously articulated conservatism of Burke was formulated in reaction to the French Revolution; similarly, the anti-liberal, anti-revolutionary policy that was a major factor in European international relations during the Metternich period (1809–48) was a reaction to the political discontent aroused by demands for liberal reforms and constitutions. The Enlightenment, in fact, had resulted in the propagation of certain attitudes and ideas that were to have far-reaching political consequences during the succeeding centuries, the most significant of which were a belief in the possibility of improvement in the human condition—a belief, that is, in the idea of progress—and a concomitant disposition to tamper with or discard existing institutions or practices in pursuit of progress, a disposition that has been characterized as "rationalist." Such rationalist politics embrace a broad segment of the political spectrum, including much of liberal reformism, socialism of the welfare-state or mixed-economy variety characteristic of western Europe, and Marxist socialism. The changes that have been wrought under the banner of rationalist politics have thus been immense and point to what has been described as a dilemma of modern conservatism—the extent to which, in face of constant rationalist innovation, conservatives may be forced to adopt a merely defensive role, so that the political initiative lies always in the other camp.

The responses of conservatives to this predicament have naturally varied considerably in differing political contexts; an account of some of these responses is given below. An analysis of the role of conservatism in contemporary politics, however, cannot be confined merely to an account of the programs of political parties identified with the conservative cause, for conservatism makes its influence felt in a variety of ways less direct than through expression in party platforms. Conservatism in the 20th century has in fact been a pervasive force in the political life of those parliamentary democracies in which rationalist politics have seemed to hold sway, as well, of course, as in less liberal political climates.

Non-political manifestations of conservatism

Conservative influences operate indirectly (*i.e.,* other than via the programs of political parties) largely by virtue of the fact that, while man is undeniably a persistent innovator, there is also much in the human temperament that is naturally or instinctively conservative: among such conservative traits are the tendency to fear and avoid sudden change and the tendency to act according to habit. While these are traits of the individual, they may find collective expression in, for example, resistance to imposed political change and in a whole cluster of value preferences that contribute to the formation and stability of a particular culture. The tendency for value preferences to find expression in cultural forms and political institutions (the so-called pragmatism of the British, for example, in their unwritten constitution) constitutes a profound conservative influence in political life over and above any explicit articulation of particular conservative interests that may be undertaken by a political party, for it gives rise to practices and institutions that are products of a long process of social and political evolution and are closely related to other culture-related factors, such as religion and property relationships. The existence of such cultural restraints on political innovation constitutes in all societies a fundamental conservative bias, the implications of which have been aphoristically expressed by an English commentator, F.J.C. Hearnshaw: "It is commonly sufficient for practical purposes if conservatives, without saying anything, just sit and think, or even if they merely sit." Mere inertia, however, has rarely sufficed to protect conservative values in an age dominated by rationalist dogma and by social change related to continuous technological developments. The conservative reaction, however, is best analyzed in specific political contexts. Historians, it may be noted, cannot safely agree on there being more than four great political parties of the 20th century deserving of the name: the Conservative Party of England, the Christian Democrats of Italy and of Germany, and the Liberal Democrats of Japan.

In England, Disraeli's successor, Lord Salisbury, was prime minister in 1885, from 1886 to 1892, and from 1895 to 1902; Arthur Balfour succeeding him from 1902 to 1905. This longest era of Conservative rule was characterized by imperialism, high tariffs, and the gradual erosion of the party's working-class vote, which Disraeli had so far-sightedly nurtured by extending the franchise to the workers in 1867. The party had thereby broadened its original class basis (landed aristocracy and established church) to outflank from below and above the new commercial class and its Liberal Party. It may be said that conservatism in Great Britain since Disraeli's time has veered between a passive and largely resigned acceptance of changes introduced by its Liberal and, later, Labour opponents and a more positive conservatism, the aim of which has been to foster a social environment in which the individual is encouraged to advance his own interests without undue hindrance from, or reliance on, the state—a policy descended from the liberal individualism of the 19th century, associated particularly with the Liberal Party. This positive conservatism of liberal individualism tinged with a strong sense of social conscience was given its earliest formulation by Disraeli, who combined a desire to mitigate harsh conditions suffered by the working class under conditions of unrestrained capitalism with a belief in the value of existing institutions such as the monarchy, the church, and the class system. Disraeli's foreign policy, which emphasized the need for Britain to act constructively as a "moderating and mediatorial" power and to maintain its interest in its empire, also reflected the view that conservatism must be a force shaping events rather than merely reacting to them. These three elements—the improvement of material conditions both by encouragement of individual initiative and timely reform of abuses, emphasis on the value of traditional institutions, and belief in the need for an active foreign policy—have been recurring themes of British conservatism in the 20th century. Later conservative thinkers have elaborated on the value of divergency of personality and attitudes, the role of property as an expression of individuality, and the central role of the family in providing a stable environment in which the individual may develop.

British conservatism after Disraeli

In its less positive periods (as, for example, during the interwar period), conservatism in Britain has been identified with the defense of class privileges and of the status quo, an unconstructive opposition to socialism, and, during the 1930s, a deal-making commercialist approach to the rising Nazi menace. Faced, however, with the introduction of a mixed economy and the vast extension of state welfare services by the Labour Party after 1945, the Conservatives, when returned to power in 1951, reversed very few of their socialist predecessors' innovations, emphasizing instead their claim to be more able to administer the welfare state efficiently and to some extent outbidding their opponents, especially in areas of social policy related to their fundamental beliefs—the encouragement of a heavy program of house building being an example. The Conservative resurgence that resulted in the election of Margaret Thatcher as prime minister in 1979 inspired a more activist, if not doctrinaire, spirit, particularly in the fields of economic and fiscal policy (including, for example, the "privatization" of a number of industries nationalized under Labour governments).

Western Europe. It is of significance that the British Conservative Party has been the more ardent of the two major British parties in championing British membership of the European Economic Community (EEC), reflecting an internationalism voiced by Sir Winston Churchill when, in 1940, he appealed for a Franco-British union and, in 1946, for a European union. Originally conceived as a means by which the economies of the European countries might be integrated—so that war between them would be impossible—the nascent community assumed significance during and after the Cold War as a means of strengthening western Europe against the threat of external Communist aggression and internal subversion. Together with the military North Atlantic Treaty Organization alliance, it thus assumed a role as a bulwark of parliamentary democracy and capitalism.

In the arena of party politics, conservatism in western Europe is generally represented by two or more parties, ranging from the liberal centre to the moderate and extreme right. Three types of party may be discerned: agrarian parties (particularly in Scandinavia), Christian democratic parties, and conservative parties linked strongly with big business interests and sometimes with a markedly nationalistic outlook. Such categories are very general and are not mutually exclusive.

The Christian democratic tradition

Among parties of the right, the Christian democratic tradition has the longest continuity, the predecessors of contemporary parties having emerged during the first half of the 19th century to represent supporters of the church and the monarchy against liberal elements. Especially after World War I, business interests became a third important element. The clerical interest is strongest in the Democrazia Cristiana (DC; Christian Democrat Party) of Italy, which has dominated government since 1945. Through this party, Catholicism has set limits on policy concerning such church-related matters as divorce and contraception; in regard to other social questions, however, the party has never presented a coherent policy, largely because it comprises little more than an alliance of disparate and often conflicting interest groups.

In West Germany, a country divided between Catholics and Protestants, the church plays a far less significant role in the main conservative party, the Christlich-Demokratische Union (CDU; the Christian Democratic Union). After 1950, following debate within the party over economic and social questions, advocacy of a free-enterprise economy coupled with a strong commitment to maintain and improve social insurance and other welfare provisions became established policy. The conservative temper of the political climate in Germany since the beginning of economic recovery may be judged from the fact that since the early 1950s the main opposition party, the Sozialdemokratische Partei Deutschlands (SPD; the Social Democrats), has progressively eliminated the socialist content of its program, a congress at Bad Godesberg (1959) in fact going so far as to champion the profit motive.

France provides an exception to the general pattern of the representation of moderate conservative opinion by a Christian democratic party; the closest equivalent has been the Catholic, right-wing, Mouvement Républicain Populaire, which by the late 1960s had become little more than a political club. Instead, a large proportion of conservatives in France has supported Gaullist groups such as the Union pour la Défense de la République. Gaullist conservatism has been markedly nationalistic, involving assumptions concerning French leadership of a united Europe and emphasizing tradition, order, and the regeneration of France. Gaullists espouse divergent views on domestic social issues, however, as do non-Gaullist groups such as the Centre National des Indepéndants et Paysans. The number of conservative groups, their lack of stability, and their tendency to be identified with local issues defy simple categorization. Conservatism in France, however, as in Italy and Germany, has been the dominant political force since World War II.

Conservatism in Europe is thus revealed as a dominating political influence in the major states, finding expression in parties of very different character. These parties represent traditional bourgeois values and oppose unnecessary state involvement in economic affairs and any radical attempts at income redistribution. They are also characterized by an absence of ideology and often of even a well-articulated political philosophy, but this tends to be of little consequence in terms of their influence since they give political expression to the conservatism of temperament mentioned above as an important underlying bias in political conflict, as well as to persistent culture-related values that are of great importance in terms of continuity and stability.

Japan. The relationship between conservatism as an underlying bias related to psychological factors and cultural values and conservatism as an articulated political credo is illustrated by the history of party politics in Japan since its opening to Western influence in the middle of the 19th century. The political and social changes that took place following the Meiji Restoration (1868) were of major proportions, involving the abolition of feudal institutions and the introduction of such Western political ideas as constitutional government. But despite institutional innovations and the dislocations resulting from rapid industrialization, traditional loyalties and attitudes proved to be more important factors in shaping political developments.

Majority conservatism in Japan

Except for the period of intervention by the militarists during the 1930s and 1940s, Japan has been ruled by conservatives since the beginning of party politics in the 1880s. The conservative parties (the two most important of which merged to form the Liberal-Democratic Party in 1955) have been dominated by personalities rather than by ideology and dogma; and personal loyalties to leaders of groups within the party (factions) rather than commitment to policy have determined the allegiance of conservative members of the Diet. As one American scholar, Nathaniel B. Thayer, has described it, the factions

> have adopted the social values, customs, and relationships of an older Japan.... The old concepts of loyalty, hierarchy, and duty hold sway in them. And the Dietman (or any other Japanese) feels very comfortable when he steps into this world.

The Liberal-Democratic Party is intimately linked with big business interests, and its policies are guided primarily by the objective of fostering a stable environment for the development of Japan's free-enterprise economy; to this end, the party functions as a broker of conflicting business interests. Policy toward other Asian countries, national defense, and internal security are other conservative preoccupations.

The United States. It may be argued that the United States has no nationwide conservative or liberal parties but instead only two fluctuating, all-inclusive coalitions. Both the Democrat and Republican coalitions include interest groups occasionally labelled conservative—the racists among southern Democrats, for example, and such Republican offshoots as the local New York Conservative Party. On a journalistic level the word conservative has been used loosely for a segment of the Republican party associated with Sen. Barry Goldwater and with Ronald Reagan, a former California governor elected president in 1980.

Modern American conservatism has been highly influential in the literary and religious realm in such masterpieces of conservative outlook as Irving Babbitt's *Democracy and Leadership* (1924) or the aristocratic traditionalism of the Nobel Prize-winning novelist William Faulkner. Indeed, such figures as the novelist Herman Melville and the theologian Reinhold Niebuhr, usually independent of each other and eschewing conservative labels, have performed the nation's spiritual arithmetic, calculating the spiritual price of material progress and of a robotizing technology. Unconsciously conservative in this sense, even when under radical slogans, is the impulse among young people in the 1970s and after to conserve ecology and environment against what Melville called "the impieties of progress." These unconscious young conservers sublimate the old class-based elitism into a new value-based elitism, open to all, thereby rescuing quality (the cultural as well as physical ecology) from the parvenu plutocrats of quantity (mass culture and robot technology).

The impact of the horrors allowed at Auschwitz has purged—in effect conservatized—many modern liberals out of their most unconservative axiom: the Rousseauist doctrine of the "natural goodness" of man and the masses. For many, the real battle for the future now seems to be an alliance of such chastened liberals with conservatives in jointly defending their shared constitutional and ethical framework against extremist destroyers from a mirror-image right and left.

Conservative reticence

It is arguable that conservatism, whether its influence operates through political parties or through psychological, cultural, and institutional factors, is a far more persuasive influence in democratic societies than the rate of social and economic change and the welter of rationalist dogma would suggest. That it is often lacking in articulation and that, as critics of conservatism point out, there is a comparative lack of persuasive presentations of the conservative cause compared with the abundant literature of rationalist politics is in part a consequence of its underlying strength and in part a result of a certain coyness among the best conservative thinkers deriving from the fear that a conservatism that needs to present itself in the same terms as the doctrines it opposes is no longer conservatism or is a conservatism in retreat. In the latter part of the 20th century, however, many would say that it may be argued that the predilection of governments to extend their role in social life is so strong as to necessitate a more articulate, even aggressive, conservatism. One particularly important task of conservatives will be to emphasize that the social sciences, particularly anthropology and psychology, so long enlisted in the cause of social engineering and liberal utopianism, also reveal much about the role of tradition, custom, and evolution in the survival of societies. (Pe.V.)

BIBLIOGRAPHY

Socialism: DANIEL BELL, "Socialism," in *International Encyclopedia of the Social Sciences,* 14:506–534 (1968), one of the best short surveys of socialism from its origins to the present day; G.D.H. COLE, *A History of Socialist Thought,* 5 vol. (1953–60), the most complete and detailed general study; H.W. LAIDLER, *History of Socialism,* rev. and enl. ed. (1968), and CARL LANDAUER, *European Socialism,* 2 vol. (1959), two histories written from a social democratic perspective; IRVING HOWE (ed.), *Essential Works of Socialism* (1970), a collection of readings in many varieties of Marxist thought; C.A.R. CROSLAND, *The Future of Socialism* (1957), an assessment of the prospects for socialism by a leading member of the British Labour Party; MICHAEL HARRINGTON, "Why We Need Socialism in America," *Dissent,* 17:240–303 (1970), a defense of the relevance of socialism in the United States by a leading American socialist; JAMES JOLL, *The Second International 1889–1914* (1966), a standard but very readable history of 19th-century social democracy; GEORGE LICHTHEIM, *Marxism* (1961), a sophisticated and searching work dealing with the transformation of Marxism from the time of its founders to that of Lenin and Stalin; *A Short History of Socialism* (1970) and *The Origins of Socialism* (1969), idiosyncratic but stimulating reading; A.F. STURMTHAL, *The Tragedy of European Labor* (1943), a study of the failures of social democratic policies during the depression of the 1930s, by a sympathetic observer; ANTHONY GIDDENS, *A Contemporary Critique of Historical Materialism,* 2 vol. (1982–).

Communism: SHLOMO AVINERI, *The Social and Political Thought of Karl Marx* (1968); ISAIAH BERLIN, *Karl Marx,* 2nd ed. (1948); R.N. CAREW-HUNT, *The Theory and Practice of Communism,* 5th rev. ed. (1957, reprinted 1963); J.L.H. KEEP, *The Rise of Social Democracy in Russia* (1963), an outstanding history of the subject up to 1906; LEONARD SCHAPIRO, *The Origin of the Communist Autocracy: Political Opposition in the Soviet State: First Phase, 1917–1922* (1955); ADAM ULAM, *The Bolsheviks* (1965); FRANCO VENTURI, *Il populismo russo* (1952; Eng. trans., *The Roots of Revolution: A History of the Populist and Socialist Movements in Nineteenth Century Russia,* 1960); BERTRAM D. WOLFE, *Three Who Made a Revolution* (1948), a readable, stimulating history of Bolshevism in its formative years.

Works on Stalinism include ROBERT CONQUEST, *The Great Terror: Stalin's Purge of the Thirties* (1968); IAN GREY, *The First Fifty Years: Soviet Russia 1917–1967* (1967); LEONARD SCHAPIRO, *The Communist Party of the Soviet Union,* 2nd rev. ed. (1970), a detailed history of the Communist Party in theory and in practice up to 1968.

The world movement up to Stalin's death is treated in HAMILTON FISH ARMSTRONG, *Tito and Goliath* (1951), an excellent study of the conflict between Yugoslavia and the Soviet Union; ZBIGNIEW K. BRZEZINSKI, *The Soviet Bloc: Unity and Conflict,* rev. ed. (1961); C. BRANDT, BENJAMIN SCHWARTZ, and J.K. FAIRBANK, *A Documentary History of Chinese Communism* (1952); VLADIMIR DEDIJER, *Tito Speaks* (1953); JANE DEGRAS (ed.), *Communist International Documents, 1919–1943,* 3 vol. (1956–65), a collection of documents with commentary; HERBERT FEIS, *Churchill, Roosevelt, Stalin* (1957), a well-documented study of wartime diplomacy; GUNTHER NOLLAU, *Die Internationale: Wurzeln und Erscheihungsformen des proletarischen Internationalismus* (1959; Eng. trans., *International Communism and World Revolution,* 1961); E. REALE, *Avec Jacques Duclos: Au Banc des Accusés* (1958); DAVID REES, *Korea: The Limited War* (1964); HUGH SETON-WATSON, *The Pattern of Communist Revolution,* rev. ed. (1961), a study of the rise of Communism in eastern Europe.

Developments after Stalin are surveyed in ADAM BROMKE (ed.), *The Communist States at the Crossroads, Between Moscow and Peking* (1965); ALEXANDER DALLIN (ed.), *Diversity in International Communism: A Documentary Record, 1961–63* (1963); HELENE CARRERE D'ENCAUSSE and STUART R. SCHRAM, *Le Marxisme et l'Asie, 1853–1964* (1965; Eng. trans., *Marxism and Asia: An Introduction with Readings,* 1969); EDWARD CRANKSHAW, *The New Cold War: Moscow v. Peking* (1963); GHITA IONESCU, *The Break-up of the Soviet Empire in Eastern Europe* (1965); WALTER LAQUEUR and LEOPOLD LABEDZ (eds.),

Polycentrism (1962), a valuable collection of essays on dissent in the Communist parties; WOLFGANG LEONHARD, *Kreml ohne Stalin* (1959; Eng. trans., *The Kremlin Since Stalin,* 1962); THE RUSSIAN INSTITUTE, COLUMBIA UNIVERSITY, *The Anti-Stalin Campaign and International Communism* (1956), an annotated text of Khrushchev's secret speech in 1956, with some other documents; HUGH SETON-WATSON, *The Imperialist Revolutionaries* (1978); H. GORDON SKILLING, *The Governments of Communist East Europe* (1966); MICHEL TATU, *Le Pouvoir en U.R.S.S.* (1967; Eng. trans., *Power in the Kremlin,* 1969); DONALD S. ZAGORIA, *The Sino-Soviet Conflict, 1956–61* (1962); MILORAD M. DRACHKOVITCH (ed.), *Fifty Years of Communism in Russia* (1968), a symposium by a number of experts on changes in communist doctrine; LEONARD SCHAPIRO (ed.), *The USSR and the Future* (1962), essays by specialists on various aspects of the party program of 1961; GUSTAV WETTER, *Dialektischer und historischer Materialismus* (1962; Eng. trans., *Soviet Ideology Today,* 1966).

Anarchism: The principal anarchist classics are mentioned in the text. In recent years many have been reprinted, notably WILLIAM GODWIN, *Enquiry Concerning Political Justice and Its Influence on Morals and Happiness,* 3 vol. (1793, reprinted 1946); PETER KROPOTKIN, *Memoirs of a Revolutionist,* 2 vol. (1899, reprinted 1968), and *Mutual Aid* (1890–96, reprinted 1939); and PIERRE JOSEPH PROUDHON, *Qu'est-ce que la propriété?* (1840; Eng. trans., *What Is Property?,* 2 vol., 1898–1902, reprinted 1966). G.P. MAXIMOFF (ed.), *The Political Philosophy of Bakunin: Scientific Anarchism* (1953), is a good anthology of Bakunin's writings; and VERNON RICHARDS (ed.), *Errico Malatesta: His Life and Ideas* (1965), contains a collection of writings by Malatesta, many of which were previously unavailable in English. TOLSTOY's principal writings in the anarchist vein, *The Kingdom of God Is Within You, The Slavery of Our Time,* and *What I Believe,* are available in standard editions. HERBERT EDWARD READ's anarchist writings are mainly collected in *Anarchy and Order* (1954, reprinted 1971) and *The Cult of Sincerity* (1968); his *Education Through Art* (1963) should also be consulted. The best general accounts of anarchist thought and history are GEORGE WOODCOCK, *Anarchism* (1962); JAMES JOLL, *The Anarchists* (1965); DANIEL GUERIN, *L'Anarchisme, de la doctrine à l'action* (1965; Eng. trans., *Anarchism: From Theory to Practice,* 1970); and APRIL CARTER, *The Political Theory of Anarchism* (1971). A useful collection of essays on contemporary trends is DAVID E. APTER and JAMES JOLL, *Anarchism Today* (1970); two recent anthologies that contain writings drawn from the whole libertarian tradition are LEONARD I. KRIMERMAN and LEWIS PERRY (eds.), *Patterns of Anarchy* (1966); and IRVING L. HOROWITZ (ed.), *The Anarchists* (1964). P. ELTZBACHER, *Der Anarchismus* (1900; Eng. trans., 1908), is a classic account of leading anarchist thinkers. ALAN RITTER, *Anarchism: A Theoretical Analysis* (1981), is a lucid, sympathetic study of the writings of major anarchistic theorists.

On anarchism in Russia: PAUL AVRICH, *The Russian Anarchists* (1967); FRANCO VENTURI, *Il populismo russo* (1952; Eng. trans., *Roots of Revolution: A History of the Populist and Socialist Movements in Nineteenth Century Russia,* 1960); and DAVID FOOTMAN, *Civil War in Russia* (1961). On anarchism in Spain: GERALD BRENAN, *The Spanish Labyrinth: An Account of the Social and Political Background of the Civil War,* 2nd ed. (1950, paperback 1960); BURNETT BOLLOTEN, *The Grand Camouflage: The Communist Conspiracy in the Spanish Civil War* (1961); and HUGH THOMAS, *The Spanish Civil War* (1961). On anarchism in Italy: RICHARD HOSTETTER, *The Italian Socialist Movement,* vol. 1 of a projected 3-volume work (1958). On anarchism in the United States: JAMES J. MARTIN, *Men Against the State: The Expositors of Individualist Anarchism in America, 1827–1908* (1953); and RUDOLF ROCKER, *Pioneers of American Freedom* (1949). On anarchism in France: EUGENIA W. HERBERT, *The Artist and Social Reform: France and Belgium, 1885–1898* (1961).

Fascism: H. ROGGER and E.J. WEBER (eds.), *The European Right: A Historical Profile* (1965); S.J. WOOLF (ed.), *European Fascism* (1969); H. KOHN, *Political Ideologies of the Twentieth Century,* 3rd rev. ed. (1966), a discussion of fascism and other ideologies in their historical setting; E. NOLTE, *Der Faschismus in seiner Epoche* (1963; Eng. trans., *Three Faces of Fascism,* 1965), the best theoretical analysis of fascism, and *Die Krise des liberalen Systems und die faschistischen Bewegungen* (1968), a continuation of his earlier work, with an extensive bibliog.; G.A. BORGESE, *Goliath, Der marsch des fascismus* (1938), on the historical and intellectual roots of Italian Fascism; D.L. GERMINO, *The Italian Fascist Party in Power: A Study in Totalitarian Rule* (1959); E. WISKEMANN, *Fascism in Italy: Its Development and Influence* (1969); R. DE FELICE, *Mussolini,* vol. 1, *Il rivoluzionario, 1883–1920* (1965); A.J. GREGOR, *The Ideology of Fascism* (1969), an analysis based upon the philosophy of Giovanni Gentile, with an extensive bibliog.; F.W. DEAKIN, *The Brutal Friendship: Mussolini, Hitler and the Fall of Italian Fascism* (1962); M. BAUMONT *et al.* (eds.), *The Third Reich* (1955), an authoritative symposium by European and American scholars; W. EBENSTEIN, *The Nazi State* (1943); H. RAUSCHNING, *Die revolution des nihilismus* (1938; Eng. trans., *The Revolution of Nihilism: Warning to the West,* 1940); H. MOHNE, *Der Orden unter dem Totenkopf,* 2 vol. (1969), a detailed study of the most militant Nazi organization, the SS; A. BULLOCK, *Hitler: A Study in Tyranny* (1952); R. HILBERG, *The Destruction of the European Jews* (1961); G. BRENAN, *The Spanish Labyrinth: An Account of the Social and Political Background of the Civil War* (1943); S.G. PAYNE, *Falange: A History of Spanish Fascism* (1961); H. THOMAS, *The Spanish Civil War* (1961); C.A. MACARTNEY, *A History of Hungary, 1929–1945,* 2 vol. (1956–57); J. PLUMYENE and R. LASIERRA, *Les Fascismes Français, 1923–1963* (1963); E.J. WEBER, *Varieties of Fascism: Doctrines of Revolution in the 20th Century* (1964); H.F. MACNAIR, *The Real Conflict between China and Japan: An Analysis of Opposing Ideologies* (1938); A. DEL BOCA and M. GIOVANA, *I "figli del sole": mezzo seculo di nazifascismo nel mondo* (1965; Eng. trans., *Fascism Today: A World Survey,* 1969), a well-documented comprehensive survey of the resurgence of fascism after 1945; *Who Were the Fascists: Social Roots of European Fascism,* ed. by STEIN LARSON *et al.* (1981), fifty essays on fascism in over twenty countries.

Nationalism: The origins and development of nationalism as a political idea are discussed in detail in HANS KOHN, *The Idea of Nationalism* (1944, reprinted 1967). Another discussion of the development of nationalism, including various interpretations in modern times, is analyzed in LOUIS L. SNYDER, *Varieties of Nationalism* (1976). A sociological treatment of the subject is ANTHONY D. SMITH, *Theories of Nationalism* (1971). HUGH SETON-WATSON, *Nations and States* (1977), is a detailed, worldwide study. A new approach to the problem of nationalism was introduced by KARL W. DEUTSCH, *Nationalism and Social Communication* (1953), and *Tides Among Nations* (1979), an overview of the development of his thinking. On nationalism in the Third World, see RUPERT EMERSON, *From Empire to Nation: The Rise to Self-Assertion of Asian and African Peoples* (1960); and SELIG S. HARRISON, *The Widening Gulf* (1978). There are two valuable bibliographies: KOPPEL S. PINSON, *A Bibliographical Introduction to Nationalism* (1935); and KARL W. DEUTSCH, *Interdisciplinary Bibliography on Nationalism* (1956).

Liberalism: General works include R.D. CUMMING, *Human Nature and History: A Study of the Development of Liberal Political Thought* (1969); H.K. GIRVETZ, *Evolution of Liberalism,* rev. ed. (1963), a comparison of classical and contemporary liberalism that examines the psychological assumptions underlying the political and economic views of the classical liberals; L. HARTZ, *The Liberal Tradition in America: An Interpretation of American Political Thought Since the Revolution* (1955); K. MARTIN, *The Rise of French Political Thought,* 2nd ed. K.R. MINOGUE, *The Liberal Mind* (1963), on the contention that liberalism now reflects a moral and political consensus; W.A. ORTON, *The Liberal Tradition: A Study of the Social and Spiritual Conditions of Freedom* (1945); F.M. WATKINS, *The Political Tradition of the West: A Study in the Development of Modern Liberalism* (1948).

Classical liberalism is treated in A.V. DICEY, *Lectures on the Relation Between Law and Public Opinion in England During the Nineteenth Century* (1905); E. HALEVY, *La formation du radicalisme philosophique,* 3 vol. (1901–04; Eng. trans., *The Growth of Philosophical Radicalism,* 1928); L.T. HOBHOUSE, *Liberalism* (1911); H.J. LASKI, *The Rise of Liberalism: The Philosophy of a Business Civilization* (1936); GUIDO DE RUGGIERO, *Storia del liberalismo europeo* (1925; Eng. trans., *The History of European Liberalism,* 1927); LESLIE STEPHEN, *The English Utilitarians* (1900).

Works on contemporary liberalism include A.A. BERLE and G.C. MEANS, *The Modern Corporation and Private Property* (1933), the classic account of how the private corporation has generated "managerial" power; W.H. BEVERIDGE, *Full Employment in a Free Society* (1945), by the chief English architect of the "welfare" state; G.D.H. COLE, *Economic Planning* (1935); C.A.R. CROSLAND, *The Future of Socialism* (1957), by a leading theoretician of the British Labour Party who believes that the Keynesian revolution has made many traditional socialist attitudes obsolete; JOHN DEWEY, *Human Nature and Conduct: An Introduction to Social Psychology* (1922), the best criticism of the psychological preconceptions of the classical liberals, *Liberalism and Social Action* (1935); JOHN KENNETH GALBRAITH, *American Capitalism: The Concept of Countervailing Power,* 2nd ed. rev. (1956), *The Affluent Society* (1958), one of the first works to contend that liberal emphasis should shift from greater productivity to the question of priorities, *The New Industrial State* (1967); A.H. HANSEN, *Economic Policy and Full Employment* (1947), by a leading American exponent of Keynesianism, *Fiscal Policy and Business Cycles* (1941); S.E. HARRIS, *Economic Planning* (1949), an examination of economic planning in ten

European countries as well as in the U.S., Argentina, Japan, and India; JOHN KEYNES, *The End of Laissez-faire* (1926), *The General Theory of Employment, Interest and Money* (1935–6), the most influential economic tract of the first half of this century; W.A. LEWIS, *The Principles of Economic Planning* (1951), on the difference between direct and indirect planning, and its important implications; JAMES E. MEADE, *Planning and the Price Mechanism* (1949), a liberal socialist approach to the problems of planning and use of the price mechanism in postwar Britain; GUNNAR MYRDAL, *Beyond the Welfare State* (1960), on the social impact of economic planning in highly industrialized states and the economic relations between them and less developed economies, *Challenge to Affluence* (1963); E.V. ROSTOW, *Planning for Freedom* (1959), an analysis of the impact of the Keynesian "revolution" on government intervention in the economic sphere. See also ANDREW LEVINE, *Liberal Democracy: A Critique of its Theory* (1981).

Conservatism: Among "Burkean-conservative" works sharing an anti-extremist centre with moderate liberals are DANIEL J. BOORSTIN, *The Genius of American Politics* (1953); THOMAS I. COOK and MALCOLM MOOS, *Power Through Purpose* (1954); ERIC HOFFER, *The True Believer* (1951); ROSS J.S. HOFFMAN and PAUL LEVACK (eds.), *Burke's Politics* (1949); HENRY A. KISSINGER, *A World Restored* (1957); WALTER LIPPMANN, *The Cold War* (1947); REINHOLD NIEBUHR, *The Irony of American History* (1952), *Christian Realism and Political Problems* (1953), and *The Self and the Dramas of History* (1955); ROBERT A. NISBET, *Community and Power* (1962); CLINTON L. ROSSITER, *Conservatism in America,* 2nd ed. rev. (1962); GEORGE SANTAYANA, *Dominations and Powers* (1951); LEO STRAUSS, *Natural Right and History* (1953) and *What Is Political Philosophy? and Other Studies* (1959); FRANK TANNENBAUM, *A Philosophy of Labor* (1951); PETER VIERECK, *Conservatism Revisited* (1949; paperback rev. ed., 1965), *The Unadjusted Man* (1956), and *Shame and Glory of the Intellectuals* (1953; paperback rev. ed., 1965); ERIC VOEGELIN, *The New Science of Politics* (1952); and *Order and History,* 3 vol. (1956–58); FRANCIS G. WILSON, *The Case for Conservatism* (1951). For British Conservatives, see LEOPOLD AMERY, *The Forward View* (1935); ARTHUR BRYANT, *The Spirit of Conservatism* (1929); LORD HUGH CECIL, *Conservatism* (1912); HENRY FAIRLIE, *The Life of Politics* (1968); QUINTIN HOGG, *The Conservative Case,* rev. ed. (1959); F.J.C. HEARNSHAW, *Conservatism in England* (1933); MICHAEL J. OAKESHOTT, *Rationalism in Politics, and Other Essays* (1962); and PEREGRINE WORSTHORNE, *The Socialist Myth* (1971). BURTON Y. PINES, *Back to Basics* (1982), an account of the conservative resurgence of the 1970s in the U.S.

More extreme views, whether right-wing nationalist in politics or militantly business-oriented in economics (mainly by contributors to the contemporary New York periodical *National Review*), may be found in WILLIAM F. BUCKLEY, JR., *God and Man at Yale: The Superstitions of Academic Freedom* (1951) and *Up from Liberalism,* rev. ed. (1968); JAMES BURNHAM, *Congress and the American Tradition* (1959) and *Suicide of the West* (1964); MILTON FRIEDMAN, *Capitalism and Freedom* (1962); BARRY GOLDWATER, *Conscience of a Conservative* (1963); JEFFREY P. HART, *The American Dissent* (1966); NELLIE D. KENDALL (ed.), *Willmoore Kendall contra mundum* (1971); RUSSELL KIRK, *The Conservative Mind, from Burke to Santayana* (1953), *Prospects for Conservatives* (1956), and *Enemies of the Permanent Things* (1969); ERIK VON KUEHNELT-LEDDIHN, *Liberty or Equality* (1952); FRANK S. MEYER, *In Defense of Freedom* (1962) and *The Conservative Mainstream* (1969); THOMAS S. MOLNAR, *The Counter-Revolution* (1969); ENOCH POWELL, *Freedom and Reality* (1969); RONALD REAGAN, *The Creative Society* (1968).

Useful anthologies include PETER VIERECK (ed.), *Conservatism: From John Adams to Churchill* (paperback rev. ed. 1956); and PETER WITONSKI (ed.), *The Wisdom of Conservatism* (1971).

Socrates

Socrates of Athens, who flourished in the last half of the 5th century BC, was the first of the great trio of ancient Greeks—Socrates, Plato, and Aristotle—who laid the philosophical foundations of Western culture. As Cicero said, Socrates "brought down philosophy from heaven to earth"—*i.e.,* from the nature speculation of the Ionian and Italian cosmologists to analyses of the character and conduct of human life, which he assessed in terms of an original theory of the soul. Living during the chaos of the Peloponnesian War, with its erosion of moral values, Socrates felt called to shore up the ethical dimensions of life by the admonition to "know thyself" and by the effort to explore the connotations of moral and humanistic terms.

LIFE

Socrates was born in or about 470 BC, 10 years after the Battle of Salamis. His father, Sophroniscus, was a friend of the family of Aristides the Just, founder of the Delian League, from which the empire arose. The tale that his father was a sculptor rests on Plato's reference to the mythical sculptor Daedalus as the ancestor, or work-lineage, of Socrates. Although the philosopher's mother, Phaenarete, acted as a "midwife," this fact implies nothing about her social status.

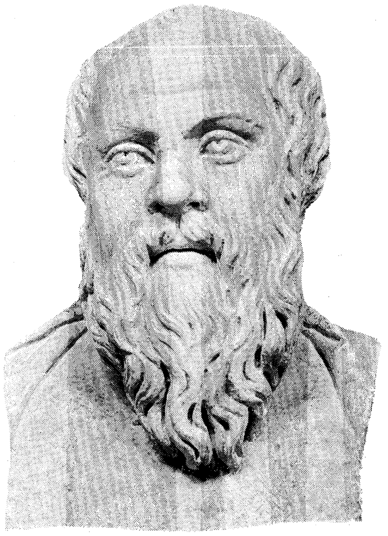
Early life and connections

The memoir writer Ion of Chios mentioned meeting Socrates at Samos in the company of the philosopher Archelaus, a pupil of Anaxagoras (Athens' first philosopher), presumably during the military operations of 441–439. The connection between the two men is also asserted by the musicologist Aristoxenus, whereas the tradition of commentaries based on Theophrastus, Aristotle's successor, calls Socrates the "disciple" of Archelaus.

Plato and Aeschines the Socratic, both writers of Socratic dialogues, agree with the military historian Xenophon in depicting him as intimate with the leading figures of the Periclean circle (Aspasia, Alcibiades, Axiochus, Callias), dominant in Athens at the time. Xenophon concurs with Plato in saying that he was well versed in both geometry and astronomy, and this representation of Socrates agrees with the narrative of Plato's *Phaedo* as well as the burlesque *The Clouds,* which was written by the playwright Aristophanes.

Socrates must already have been a conspicuous figure at Athens when Aristophanes and Ameipsias both made him the subject of their comedies in 423, and, because they made a special point of his neediness, he had probably suffered recent losses. (The marked poverty of his old age is said in Plato's *Apology* to have been caused by his preoccupation with his mission to mankind.)

Socrates was married, apparently late in life, to Xan-



Socrates, herm with a restored nose probably copied from the Greek original by Lysippus, *c.* 350 BC. In the Museo Archeologico Nazionale, Naples.

By courtesy of the Soprintendenza alle Antichita della Campania, Naples thippe, by whom he left three sons, one an infant. Xenophon speaks of her high temper; there is no evidence, however, that she was a "shrew"; the sons, according to Aristotle, proved insignificant.

Socrates' record for endurance was distinguished. He served as a hoplite, perhaps at Samos (440), and at several stations during the Peloponnesian War. (At Potidaea he saved the life of Alcibiades.) In politics he took no part, knowing, as he told his judges, that office would mean compromise with his principles. Once at least, in 406–405, he was a member of the Boule, or legislative council, of 500; and at the trial of the victors of Arginusae, he resisted, at first with the support of his colleagues, afterward alone, the unconstitutional condemnation of the generals by a collective verdict. He showed the same courage in 404, when the oligarchy of the Thirty Tyrants in Athens, wishing to implicate honourable men in their proceedings, instructed him and four others to arrest Leon, one of their victims. Socrates disobeyed, and he says in Plato's *Apology* that this might have cost him his life but for the counterrevolution of the next year. (For the background to these events see GRECO-ROMAN CIVILIZATION.)

Indictment, trial, and death

In 399 Socrates was indicted for "impiety." The author of the proceedings was the influential Anytus, one of the two chiefs of the democrats restored by the counterrevolution of 403; but the nominal prosecutor was the obscure and insignificant Meletus. There were two counts in the accusation, "corruption of the young" and "neglect of the gods whom the city worships and the practice of religious novelties." Socrates, who treated the charge with contempt and made a "defense" that amounts to avowal and justification, was convicted, probably by 280 votes against 220. The prosecutors had asked for the penalty of death; it now rested with the accused to make a counterproposition. Though a smaller, but substantial, penalty would have been accepted, Socrates took the high line that he really merited the treatment of an eminent benefactor: maintenance at the public table. He consented only for form's sake to suggest the small fine of one mina, raised at the entreaty of his friends to 30.

The claim to be a public benefactor incensed the court, and death was voted by an increased majority, a result with which Socrates declared himself well content. As a rule at Athens, the condemned man "drank the hemlock" within 24 hours, but, in the case of Socrates, the fact that no execution could take place during the absence of the sacred ship sent yearly to Delos caused an unexpected delay of a month, during which Socrates remained in prison, receiving his friends daily and conversing with them in his usual manner. An escape was planned by his friend Crito, but Socrates refused to hear of it, on the grounds that the verdict, though contrary to fact, was that of a legitimate court and must therefore be obeyed. The story of his last day, with his drinking of the hemlock, has been perfectly told in the *Phaedo* of Plato, who, though not himself an eyewitness, was in close touch with many of those who were present.

Main sources of information. Socrates wrote nothing; therefore, information about his personality and doctrine has to be sought chiefly in the dialogues of Plato and in the *Memorabilia* of Xenophon. As both men were nearly 45 years younger than Socrates, they could speak from firsthand knowledge about only the last 10 to 12 years of his life.

Xenophon, whose relations with Socrates seem not to have been close, has even been suspected of drawing from Plato. His admitted deficiencies in imagination and capacity for thinking do not make him the more faithful exponent of a philosophical genius. Moreover, Xenophon's apologetic purpose calls for some discounting. His most valuable statements are those that appear to be most at variance with his main thesis, viz., that the prosecutors of Socrates were mistaken even from their own point of view.

Plato's depiction of Socrates

Plato's more vivid picture has been suspected on the grounds that he used Socrates as a "mouthpiece" for speculations of his own: the theory of "Ideas" or doctrine of "Forms" is thus held to have been originated by Plato. There are serious reasons for denying this assumption, though they have not yet convinced many scholars; in any case, to employ it, without investigation, to discredit Plato's testimony begs the question.

In some important respects, Plato's testimony is confirmed by the extant writings of Aeschines Socraticus. *The Clouds* of Aristophanes yields valuable information about Socrates in his middle 40s, though allowance must be made for the work's character as a burlesque. It should be compared carefully with the autobiographical statements put into the mouth of Socrates in the *Phaedo,* which, though not "contemporary evidence," are clearly meant to express Plato's bona fide belief about his master's intellectual history. Whichever way the evidence is interpreted, however, a Platonic view of Socrates is available in the *Dialogues* that is valuable in its own right; hence the following discussion will draw heavily upon the *Dialogues* as the primary source for the portrait of Socrates.

Personal characteristics. Though Socrates was a good fighting man, his outward appearance was grotesque. Stout and not tall, with prominent eyes, snub nose, broad nostrils, and wide mouth, he seemed a very Silenus. But, as his friends knew, he was "all glorious within," "the most upright man of that day" (Plato, *The Seventh Letter* [324*e*]). His self-control and powers of endurance were exemplary; "he had so schooled himself to moderation that his scanty means satisfied all his wants."

But Socrates was no self-tormenting ascetic: he "knew both how to want and how to abound" and could be the soul of the merriment at a gay party. He had no sympathy with the slatternliness of his friend Antisthenes nor with the godly dirtiness often affected by the followers of Pythagoras (the philosopher of number). There was nothing of the complacent self-righteousness of the Pharisee nor of the angry bitterness of the satirist in his attitude toward the follies or even the crimes of his fellowmen. It was his deep and lifelong conviction that the improvement not only of himself but also of his countrymen was a task laid upon him "by God," not to be executed, however, with a scowling face and an upbraiding voice. Like St. Francis Xavier, he understood that to win men's souls one must be "good company." Conscious of his own infirmities, he felt a profound sympathy for the intemperate.

Socrates was a true patriot who felt that he could best prove his devotion to Athens by setting his face resolutely against the attractions of specious and popular, but deadly false theories of public and private morality. When the city brought him to trial and threatened him with death, his sense of civic duty forbade him to escape into exile either before or after the trial. It was his very patriotism that made him an unsparing critic of the Athenian "democracy" and so led to his being condemned to death.

The Socratic irony

Nothing was more marked in his character than an unusually keen appreciation of the comic in human nature and conduct that protected him at once against sentimentality and against cynicism. This is what his opponents in Plato call his "irony" and treat as an irritating affectation. "Intellectually the acutest man of his age, he represents himself in all companies as the dullest person present. Morally the purest, he affects to be the slave of passion" (W.H. Thompson). No doubt, in part, this irony was "calculated"; it "disarmed ridicule by anticipating it." But its true source is the spontaneous sense of fun that makes its possessor the enemy of all pretentiousness, moral or intellectual. And it is certain that, though the purity of Socrates is beyond question, he really had an ardent and amorous temperament.

Religious beliefs. Socrates was clearly a man of deep piety with the temperament of a mystic. He regarded mythology, with its foolish or immoral tales about gods, as a mere invention of the poets. But he found it easy to combine his own strong belief in God as ruler of the world with the view that, in practice, one could worship God in the way prescribed by "the usage of the city." God's existence is shown, he held, not only by the providential order of nature and the universality of the belief in him but also by warnings and revelations given in dreams, signs, and oracles. The soul of man partakes of the Divine; and, as Plato argued in the *Phaedo,* Socrates believed in the soul's immortality. Aristophanes makes Socrates combine the parts of "infidel" physicist and hierophant of a mysterious private faith and, in *The Birds,* presents him as presiding at a fraudulent séance. He was regular, says Xenophon, in prayer and sacrifice, though he held that, because only the gods know what is good for a man, his prayer should simply be "give me what is good." It is clear from Plato that Socrates was quite familiar with Pythagorean and Orphic religious ideas—with the doctrine of the divine origin and destiny of the soul, for example—though he regarded the ordinary Orphic mystery monger with healthy contempt.

The evidence that Socrates had a markedly "mystical" temperament is abundant. Plato tells of his curious "rapts," in one of which he stood spellbound for 24 hours in the trenches. The accounts of the philosopher's "divine sign" tell the same story. This, according to Plato, was a "voice" often heard by Socrates from childhood. It forbade him to do things but never gave positive encouragement. According to Plato, it merely gave prognostications of good or bad luck, and the occasions of its occurrence were often "very trivial." Thus, it was neither an intuitive conscience nor a symptom of mental disorder but an interior psychic audition.

The test of the Delphic oracle

Mode of life. Socrates seemed to spend all his time in the streets, the marketplace, and, more particularly, the *gymnasia.* He cared little for the country. Though he frequented by choice the society of young men of promise, he also talked freely to politicians, poets, and artisans about their various callings, their notions of right and wrong, the familiar matters of interest to them. The object of all this dialogue was to test the famous oracle of Apollo at Delphi, which had pronounced him the wisest of men. This pronouncement was made before Socrates had become conscious of his mission to his fellowmen: even at that early date, it is implied, he had the highest of reputations in circles interested in wisdom. (This early date is attested by the fact that the Eleatics from Megara and the young pupils of the Pythagoreans from Thebes and Phlious who were attached to Socrates must have formed their connection with him before the Peloponnesian War.)

Socrates set himself to convict "the god" of falsehood. But finding that those who thought themselves wise were unable to give any coherent account of their wisdom, Socrates had to admit that he was wiser than others, just because he alone was aware of his own ignorance. This account is plainly tinged with the usual "irony." Socrates took the Delphic oracle seriously enough to probe into its real import. He believed himself charged with a mission from God to make his fellowmen aware of their ignorance and of the supreme importance of knowledge of what is for the soul's good. This is proved by his declaration that he was more than ready to face instant death rather than to neglect his commission.

The poverty in which this mission had involved him and the austerity of the rule of life that it entailed were notorious. Summer and winter, Socrates' coat was the same; he had neither shoes nor shirt. "A slave who was made to live so," the Sophist Antiphon said, "would run away." This self-imposed life of hardships was the price of his spiritual independence.

His message, however, was variously received. Some of those whose false pretensions were exposed by his trenchant criticizing regarded him with ill will; many thought him an officious busybody. Among the younger men, many merely thought it good sport to see their elders silenced. Others, such as Alcibiades and Critias, deliberately attached themselves to him for a time "for private ends," believing that to learn the secret of so acute a reasoner would be the best preparation for success in the law courts, the council, and the assembly. Others sincerely hoped by associating with him to become good men and true, capable of doing their duty by house and household, by relations and friends, by city and fellow citizens. Finally, there was an inner circle that entered more deeply into Socrates' principles and transmitted them to the next generation. But these were not "disciples" united by a common doctrine. The bond of union was a common reverence for a great man's intellect and character. It was, in the main, this group—many from states that had been enemies of Athens in the recent war—that collected around Socrates on the day of his death.

His association with political enemies

The accusation and its causes. The explanation of the attack made on Socrates is simple. He had been on terms of close friendship with the two men whose memories were most obnoxious to the democrats: Critias, the fiercest spirit among the extremists of the "terror" of 404; and Alcibiades, whose self-will had done so much to bring about the downfall of the Athenian empire. The charge of "educating Alcibiades" was made prominent in the pamphlet written a few years after the trial by the Sophist Polycrates, in justification of the verdict. More than half a century later, the orator Aeschines reminds his audience that Socrates had been put to death because he was believed to have educated Critias. In point of fact, it was absurd to make Socrates responsible for the ambitions of Alcibiades, and, as he reminded his judges, he had disobeyed an illegal order from Critias and his colleagues at the risk of his life. But it is natural that he should have had to suffer for the crimes of both men, the more so because he had been an unsparing critic of democracy and of the famous democratic leaders and, furthermore, had not, like the advanced democrats, withdrawn from Athens during the "terror."

Socrates was, in fact, suspected of using his great abilities and gifts to pervert his younger associates from loyalty to the principles of democracy, and the convinced democrats who had recovered the city in 403 were unwilling, as J. Burnet has said, "to leave their work at the mercy of reaction." The motives of Anytus, an upright, unintelligent democrat, are thus quite explicable: from his point of view, Socrates would be at the best a moderate oligarch, and democrats who remembered the career of the statesman Theramenes, who had tried to mix oligarchy and democracy, could not be expected to make a fine distinction between the moderate oligarch and the traitor.

The real grounds for the attack could not be disclosed in the indictment because of the amnesty that had terminated the struggle, of which Anytus himself had been a main promoter. Hence, the charge took the form of a vague accusation of "corruption of the young." Probably for the same reasons, Anytus was ashamed to appear as the principal in the matter and put forward the obscure Meletus, who might venture on "indiscretions" more openly. If this was the same Meletus who prosecuted Andocides on the same charge of "impiety," he must have been a half-witted fanatic—and this may explain why the charge of irreligion was added. Xenophon suggests that the allusion was to the "divine sign," but this cannot be correct. Meletus said nothing about the "sign" at the prosecution, and Socrates is speaking with his "usual irony" when he pretends to guess that the mention of "religious novelties" in the indictment referred to the "sign." In the *Apology,* Socrates says that the prosecution is, no doubt, relying on memories of Aristophanes' *The Clouds,* where he had been made to talk "atheism" as part of the burlesque on men of science.

Meaning of the charge against him and his conviction

But there must have been more behind the charge. It seems likely that the prosecution of Andocides revived the old scandal of the "profanation of the mysteries" that had thrown Athens into a ferment on the eve, in 415, of the Sicilian expedition. The two chief victims, Alcibiades and his uncle Axiochus, had both been among the intimates of Socrates, and there is reason to think that others of his friends were affected. If this is what lay behind the charge, it can be understood why its real meaning seems never to have been explained: for in view of the terms of the amnesty, the matters in question were not within the competence of the court.

Socrates himself treats the whole matter with contempt. His defense consists in narrating the facts of his past life, which had proved that he was equally ready to defy the populace and the Thirty in the cause of right and law, and in insisting on the reality of his mission from God and his determination to discharge it, even at the cost of life. The prosecutors had no desire for blood. They counted on a voluntary withdrawal of the accused from the jurisdiction before trial; the death penalty was proposed to make such a withdrawal certain. Socrates himself forced the issue by refusing at any stage to do anything involving the least shade of compromise. The prosecution had raised the question whether he was a traitor or, as he held himself to be, an envoy from God; Socrates was determined that the judges should give a direct verdict on the issue without evasion. This is not only what makes him a martyr but also what forbids us to call Anytus a murderer.

DOCTRINE AND METHOD

Socrates was a man of the Periclean age, which witnessed one of the periodic "bankruptcies of science." Cosmological speculation, which had been boldly pursued from the beginning of the 6th century, seemed to have led to a chaos of conflicting systems of thought. The Rationalist Parmenides of Elea had apparently cut away the ground from science by showing that the real world must be quite unlike anything that the senses reveal and that, consequently, the interpretation of the world by familiar analogies is inherently fallacious; and his pupil Zeno of Elea seemed to have shown that even the postulates of mathematics are mutually contradictory. Thus, the ablest men, such as the Sophists Protagoras and Gorgias, had turned away from the pursuit of science and concerned themselves not with truth but with making a success of human life.

Socrates' early interest in "natural science"

Socrates, as a young man, was enthusiastically interested in "natural science" and familiarized himself with the various current systems—with the Milesian cosmology with its flat Earth and the Italian with its spherical Earth and with the mathematical puzzles raised by Zeno about "the unit" (*i.e.,* the problem of continuity). There was a complete lack of critical method. For a moment, Socrates hoped to find salvation in the doctrine of Anaxagoras that "Mind" is the source of all cosmic order because this seemed to mean that "everything is ordered as it is best that it should be," that the universe is a rational teleological system. But on reading the book of Anaxagoras, he found that the philosopher made no effective use of his principle; the details of his scheme were as arbitrary as those of any other.

The Socratic "hypothesis." After this disappointment, Socrates resolved from then on to consider primarily not "facts" but *logoi,* the "statements" or "propositions" that one makes about "facts." His method would be to start with whatever seemed the most satisfactory "hypothesis," or postulate, about a given subject and then consider the consequences that follow from it. So far as these consequences proved to be true and consistent, the "hypothesis" might be regarded as provisionally confirmed. But one should not confuse inquiry into the consequences of the "hypothesis" with proof of its truth. The question of truth could be settled only by deducing the initial "hypothesis" as a consequence from some more ultimate, accepted "hypothesis."

The doctrine of Forms. According to Plato, Socrates next proceeded to take it as his own fundamental "hypothesis" that every term (such as "good," "beautiful," "man") that has an unequivocal denotation directly names a self-same object of a kind inaccessible to sense perception and apprehensible only by thought. Such an object Socrates calls an *Idea* or *Eidos; i.e.,* a Form. The sensible things on which a man predicates beauty, goodness, humanity, have only a secondary and derivative reality; they *become* this or that for a time, in virtue of their "participation" in the Form.

Whether the Forms are Socratic

Scholars in the 19th century usually assumed that this doctrine of Forms was consciously devised by Plato after the death of Socrates. The chief argument for this view is based upon the observation of Aristotle that Socrates rightly "did not separate" the universal from the particular as, it is apparently implied, Plato did. He might equally have meant, however, that the doctrine of the *Phaedo* does not itself involve the kind of "separation" to which he objects in the Platonic theory. On the other side, the doctrine is expressly said in the *Phaedo* to be a familiar one, which Socrates "was always" repeating; and, if untrue, it is hard to see what could be the point of such a mystification and harder to understand how Plato could have expected it to be successful, especially as most of the personages of the *Phaedo* were certainly still alive. If true, however, one must be prepared to admit the possibility that he is also reproducing the thought of Socrates in the *Symposium* and *Republic,* in which he speaks of a supreme Form, that of Beauty, or Good, the vision of which is the far-off goal of all intellectual contemplation. Unfortunately, no complete separation of the Socratic and the Platonic is possible.

Logical methods. On the logical side, both Plato and Xenophon bear out the remark of Aristotle that Socrates may fairly be credited with two things: "inductive arguments" and "universal definitions." The "universal definition" is an attempt to formulate precisely the meaning of a universally significant predicate—*i.e.,* to apprehend what the *Phaedo* calls a Form. And it is from the practice of Socrates, who aimed at the clarification of thought about the meaning of moral predicates as the first indispensable step toward the improvement of practice, that the theory of logical division and definition, as worked out by Plato and Aristotle, has arisen.

The "inductive arguments" mean the characteristic attempts to arrive at such formulations by the consideration of simple and striking concrete illustrations, the perpetual arguments about "shoemakers and carpenters and fullers," which the fashionable speakers in Plato profess to think vulgar. Induction, on this view of it, is not regarded as a method of proof; its function is that of suggestion: it puts the meaning of a proposed "definition" forcibly and clearly before the mind. The justification of the definition, then, has to be sought in a consideration of the satisfactoriness of the "consequences" that would follow from its adoption. Socrates himself sought for his "definitions" principally in the sphere in which he was most interested: as Aristotle says, he concerned himself with the "ethical," character and conduct, both private and public, not with "nature" at large.

Ethics and politics. With Socrates the central problem of philosophy shifted from cosmology to the formulation of a rule of life, to the "practical use of reason." As the *Apology* relates, the specific message from God that Socrates brought to his fellowmen was that of the "care" or "tending" of one's "soul," to "make one's soul as good as possible"—"making it like God," in fact—and not to ruin one's life, as most men do, by putting care for the body or for "possessions" before care for the "soul"; for the "soul" or *psychē* is that which is most truly a man's self.

His original view of the soul

Socrates' view of the soul stands in sharp contrast with the Homeric and Ionian view of the *psychē* as "the breath of life," which is given up when the man "himself," his body, has perished, and also with the view prevalent in circles influenced by Orphic-type religions, according to which the soul is a sort of stranger loosely inhabiting the body, which "sleeps while the body is active, but wakes when the body sleeps"; instead, the soul came in the 4th century to be viewed as the normal walking personality, the seat of character and intelligence, "that," as Socrates says in Plato, "in virtue of which we are called wise or foolish, good or bad." And as this usage of the word first appears in writers who are known to have been influenced by Socrates (Isocrates, Plato, and Xenophon), it may fairly be ascribed to his influence. Thus the soul *is* the man.

A man's happiness or well-being, in Socrates' view, depends directly on the goodness or badness of his soul. No one ever wishes for anything but true good—*i.e.,* true happiness. But men miss their happiness because they do not *know* what it is. For real good they mistake things that are not really good (*e.g.,* unlimited wealth or power). In this sense, "all wrong-doing is involuntary." Men need to *know* true good and not confuse it with anything else, so as to keep from *using* strength, health, wealth, or opportunity wrongly. If a man has this knowledge, he will always act on it, since to do otherwise would be to prefer known misery to known happiness. If a man really knew, for instance, that to commit a crime is worse than to suffer loss or pain or death, no fear of these things would lead him to commit the crime. To the professional Sophist, "goodness" is a neutral "accomplishment" that can always be put to either of two uses, a good one or a bad one. To Socrates, in contrast, knowledge of good is the one knowledge of which it is impossible to make an ill use; the possession of it is a guarantee that it will always be used properly. Thus, Socrates becomes—as against the relativism of Protagoras—the founder of the doctrine of an absolute morality based on the conception of a felicity that is the good not of Athenians or Spartans or even of Greeks but of man as man, as part of universal humanity.

His political philosophy

Politics, from this point of view, is the statesman's task of "tending" the souls of all his fellow citizens and making them "as good as possible." The knowledge of good is also the foundation of all statesmanship. The radical vice of ancient democracy, according to Socrates, is that of putting society in the hands of men without true insight and with no adequate expert knowledge. His main criticism, however, is that, though in some departments democracy takes the advice only of a qualified expert, on questions of morality and justice it treats any one citizen's opinion as of equal value with another's.

Even a Themistocles or a Pericles plainly had no knowledge of true statesmanship: they gave the populace the things that tickled its taste, such as a navy and a commerce; but they were no "physicians of the body politic," for they did not promote "righteousness and temperance," the spiritual health of the community. Socrates maintained that he alone deserved the name of statesman, because he understood, as the men of action did not, that knowledge of the absolutely good is the necessary and sufficient condition of national well-being and felicity. Indeed, Plato's *Republic* may fairly be viewed as a picture of life in a society governed by this Socratic conviction. How far any of the special regulations of the *Republic* embody actual convictions of Socrates is more than can be said, though it is significant that the *Aspasia* of Aeschines represents Socrates as maintaining one of Plato's "paradoxes," the capacity of women for war and for politics.

Socrates exerted a temporary influence on a group of men who have come to be known as "the minor Socratics," among whom the most important were Antisthenes of Athens and Eucleides of Megara, with whom the Cynics and the Megarian school were connected. It was mainly through his influence on Plato, however, who took up the thought of Socrates and continued it in his own life's work, that Socrates' efforts bore their full fruit for subsequent ages (see PLATONISM and PHILOSOPHICAL SCHOOLS AND DOCTRINES: *Platonism*). (A.E.Ta.)

BIBLIOGRAPHY

General studies: A thorough and balanced introduction to Socrates is W.K.C. GUTHRIE, *A History of Greek Philosophy,* vol. 3, *The Fifth-Century Enlightenment,* pp. 323–507 (1969). Socrates' cultural and philosophical context is concisely described in FRANCIS M. CORNFORD, *Before and After Socrates* (1932, reprinted 1979); and in W.K.C. GUTHRIE, *The Greek Philosophers: From Thales to Aristotle,* ch. 4 (1950, reissued 1975), on the reaction toward humanism. Older classical accounts by JOHN BURNET, *Early Greek Philosophy,* 4th ed. (1930, reissued 1963); A.E. TAYLOR, *Socrates* (1932, reissued 1979); and CONSTANTIN RITTER, *Sokrates* (1931, in German), ascribe to Socrates himself much of Plato's formal theory of Ideas while also giving sympathetic portraits of his personality and thought.

The historical Socrates: Many studies centre on the historical value of the principal informants about Socrates: Plato and Xenophon, Aristotle and Aristophanes. The main source materials and problems are presented by RICHARD L. LEVIN and JOHN BREMER, *The Question of Socrates* (1961); a brief informative survey of positions comes from C.J. DE VOGEL, *Philosophia,* part 1, *Studies in Greek Philosophy,* ch. 5 (1970). Reliable accounts of the Socrates–Plato relationship are given by GUY C. FIELD, *Plato and His Contemporaries,* 3rd ed. (1967); and of Socrates' last days by COLEMAN PHILLIPSON, *The Trial of Socrates* (1928). Historical facts are skeptically regarded and the functions of myth and political polemic on the part of the sources are stressed by OLOF A. GIGON, *Sokrates: Sein Bild in Dichtung und Geschichte,* 2nd ed. (1979); A.H. CHROUST, *Socrates, Man and Myth: The Two Socratic Apologies of Xenophon* (1957), with special criticism of Xenophon and Antisthenes; and the two reporting dissertations by V. DE MAGALHÃES-VILHENA: *Le Problème de Socrate* (1952), and *Socrate et la légende platonicienne* (1952). There are some scholarly defenses and methodological interpretations of the historicity of the major Socratic sources: AUGUSTE DIÈS, *Autour de Platon,* 2 vol. (1927, reprinted 1975); T. DEMAN, *Le Témoignage d'Aristote sur Socrate* (1942); KENNETH J. DOVER's introduction and commentary to his edition of Aristophanes, *Clouds* (1968); and HUGH TREDENNICK's introduction to his translation of Xenophon, *Memoirs of Socrates and the Symposium (The Dinner Party)* (1970).

Socrates in history: The personality, ideals, and examples of Socrates are a perennial inspiration for subsequent thinkers and traditions. His influence is visually manifested in the many illustrations and texts of MICHELINE SAUVAGE, *Socrates and the Human Conscience,* (1960; originally published in French, 1956); as well as in the collection of testimonies gathered by HERBERT SPIEGELBERG in *The Socratic Enigma* (1964). His educational impact is measured by WILLIAM K. RICHMOND, *Socrates and the Western World* (1954), while his deep impress upon Christian religious awareness and Existential thinking is affirmed in ROMANO GUARDINI, *The Death of Socrates* (1948, reissued 1965; originally published in German, 1943). Samples of Socrates' relationships with modern philosophy are: SØREN KIERKEGAARD, *The Concept of Irony, with Constant Reference to Socrates* (1965, reissued 1968; originally published in Danish, 1841), reformulating his view of irony in terms of Romanticism and Hegel; LEONARD NELSON, *Socratic Method and Critical Philosophy* (1949, reprinted 1965), giving a Neo-Kantian interpretation of his maieutic method; and LASZLO VERSÉNYI, *Socratic Humanism* (1963, reprinted 1979), showing Socrates' implications for modern reflective and Existential humanisms; and WERNER J. DANNHAUSER, *Nietzsche's View of Socrates* (1974).

Philosophy: GREGORY VLASTOS (ed.), *The Philosophy of Socrates* (1971, reissued 1980), a collection of scholarly essays; JAMES BECKMAN, *The Religious Dimension of Socrates' Thought* (1979), and GERASIMOS F. SANTAS, *Socrates: Philosophy in Plato's Early Dialogues* (1979), detailed analyses of the *Apology* and *Crito.* See also RICHARD D. MCKIRAHAN, *Plato and Socrates: A Comprehensive Bibliography, 1958–1973* (1978).

Soil Organisms

The soil, often considered to be lifeless and inert, contains vast numbers of living organisms, which range in size from microscopic single cells to small mammals. One square metre of rich soil may contain more than 1,000,000,000 organisms.

This article deals with the importance of soil organisms, their general features, and the roles of microorganisms and animals in soil. It is divided into the following sections:

IMPORTANCE OF SOIL ORGANISMS

As decomposers. Soil organisms are important because they improve soil fertility by breaking down plant and animal tissues; during this process, the nutrients that are released and the minerals that are fixed are incorporated into the soil. Soil flora and fauna break down dead plant material into a finely divided organic complex called humus; the exact composition of humus is unknown, but it probably contains a small proportion of water-soluble organic substances, many phenolic compounds, some organic phosphates, and a polyuronoid substance containing complex sugars. It consists of about 60 percent carbon; the ratio of carbon to nitrogen is about 10 to 1 (*i.e.,* the nitrogen content is about 6 percent). Soil animals incorporate the humus into soil, where it is gradually broken down by microorganisms that release its constituents for plant nutrition.

Humus

As pests. Some soil organisms are pests of crops or animals. Growth of a single crop favours the proliferation of specific soil animals that feed on roots; these organisms then become pests and may kill the plants or cause decreased crop yields. Among the soil animals that feed on, or in, roots are nematodes, slugs and snails, symphylids, beetle larvae, fly larvae, caterpillars, and root aphids. Millipedes and springtails may occasionally attack growing roots, although normally they feed on decaying matter. Some soil organisms cause rots such as those in potato wart disease and potato blight (see DISEASE: *Diseases of plants*). Some soil microorganisms release substances that inhibit the growth of higher plants, and some soil animals are intermediate hosts for organisms that cause animal diseases, particularly those caused by nematode or helminth parasites.

Effects of pesticides

Pesticides and other soil pollutants affect the populations of many soil organisms. Some soil invertebrates are susceptible to pesticides. The active predatory species often are killed first, allowing the numbers of their prey to increase; for example, populations of springtails increase when their mite predators are killed by DDT. Other agricultural practices decrease the populations of soil organisms. Although the ecological balance of the soil is changed by pesticides, the total numbers or weight of soil organisms is not always decreased. Persistent pesticides affect the fauna over long periods of time. Herbicides, which may kill the soil microflora, often affect the soil fauna indirectly by killing some of their plant foods.

GENERAL FEATURES OF SOIL ORGANISMS

Types of soil organisms. *Protists.* Many of the microorganisms in the soil participate in breaking down organic matter, in soil formation and fertility, and in returning mineral elements to the soil. The smallest and most numerous soil microorganisms are bacteria. These spherical, rod-shaped, or spiral cells usually reproduce by simple division and survive unfavourable periods as resistant spores; they are more numerous in alkaline than in acid soils. The actinomycetes resemble both bacteria and fungi; actinomycete spores, although similar to those of bacteria, germinate into very fine colourless mycelia (or threads) that resemble those of fungi. Such fungi as molds, mildews, or mushrooms are usually larger and more variable in form than either bacteria or actinomycetes; they prefer acid soils. They also have resting spores, which germinate into actively growing, straight or branched, white or colourless mycelia that produce fruiting bodies on which spores develop. There are also algae, which are found as motile single cells or nonmotile filaments and contain pigments (*e.g.,* chlorophyll); they are most abundant near the soil surface (see Figure 1).

Microfauna. The soil microfauna, which are usually defined as animals less than 100 microns long (1 micron = 0.001 millimetre or about 0.00004 inch), include single-celled protozoans, some smaller nematodes, small flatworms, rotifers, and tardigrades (eight-legged arthropods); the microfauna feed on microorganisms. The abundant soil protozoans range from the shapeless amoeba to forms with whiplike flagella or tiny hairlike cilia; all protozoans have a nucleus, multiply by cell division, and form resting cysts. Much of the microfauna is confined to water films in litter and soil.

Mesofauna (Meiofauna). Other soil animals, somewhat larger than microfauna, have been placed in a heterogeneous group called the mesofauna. Many of them are nematodes, which are bisexual, survive unfavourable conditions as eggs in resistant cysts, and feed on microorganisms, other soil animals, decaying organic matter, or living plants. The mites (Acarina), which are the most numerous arthropods in soil, have eight legs when adult and feed on plant or animal material; some are important in breaking down organic matter. Springtails (Collembola), which are wingless insects commonly found in soil, feed on decaying organic matter and micro-organisms; collembolans that live near the soil surface have a springing organ, which enables them to leap. Small, wingless, insect-like organisms called proturans lack antennae and eyes and probably feed on fungi; members of another arthropod group, the pauropods, have nine pairs of legs when mature and feed on decaying plant material.

Macrofauna. The most common soil animals in the group known as the macrofauna are potworms (Enchytraeidae), small, white, segmented worms that feed on fungi, bacteria, and decaying plant material. The myriapods include the small white symphylids, which have 12 pairs of legs and feed on plant roots or decaying material; the centipedes (Chilopoda), which are many-legged, long, slender arthropods that prey on other soil animals; and the millipedes (Diplopoda), which have many legs arranged in pairs and feed principally on decaying plant material. Among the other macrofauna are slugs and snails, which eat mostly plant tissues and organic debris. Among the soil-inhabiting insects are fly larvae, and beetles and their larvae; some of these are predatory but most feed on decaying matter.

Megafauna. The megafauna are the largest soil organisms and include the larger earthworms, which pass both soil and organic matter through their guts, and several vertebrates (*e.g.,* moles, mice, hares, rabbits, badgers, gophers, snakes, lizards), whose food habits vary. The earthworm is important in maintaining soil fertility.

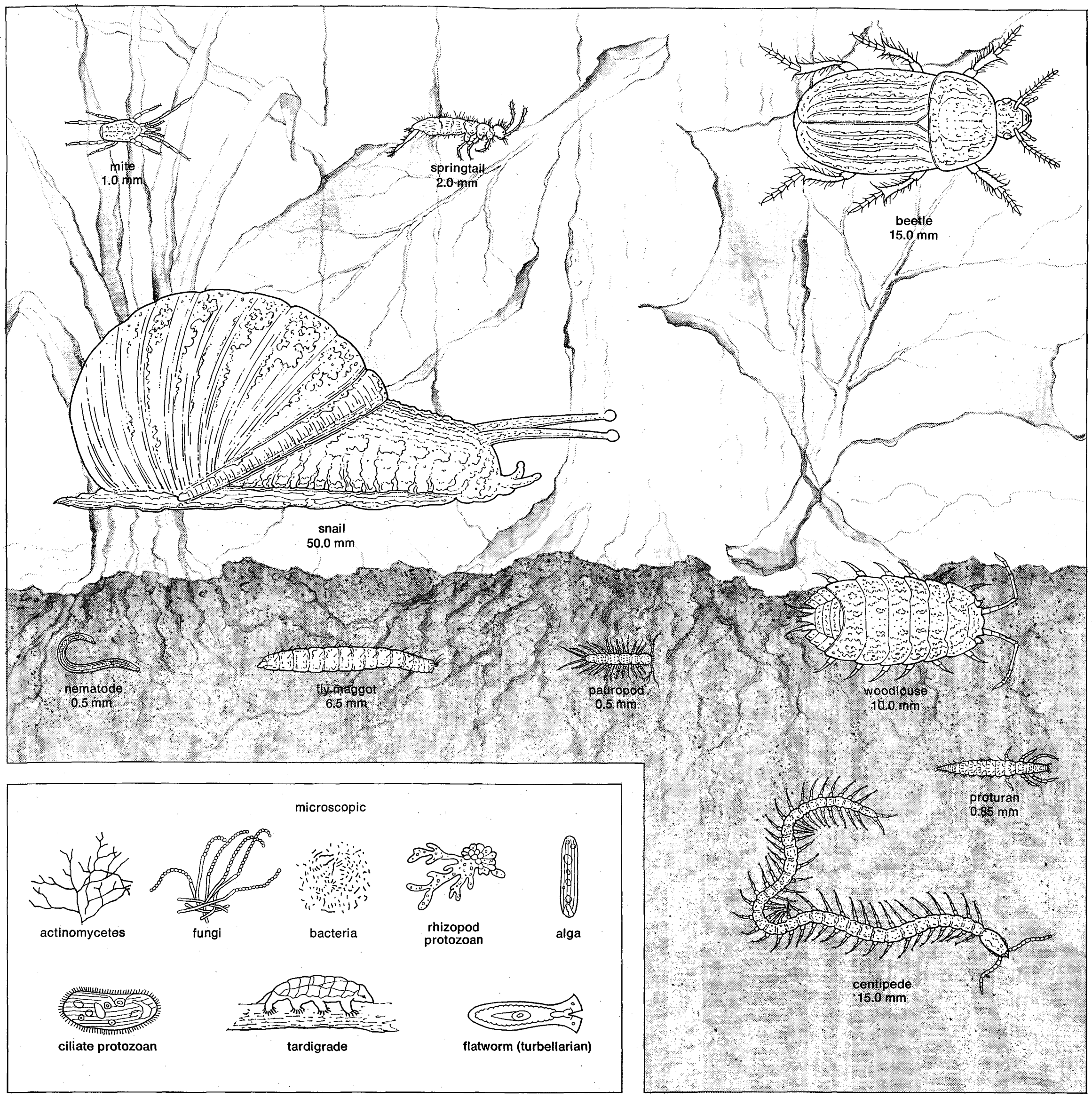


Figure 1: Types of soil organisms.

Vertical distribution of soil organisms. The vertical distribution of soil organisms is not uniform; most of them are concentrated in the litter layer and the top two inches of soil in woodlands. It has been calculated that as much as 90 percent of the soil fauna may be in this upper stratum. Many soil invertebrates move up and down through the soil either daily, seasonally, or in response to changes in temperature or moisture in the upper soil. Some invertebrates penetrate deep into the subsoil, usually moving through cracks and fissures. The vertical distribution of microorganisms follows that of organic matter; as a result, much of the soil microflora is concentrated close to the soil surface.

Movement through soil

THE ROLE OF MICROORGANISMS IN SOIL

The most important functions of soil microorganisms are to break down plant and animal organic matter and to help convert it into a form usable by plants. Heterotrophic soil microorganisms, which derive their carbon and energy from organic materials, are concerned mainly with the breakdown of organic matter, the carbon cycle, and nitrogen fixation. Autotrophic microorganisms, which obtain carbon from carbon dioxide and energy from the oxidation of simple organic compounds, form nitrites and nitrates and oxidize sulfur and iron compounds. Most microorganisms produce carbon dioxide, which dissolves in water to form a weak acid (carbonic acid); this carbonic acid dissolves relatively insoluble soil minerals, which thereby makes these minerals available to plants as nutrients.

Since most functions of soil microorganisms are beneficial, soils with large populations are usually naturally fertile. Microorganisms are more numerous and more active in woodland soils than they are in pasture soils; these in turn have a greater microbial activity than fallow or cultivated soils.

The carbon cycle. One of the most important functions of soil microorganisms is to recycle the carbon that is bound as complex substances in decaying plant and animal matter. Green plants convert carbon dioxide and water into organic plant substances. Animals eat these plants and convert their constituents into animal tissues. When plants and animals die, microorganisms begin the decay process and complete the carbon cycle by liberating carbon dioxide to the atmosphere. They also liberate minerals (for example, nitrogen and sulfur) and such complex molecules as proteins, polysaccharides, and nucleic acids into the soil.

The nitrogen cycle. Nitrogen, because it is an important component of proteins, is an indispensable element for the growth of microorganisms, plants, and animals. The availability of nitrogen in soil, in a form that is useful to plants, quite often influences natural vegetation and crop yields.

One essential role of soil microorganisms in the nitrogen cycle is to break down complex nitrogenous compounds (*e.g.*, proteins, polypeptides, nucleic acids) in soil and humus to simple inorganic compounds (ammonium ions, nitrite, and nitrate) that can be used by plants. Proteins can be decomposed by certain bacteria and fungi. Bacteria usually are most active in alkaline soils, whereas fungi are more active in acid ones. Microorganisms may release nitrogen in the form of ammonium ions (ammonification); these ions may be oxidized to nitrates by other microorganisms in a process that is known as nitrification.

Some microorganisms (*e.g., Nitrosomonas* species) reduce nitrates to nitrites; others (*e.g., Nitrobacter* species) convert nitrites to ammonium ions; still others convert nitrites or ammonium ions back to gaseous nitrogen (denitrification). Soils may be depleted of their nitrogen in this way if no organisms are present to convert the nitrogen back into compounds useful to plants.

Nitrogen fixation

Another essential role of soil microorganisms in the nitrogen cycle is to return (or fix) nitrogen from the atmosphere; this can occur either by the activity of free-living microorganisms or as a result of a symbiotic association between plants and microorganisms.

Free-living microorganisms capable of utilizing atmospheric nitrogen include bacteria (*e.g., Azotobacter, Clostridium*) and blue-green algae (*e.g., Anabaena*). Since many of these bacteria also can use ammonium ions and other nitrogenous compounds as nitrogen sources, they fix nitrogen only when alternative sources are not available. Their transformations are not as important as the symbiotic associations between plants and bacteria described below.

The genus *Rhizobium* causes the formation of root nodules on suitable leguminous plants as well as some trees and shrubs. The roots of the plants secrete substances that stimulate growth and multiplication of *Rhizobium* in the soil around the roots. During nodule formation, atmospheric nitrogen is fixed by these bacteria.

The sulfur cycle. Sulfur is an essential component of plant and animal tissues. Although abundant in nature, as much as three-fourths of soil sulfur may be bound in a form that is unavailable to plants. The most important forms of sulfur for plant nutrition are sulfate salts, although some sulfur in the atmosphere may be taken into plants through the leaves as sulfur dioxide gas. Soil microorganisms release sulfur from organic compounds and form inorganic compounds; hydrogen sulfide, for example, is released into the atmosphere. Some autotrophic microorganisms oxidize inorganic forms of sulfur into elemental sulfur, then to sulfurous acid, and, finally, to sulfuric acid.

Sulfates formed by the combination of sulfuric acid and calcium, magnesium, sodium, and potassium provide food for plants. Sulfates may be reduced to sulfites by specific bacteria in transformations that are similar to those that occur in the nitrogen cycle. In well-aerated soil, sulfates are formed readily; in oxygen-depleted soils, however, as are often common in swamps, hydrogen sulfide is produced.

(For additional information about biogeochemical cycles, see BIOSPHERE.)

Adverse effects of soil microorganisms. Although most soil microorganisms are beneficial, some are harmful. There are pathogens that cause root rots of higher plants; in greenhouses, for example, "soil sickness" may result from large populations of pathogenic organisms.

Microorganisms sometimes compete with higher plants for available nutrients, particularly nitrogen and phosphorus. The addition of carbon-rich organic matter to soil often causes significant increases in the numbers of microorganisms, which utilize the nitrates available in the soil, thus depriving higher plants of nitrogen.

Toxins

Another harmful effect of soil microorganisms is their production, particularly under oxygen-poor conditions, of substances that inhibit or prevent the growth of higher plants and other microorganisms. These toxic substances (*e.g.,* methane, hydrogen sulfide, phosphine, skatole, indole, various organic acids) are found most commonly in woodland soils and in lesser amounts in pastures and plowed and cultivated soils.

Their concentrations are highest in summer and autumn and lowest in the spring. Soil can be detoxified by washing or flooding, by adding lime, by using fertilizers, and by cultivating frequently.

THE ROLE OF ANIMALS IN SOIL

Animals are usually not important in maintaining the fertility of agricultural soils. In woodlands, however, soil animals are essential; they maintain the cyclic breakdown of organic material, return organic matter to the soil, and improve soil structure. Soil animals feed on decaying organic matter, living plants, animal excreta, or other animals. The number of animals in a given area of soil is dependent on the availability of food, temperature, moisture, and the degree of aeration and drainage; several hundred million animals (with a fresh weight of as much as 900 kilograms [one ton]) are probably a valid estimate for the number of animals in an acre of soil (see Figure 2).

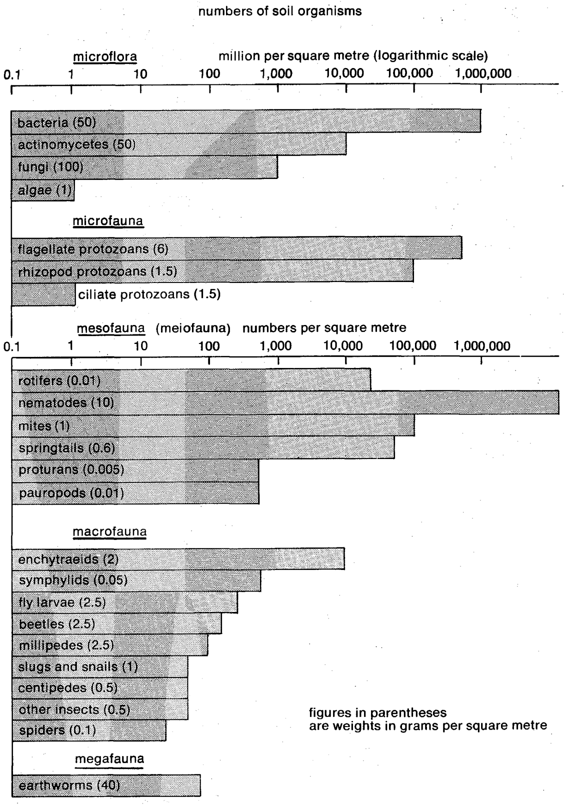


Figure 2: Characteristics of populations of soil organisms.

Formation of humus. Although they contain microorganisms when they fall, leaves usually are unpalatable to soil animals until they have weathered for several weeks in a litter layer. During this period, after water-soluble materials (*e.g.,* polyphenols) are leached out of the leaves, they are invaded by fungi and other microflora (called primary microorganisms) and become damp and soft in texture. These changes render the leaf tissues palatable to various soil animals.

A period of intense invertebrate activity occurs; the invertebrates that attack litter at this early stage have been called primary soil animals (see Figure 3). They include millipedes, wood lice, fly larvae, springtails, oribatid mites, enchytraeid worms, and earthworms. Earthworms are important; the burrowing species, such as the "night crawler" (*Lumbricus terrestris*), break up large quantities of leaves before pulling them down into their burrows as food. During this fragmentation of the litter, the invertebrates initiate the conversion of leaf material into a medium suitable for the growth of microflora; for example, invertebrates increase the surface area of leaf tissues by feeding on them. The organic constituents of leaf tissues do not undergo much change, however, as they pass through the invertebrate intestines.

Some chemical changes occur during fragmentation of the litter by soil animals. A few complex molecules are broken down into simpler molecules; for example, certain invertebrates, with the help of symbiotic organisms in their guts, can break down complex substances such as cellulose, keratin, and chitin. The leaf litter and the feces of the

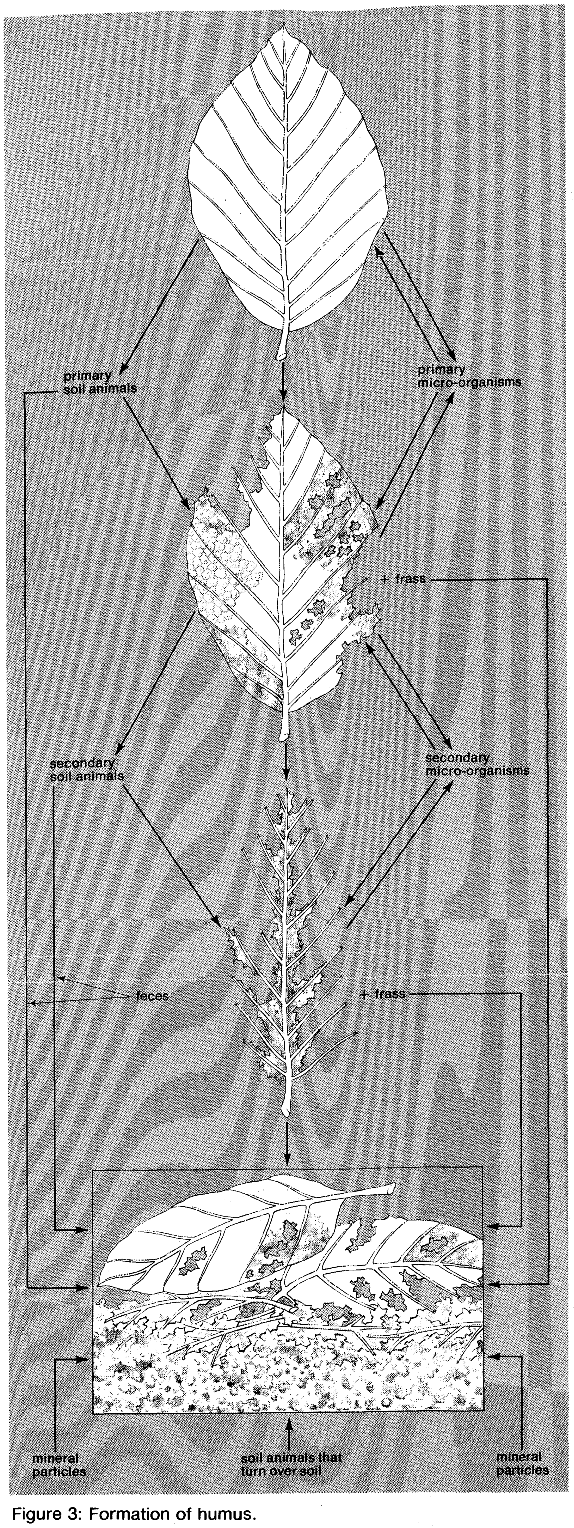


Figure 3: Formation of humus.

animals are now suitable substrates for the growth of other microorganisms, and a second wave of microbial activity occurs. These microorganisms act on the organic matter to make it palatable to other soil animals (the secondary decomposers), which flourish only on predigested material.

The secondary animals, which feed on this and also on microorganisms, include springtails, oribatid mites, and enchytraeid worms. Some secondary animals may also be primary decomposers, but different species usually are involved. The arthropods produce distinctive feces in the form of fine pellets, and enchytraeid worms produce amorphous black feces. During this stage algae, protozoans, nematodes, and rotifers invade the fine water films that form on the surfaces of leaf fragments and animal feces. (Many activities of the soil fauna and flora are dependent on moisture; if the litter layer dries, many soil organisms encyst or enter a resting stage.)

Complementary roles of organisms

The roles of the soil fauna and microflora in decomposing organic matter are complementary and interrelated. If soil animals are numerous, bacteria also are abundant. Consumption, digestion, and excretion of leaf litter by soil animals alternates with increases in the population of microorganisms; as a result the proportion of humic substances in the litter increases and mixes thoroughly with other materials.

The soil fauna ultimately destroy the mesophyllic tissues of the leaves, leaving only veins and toughened leaf margins. The rate at which destruction of the mesophyll continues is dependent on the moisture and temperature of the litter layer; destruction is rapid when the litter is moist and temperatures are high.

The organic matter of leaves, therefore, is continuously eaten and excreted by soil animals, acted upon by microorganisms, and mixed with other soil materials. As this process proceeds, the cellulose, protein, and sugar content of the humus diminishes; the ratio of carbon to nitrogen decreases as starches and sugars in the litter are broken down to carbon dioxide and water. Mineral elements such as sodium, potassium, and magnesium are leached away gradually; eventually the humic substances that remain may represent as little as one-fourth of the organic matter in the original leaf tissues. The remainder of the content of the humus at this stage is lignin (a substance related to cellulose) and protein; fats, waxes, and other residual materials also are present.

Humus increases the porosity and the water-holding capacity of soil and also acts as a reservoir for essential plant nutrients; these substances, released gradually, are made available to plants by the action of soil microorganisms.

Soil fertility and structure. After the formation of humus, soil animals play an important role in soil fertility and structure. The incorporation of humus from litter into soil is a gradual process dependent on the activities of soil animals. In soil that contains small animals and a few microorganisms, most of the leaves and wood fragments are not decomposed and lie on the soil surface; below the surface is a distinct humus layer, and soil horizons are defined; such a soil is said to be in a mor, or raw humus, condition.

When the soil flora and fauna are abundant, however, decomposition of plant material is rapid, and the humus is mixed into the soil so that layers or horizons are hard to define; such a soil is classified as a mull humus form.

Turnover of soil

The earthworms are important soil animals in temperate soils; they create the mull condition by turning over the soil—*i.e.,* taking organic matter deep into the soil and bringing subsoil to the surface as "casts" (*i.e.,* excrement). Estimates as to the amount of soil turned over in this way by earthworms vary from 7,000 to 54,000 kilograms per hectare (three to 24 tons per acre) per year, with an average of 10,000 kilograms (11 tons). This is equivalent to completely turning over all the soil on the earth to a depth of 2.5 centimetres (one inch) every 10 years.

Soil also is turned over by earthworm species that do not burrow or cast soil to the surface. Moles, which feed mainly on earthworms, also turn over large quantities of soil as they tunnel. In tropical soils, termites, as they construct mounds or gallery systems, are more important than earthworms in turning over soil. Many other animals distribute soil by tunnelling; in particular millipedes, wood lice, ants, and mollusks can turn over large amounts. An important aspect of soil turnover by animals is that it distributes microorganisms to suitable substrates.

Soil animals also aerate and drain the soil as they dig their tunnels or burrows. Earthworms, termites, ants, wood lice, millipedes, centipedes, and mites break open soil that has been compacted and improve its aeration as they move through it.

Soil without animals lacks "aggregates." Mineral soil particles (particularly clay and lime) and organic matter are cemented together as they pass through the intestines of soil animals; when these substances are excreted they are in the form of water-soluble aggregates; millipedes, wood lice, and earthworms are particularly valuable in forming such aggregates.

Soil animals contribute to the turnover of soil nitrogen; their feces contain nitrogen in forms such as ammonia, urea, uric acid, and other compounds that are readily converted to substances that plants can utilize. It has been calculated that the amount of nitrogen excreted by an average earthworm population is about equal to the amount removed by an average crop of hay—*i.e.,* about 56 kilograms of nitrogen per hectare (50 pounds per acre); however, soil microorganisms contribute much greater quantities of nitrogen. Organic matter bound in the tissues of soil animals is perhaps as much as 2,250 kilograms per hectare (one ton per acre), expressed as fresh weight. This is a reservoir of nutrients that is released gradually to the soil as the animals die.

BIBLIOGRAPHY. M. ALEXANDER, *Introduction to Soil Microbiology* (1961), a good introductory text that emphasizes soil transformations by microflora; A. BRAUNS, *Praktische Bodenbiologie* (1968), an up-to-date, well-illustrated account in German of practical soil biology and ecology; N.A. BURGES and F. RAW, *Soil Biology* (1967), a comprehensive text on the biology of soil invertebrates and microflora; J. DOEKSEN and J. VAN DER DRIFT (eds.), *Soil Organisms* (1963), a collection of original papers that emphasize the interrelationships between the soil fauna and microflora; C.A. EDWARDS *et al.,* "The Role of Soil Invertebrates in Turnover of Organic Matter and Nutrients," in *Studies in Ecology,* vol. 1, ed. by D.E. REICHLE (1970), a recent article emphasizing energy changes during litter breakdown; T.R.G. GRAY and D. PARKINSON (eds.), *The Ecology of Soil Bacteria: An International Symposium* (1967), a collection of important papers on soil bacteria; D.K.M. KEVAN, *Soil Animals* (1962), an introductory text; G. MULLER, *Bodenbiologie* (1965), good summaries in German of the various elemental cycles; E.J. RUSSELL, *The World of the Soil* (1957), a popular, readable account that emphasizes the importance of soil organisms. See also STEPHEN D. GARRETT, *Soil Fungi and Soil Fertility: An Introduction to Soil Mycology,* 2nd ed. (1981).

(C.A.E.)

Soils

Soil may be defined as the fine earth covering land surfaces that has the important function of serving as a substratum of plant, animal, and human life. Soil acts as a reservoir of nutrients and water, and absorbs and oxidizes the injurious waste substances that plant growth accumulates iu the rhizosphere (*i.e.,* the root zone). These functions of soil are possible because it contains clay minerals and organic substances (clay and humus form the finer part of soil) that absorb both ions (electrically charged atoms) and water.

This article treats soil formation and the influential factors that are involved, soil profiles, soil properties and their dependence on geographic factors, soil classification, and the geographical distribution of soil. Soil-plant relations are treated in the article ECOSYSTEMS. For coverage of the processes of rock weathering and the products that are involved, see GEOMORPHIC PROCESSES and MINERALS AND ROCKS: *Clay minerals.* The interaction of soils and climate is discussed in CLIMATE AND WEATHER. See also HYDROSPHERE for the role of soils in the world's water system.

This article is divided into the following sections:

GENERAL CONSIDERATIONS

Soil usually consists of a sequence of chemically and biologically differentiated layers, called horizons, that have been formed by the action of natural forces on the unconsolidated residue (regolith) of rocks and minerals on the Earth's surface. The regolith itself is the result of weathering of the original massive rocks at the surface and may be considered to be incipient soil. In the process of weathering, fracturing occurs along planes of weakness in the rocks and their mineral components react chemically with water, oxygen, and organic, carbonic, nitric, and sulfuric acids that are derived from the atmosphere and living organisms. Through the continuation of these chemical, physical, and biological actions and reactions, the regolith material is transformed into soil that exhibits one to four master horizons called, from the surface downward, A, B, and C horizons (the soil profile), and the underlying consolidated rock, R.

Regolith and the four horizons

Surface and subsurface horizons are distinguished by the differences in conditions at various depths. Some materials accumulate in the soil and some substances decompose or are dissociated to release compounds that may be dispersed, dissolved, or transported from one horizon to another; they may even be carried away from the immediate locality. Solutions are present in the soil layers, and these vary in composition and concentrations, and thus in the chemical changes they effect. These several conditions determine the abundance and variety of living organisms, from microscopic ones to vertebrates, which in turn influence the further development and the properties of the soil. In the soil profile that results, the horizons differ from each other in characteristics such as colour, chemical composition, particle sizes and distribution, and structure (*i.e.,* arrangement of the particles in groups, aggregates, or independent units).

The A horizon extends from the ground surface to the unmodified regolith or consolidated rock. It is the horizon in which weathering is most intense, with partial removal of the resulting products and a zone of accumulation of organic material. In a typical high-productive soil, the A horizon is rich in humus (more than 1 percent carbon) to a depth of 15 centimetres (6 inches) or more; it contains the greater part of immediately available plant nutrients. Soils of low productive capacity tend to have A horizons that are poor in humus and plant nutrients. The B horizon lies immediately beneath the A horizon and may reach a depth of 65 to 90 centimetres (26 to 35 inches). It is a zone of more moderate weathering in which there is an accumulation of many of the products removed from the A horizon. Sometimes the B horizon is very clayey and impermeable, constituting a serious impediment to plant growth. In a productive soil water, air, and root penetration is easy to a depth of 75 centimetres (30 inches) or more and the water-holding capacity of the entire layer is 15 centimetres (6 inches) of available water, or more. The C horizon contains the parent materials, from which the A and B horizons are formed. The R layer is not part of the soil proper; it is an underlying layer that has properties much different from those of the C horizon immediately above it and that influences the soil by its position. Because of soil development factors that will be discussed in following sections of this article, some master horizons can be missing in certain soil profiles. For example, erosion may remove the A horizon; shortness of time or absence of key factors may preclude development of a B horizon; a shallow, weathered layer may be transformed completely into A and B horizons, so that no parent materials are left to constitute a C horizon; or conversely a deep regolith can screen the A layer to the point that the latter does not influence the properties of the overlying soil horizon.

There are many functions provided by soil that are important to man. Agriculture and animal husbandry produce more than 90 percent of man's food; together with forestry these soil-connected activities produce many other materials needed by man, such as wood, cellulose, textile fibres, and leather. The utilization of soil by agriculture, animal husbandry, and forestry is often termed "soil exploitation." Soil is necessary for dwellings, highways, airports, and recreation areas, and it also provides road fill and material for water retention structures and many other essential functions.

PROCESSES OF SOIL FORMATION

The weathering process

Clay formation. The most fundamental soil forming process is weathering. During the weathering process the crystals of feldspars and other silicate minerals are broken up, releasing chemical compounds such as bases, silica, and oxides of iron and aluminum (specifically the sesquioxides: Fe_2O_3 and Al_2O_3). Leaching (*i.e.*, washing or draining away by percolation of water through the soil) removes most of the bases and some of the silica, and the remaining silica combines with alumina and forms crystalline clays.

The kind of crystalline clay produced depends on leaching intensity. Rapid leaching leaves little silica to combine with alumina and results in what are known as 1:1 clays consisting of a tetrahedral (silica) and octahedral (alumina) layer; slow leaching leads to the formation of 2:1 clays, consisting of one octahedral (alumina) layer sandwiched between two tetrahedral (silica) layers. In neither case is the result solely one of the two types, but 1:1 clay is predominant after rapid leaching, and 2:1 clay more abundant when leaching is slow (see Figure 1).

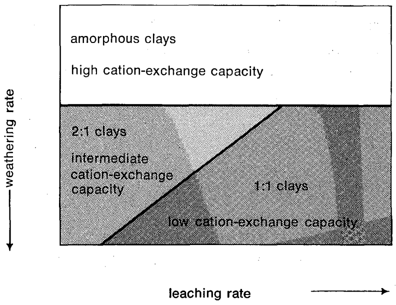

amorphous clays
high cation-exchange capacity
2:1 clays
intermediate cation-exchange capacity
1:1 clays
low cation-exchange capacity
weathering rate
leaching rate

Figure 1: Influence of weathering and leaching rates on clay mineralogy and cation-exchange capacity; rates increase in the direction shown by the arrows (see text).

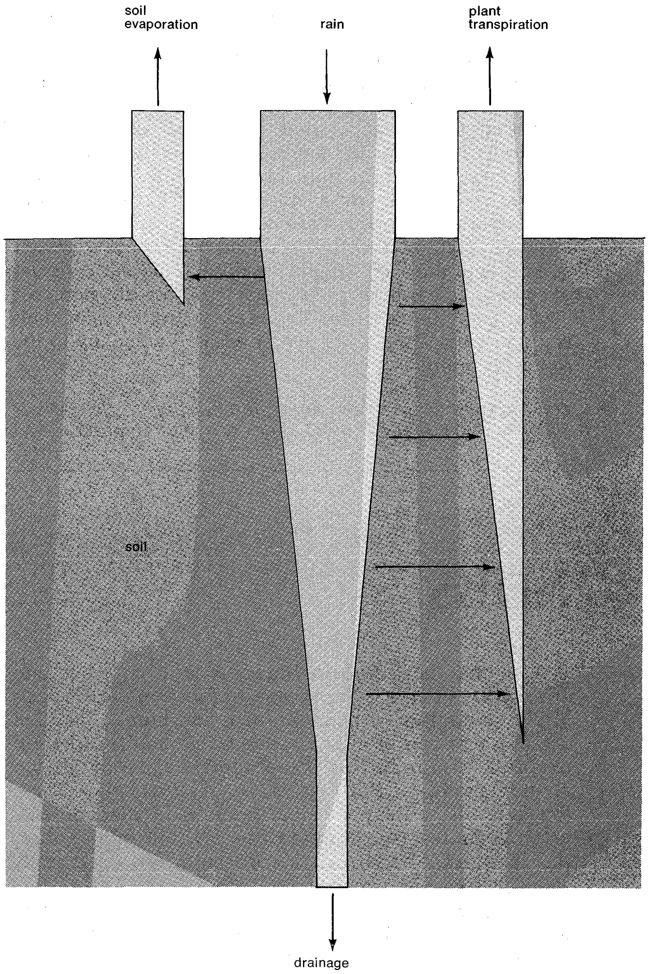


Figure 2: *Water movement in the soil.*
In general, water moves in the soil in only one direction—it descends. While descending it is absorbed by plant roots, and the substances it contains precipitate; in this way various substances are transported from the higher layers (horizons) to the lower ones. Under moderate rainfall, the quantity of water that is lost by drainage is little, if any, and only very soluble substances are eliminated.

Ions are atoms or groups of atoms that have become electrically charged through the loss or gain of electrons; ions of positive charge are cations and those of negative charge, anions. The phenomenon known as cation exchange is the result of the differences in the force with which clay holds different cations. The surfaces of clay minerals contain many negative charges, which are balanced electrically by the adsorption (*i.e.*, taking up on the surface, as opposed to taking up and distributing throughout the body of an absorbent) of cations. The tightness with which the soil holds cations varies; it is greatest for hydrogen cations and decreases for other cations in the following order: calcium, magnesium, potassium, sodium. Because of the differences in attractive forces, adsorbed cations may be removed and replaced by other cations. Because clay formation is very slow, for long time periods the silica and alumina constitute an amorphous (structureless) clay that has a cation-exchange capacity several times greater than that of crystalline clays. Alumina-rich 1:1 clays have low cation-exchange capacity. Silica-rich 2:1 clays have higher cation-exchange capacity, but lower than that of amorphous clays. Thus young soils and lower horizons that have undergone little weathering have high cation-exchange capacities per unit of clay, as do volcanic soils, because their weathering has been so rapid that clay formation has not had time to take place. Such capacity is lower in old soils in dry or cold climates, and it is lowest in tropical soils.

The most significant aspect of weathering is its relationship to clay crystallization or leaching, rather than its rapidity. Minerals differ greatly in their weatherability, and they have been classified in this respect. Clay crystallization seems to be accelerated by a basic medium, especially magnesium. Leaching is an irreversible process (see Figure 2), and thus the effect of a humid season cannot be undone by a dry season. What is important is leaching rainfall, Ln., which is the difference between rainfall and potential evapotranspiration (*i.e.*, return to the air by direct evaporation or by transpiration of vegetation, or both) during the humid season. Leaching rainfall is low in temperate countries, even when they are humid (as in London and Moscow), and high in tropical countries, even when they have a long dry season (as in Kano, Nigeria). The prevalence of 1:1 clays in tropical countries is more directly the result of leaching rainfall (see Table 1) than of the great age of the rocks.

The character of clay

Rocks subjected to weathering, young soils, and soils formed from easily weatherable materials (*e.g.*, volcanic ashes, lava, or basalt) are rich in amorphous clays of high cation-exchange capacity. Such amorphous clays are mixed with 1:1 clays or 2:1 clays, depending on leaching intensity. Old soils and soils formed from materials that weather with difficulty are poor in amorphous clays and rich in 1:1 clays or 2:1 clays according to whether leaching is rapid or slow. Soil usually contains a great variety of clays that differ in their relative proportions.

In addition to the absorbing capacity for water and ions, which is the most important soil property, clay mineralogy affects such physical properties of soil as its structure and permeability to air and water. A soil rich in amorphous clays has a low apparent density and is very permeable to air and water; decay of organic matter is slow and humus content is unusually high. Soils rich in 1:1 clays usually have a stable structure and swell very little when moistened; on the other hand, a soil rich in 2:1 clays, more especially montmorillonite, has a weak structure, low permeability, and swells easily when moistened.

Tropical soils are usually dominated by 1:1 clays, whereas those of temperate countries contain chiefly 2:1 clays. But pedology, the study of soils, began in temperate countries and is consequently based on what is observed in soils with 2:1 clays; it was some time before the fundamental difference between tropical and temperate soils was realized and understood. For analogous reasons, and because amorphous clays were lost when preparing clay samples for analysis, understanding of volcanic soils has been much delayed. It is only recently that they have been recognized as a special group. Even now the volcanic soils of dry climates are little known.

Mica is a platy silicate mineral that is directly altered (*i.e.*, changed in mineralogical composition) to clays. The first clays formed have high cation-exchange capacity; later 2:1 or 1:1 clays dominate, so that the foregoing outline of the weathering process is applicable to soils formed from materials rich in mica.

Experiments show that weathering is more rapid in a medium containing organic acids. Hence, when soil is covered by raw humus, as in podsolic regions, amorphous clays are formed as a result of the rapid weathering.

By eluviation (the transposition of soil material from one horizon to another) an illuvial horizon is formed by the deposition of materials leached out of an overlaying layer. The illuvial horizon of podsols, called the spodic horizon, is rich in amorphous clays and resembles ando (volcanic ash) soil in many respects.

Clay mineralogy is usually studied by X-rays and techniques of differential thermal analysis, which provide information on structural changes during heating. Soils may be classified on the basis of the relationship between cation-exchange capacity and clay content. Gradations between ando soils and those rich in the clay mineral kaolinite have cation-exchange capacity similar to those of soils dominated by 2:1 clays, but when conditions favour the formation of 2:1 clays, or the soil is young, volcanic soils exhibit very high cation-exchange capacities.

Humification. Second only to clay in importance as a soil constituent is organic matter, which contributes significantly to soil absorbing capacity. Roots die continuously, vegetation and crop residues fall on the soil surface and decay, and the organic leaching products enter the soil. Part of these residues is mixed with soil by organisms living in the soil or by tillage operations; some algae produce organic matter that also is added to soil.

Types of humus

All these residues are transformed by microorganisms into a mixture of organic substances called humus. Two kinds of humus may be distinguished: (1) mild humus is dark in colour, well saturated with bases, especially calcium, rich in humic acids (of high molecular weight), and serves to stabilize clay; (2) raw humus is more red in colour, less basic, rich in fulvic acids (of low molecular weight), and favours dispersion of clay. Soil also contains organic matter that has not yet been humified and in certain cases peat, which is more or less carbonized material that still retains its original structure.

Table 1: Leaching Rainfall
(in millimetres)

	Ln
Bahía Blanca, Argentina	0
Damascus	10
Tobolsk, R.S.F.S.R.	10
Reno	10
Eureka, Northwest Territories, Canada	20
Tehrān	20
Irkutsk, R.S.F.S.R.	20
Barrow, Alaska	30
Saskatoon, Saskatchewan, Canada	40
Baker Lake, Northwest Territories, Canada	40
Prague, Czechoslovakia	40
Harbin, China	40
Salt Lake City	40
Anadyr, R.S.F.S.R.	40
Rosario, Argentina	50
Nizhnekolymsk, R.S.F.S.R.	50
Ankara	50
Omsk, R.S.F.S.R.	50
Resolute, Northwest Territories, Canada	60
Madrid	60
Lincoln, Nebraska	60
Santiago, Chile	70
Corrientes, Argentina	80
Athens	80
Tripoli, Libya	80
Tientsin	90
Rabat, Morocco	90
Niamey, Niger	100
Bucharest	100
Kiev	110
Des Moines, Iowa	120
Paris	130
Warsaw	130
Budapest	140
London	160
Berlin	170
Moscow	220
Jerusalem	230
Cienfuegos, Cuba	240
Stockholm	240
Oslo	260
Mexico City	270
Chicago	270
Rio de Janeiro	280
Amsterdam	310
Naples	320
Kano, Nigeria	362
Kiel, West Germany	410
New Orleans	460
New York	470
Nāgpur, India	517
Bamako, Mali	520
Kinshasa, Zaire (Congo, Dem. Rep. of the)	640
Yaoundé, Cameroon	690
Djakarta	690
Wellington, New Zealand	690
Canton	820
Tokyo	870
New Delhi	1,040
Valentia, Ireland	1,150
Bombay	1,679
Rangoon	2,000
Valdivia, Chile	2,120
Cochin, India	2,462
Conakry, Guinea	3,740
Monrovia, Liberia	3,860
Cherrapunji, India	10,922

Vegetation determines the kind of humus a soil contains. Grasses produce mild humus rich in humic acids, and conifers produce raw humus rich in fulvic acids. But the base content of soils is also important in determining humus type; for example, calcium-rich soils produce milder humus than do hydrogen-saturated soils. Waterlogging also has an effect, favouring production of raw humus. Root growth is more abundant near the soil surface, and aerial residues accumulate on the surface; thus, organic-matter content decreases with depth.

The effects of leaching. Except for the amount that is immediately evaporated from the surface, rainwater moves through the soil in only the downward direction, and while descending it is absorbed by plant roots. Excess water may be eliminated by drainage, so that even in dry climates the upper layers of soil are somewhat leached. Leaching decreases with depth (see Figure 2). Descending water dissolves various substances, which are precipitated when the water is absorbed by roots. Descending water transports substances from the higher layers to the lower ones, and when some of the water is lost by drainage, the substances it carries are lost as well. More soluble substances are more easily leached away from a layer and are accumulated at greater depth. All chlorides and sodium, potassium, and magnesium sulfates are more soluble than gypsum, which is in turn more soluble than calcium and magnesium bicarbonates.

When there is no drainage, all substances remain in the profile, stratified according to their solubility (see Figure 3)—chlorides and sulfates at the lowest part of the rhizo sphere (root zone of plants), gypsum above them, and calcium carbonate at a higher level. With more leaching the profile deepens, salts are leached away, gypsum accumulates in the lowest part of the rhizosphere, and lime above it. Still further leaching removes first the gypsum, while the lime accumulates in the lower part of the rhizosphere; then it removes the lime, and the soil becomes

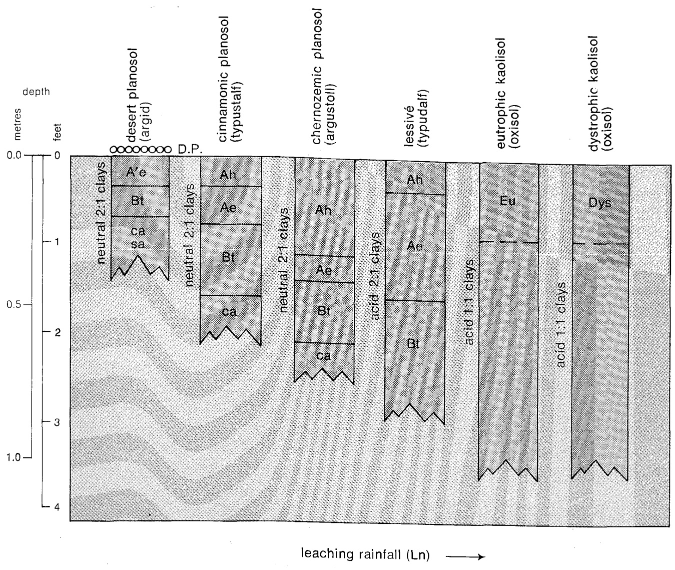


Figure 3: *Influence of leaching rainfall (Ln) on soil formation.* Leaching rainfall increases from left to right, as shown by the arrow. D.P., desert pavement; A'e, surface horizon poor in humus and eluvial; Bt, textural B; ca, lime accumulation; sa, salts accumulation; Ah, surface horizon rich in humus; Ae, eluvial surface horizon; Eu, eutrophic upper soil; Dys, dystrophic upper soil (see text).

acidified. Because silica is less soluble than lime, 1:1 clays are usually formed in an acid medium, and are usually acid. But an acid soil may later be basified, and 1:1 clays are sometimes inherited from parent material, as in the case of allochthonous soils (*i.e.,* those transported from other environments).

Base saturation acidity

Dispersion mechanisms of soil constituents. *Biological transport of elements.* Plants absorb various elements from the profile and deposit them, with their residues, on the soil surface, a process that counteracts leaching. In soils rich in bases leaching is more important. This accounts for the fact that base saturation and pH (the standard measure of acidity, related to the concentration of hydrogen ions; a pH value of 7.0 is neutral, whereas that from 1.0 to 7.0 is acid and that from 7.0 to 14.0 is alkaline) usually increase with depth. In acid soils, leaching is difficult because cations are firmly retained by soil colloids, and biological transport to the surface prevails; base saturation decreases with depth. In the kaolisols of the humid tropics, the upper 25 centimetres (10 inches) of soil often contain more bases than the rest of the profile to a depth of 150 centimetres (59 inches). The process is selective, however; *e.g.,* sodium and magnesium are less absorbed by plants than are potassium and calcium.

Addition of constituents by running water and wind. Floodwaters always contain salt, and soil, when flooded, may become saline (*i.e.,* high in chlorides and sulfates of sodium, potassium, calcium, and magnesium) unless some of the water is lost by drainage. Floods sometimes deposit fine earth or coarse materials on the soil surface. Sea winds often bring salts, and some soils, such as those of New Zealand and the Falkland Islands (Islas Malvinas), have probably been salinized in this way. In other cases winds bring dust or volcanic ash, additions that may counteract leaching and modify the soil profile.

Subirrigation. In some soils there is a water table one to two metres below the rhizosphere. During the dry season water rises by capillary action from this water table to the rhizosphere, where it is absorbed by plant roots, and the substances it contains accumulate. In this way a horizon of salts, gypsum, lime, silica, or iron accumulation is formed, depending upon the substances that are contained in subirrigation water. When the accumulation takes place in a layer poor in clay, the lime, silica, or iron are deposited on the surface of sand or gravel and may cement them, forming a hardpan (cemented layer impenetrable by roots). Many calcareous horizons, duripans, fragipans, and iron pans have this origin, but they also may be cemented by substances that descend in the profile. When the water table is shallow, the hardpan may reach the soil surface.

Solonization. Sodium clay is easily dispersible, and in sodium soils it is transported from the surface to some depth. Sodium soils swell when moistened, however, which impedes clay movement, so that clay illuviation takes place near the surface rather than at depth. This process is called solonization, or solonetization, and results in sodium-saturated soils (solonetz) that have a richer clay horizon at depth than at the surface.

Transport of chemicals in soils

Leaching activated by vegetation, however, produces organic acids and transports calcium from the lower layers to the surface. Sodium then is replaced by calcium or hydrogen in the eluvial layer and the upper part of the illuvial one. Clay illuviation thus takes place at a greater depth, and soil becomes planosol or solod (calcium- or hydrogen-saturated soils with considerable difference in clay content between the eluviated surface and the clay illuvial horizon).

Planosols can also be formed directly from a slightly sodium-saturated soil; low sodium saturation is sufficient to facilitate clay dispersion and permits clay movement to greater depth. Solonetz are often formed from sodium-saline soils, and planosols from easily weatherable materials, such as volcanic ash, under conditions that impede leaching (*e.g.,* dry climate or poor drainage or both).

Iron and clay illuviation. Fulvic acids react with iron and may cause its eluviation and that of clay. For that reason, soils formed under forest vegetation, which produces humus rich in fulvic acids, are often leached; there is a difference in clay content between the surface horizon and the B horizon in which the clay is illuviated. Moreover, such soils are relatively rich in free iron, and a difference in free iron content is often observed between the surface horizon, which is poorer, and a lower enriched horizon. Because dispersed iron is easily precipitated by bases or concentration of soil solution, soils developed under conditions that favour the formation of fulvic acid-rich humus are usually rich in iron that is finely precipitated with organic matter, which gives them a brown colour (brunisolic soils).

Podsolization. Weathering in a strongly acid medium produces an eluviation of the sesquioxides (Fe_2O_3 and Al_2O_3) released by weathering. Because weathering is very rapid amorphous clays of high cation-exchange capacity are formed. The soil formed has an upper horizon rich in unweathered minerals (sand), and an illuvial one rich in amorphous clays ("spodic" horizon). Such soils are called podsols. Although vegetation and other conditions may favour the accumulation of raw humus, its acidity is neutralized when soil contains lime or when weathering releases bases in sufficient quantity. Podsolic weathering is therefore observed in sandy materials rich in quartz and other minerals of low weatherability. The organic matter that produces eluviation of clay or sesquioxides decays and cannot accumulate in the illuvial horizon; but when this horizon is waterlogged or when its temperature is very low, an accumulation of organic matter takes place and a humus illuvial horizon is formed. Such soils are called humus podsols.

Rubification. When soil is thoroughly dried from time to time, precipitates of iron and organic matter cannot accumulate and the organic matter disappears by decay, causing irreversibly dehydrated iron sesquioxides to form. This process is known as rubification, and the resulting soils, called cinnamonic, are rich in fine crystallites of dehydrated iron sesquioxides.

Formation of cemented pans

Ferrugination. Iron sesquioxides adhere firmly to sand grains and gravel, give them a red colour, and may cement them to form an iron pan. This process is most common in soils that have been formed from materials that release much iron upon weathering and leave much sand, whereas humic horizons are seldom ferruginized, probably because organic acids and reductive processes remove the iron coating from sand grains. Fulvic acids favour iron eluviation and the formation of ferruginized horizons at some depth; thus, fragipans and other iron-cemented pans are common in podsols and brunisolic soils. The ferruginous horizons of tropical soils are formed similarly, fulvic acids contributing to their formation, but reductive processes can also be responsible for iron eluviation. Periodic soil drying may consolidate iron coatings, especially in climates in which a humid and a dry season alternate. It seems that iron coatings protect sand and gravel grains from weathering, and they are the cause of many "stone lines" parallel to soil surface.

Segregation of iron. In soil that is rich in iron and poor in 2:1 clays that retain it, iron tends to segregate, forming irregular concretions and nodules. After repeated moistenings and dryings, the nodules become as hard as stone and may form a pan cemented with iron and gravel (laterite). Waterlogging seems to further iron segregation, and alternative moistenings and dryings harden the nodules or pan. This process resembles ferrugination, the difference being in degree of segregation.

Gleization. Under waterlogging conditions, iron becomes ferrous and has a valence of 2+. (Valence is an expression of the combining power of an element, which is determined by the number of electrons in its outer or valence shell of electrons; valence is represented by a number that corresponds to the number of electrons an atom of the element can accept [if the number is positive] or donate [if the number is negative]). Ferrous soil has gray-blue colours, is called gley, and the process is called gleization. When waterlogging is less severe and transitory, ferrous iron is oxidized (*i.e.,* loses an electron, with a resulting change in valence to 3+, which is that of ferric iron, F^{3+}), and ruggy mottles (spots) are produced in the soil. Because ferrous iron is more mobile than ferric iron, gleization favours iron eluviation, resulting in a leached upper horizon; iron accumulates at a lower depth or is leached away. Or the iron may ascend by capillarity from the water table to the lower part of the rhizosphere.

Development of soil characteristics. *Soil profile and horizons.* Humification, leaching, clay eluviation, and other soil-forming processes all create differences between soil horizons. Soil profiles vary considerably in horizon types and their thicknesses.

Soil properties. Both water-holding capacity and cation exchange depend on clay content, clay mineralogy, humus content, kind of humus, and nature of absorbed ions. Mineral constituents larger than clay also have some water-holding capacity, and in some cases silt exhibits considerable ion exchange capacity. This may be attributed to its formation by aggregates of amorphous clays or by amorphous clay that adheres to silt particles.

Nature of soil texture and structure

Soil texture (*i.e.,* the relative proportion of mineral constituents classified according to their size) determines to a considerable extent soil porosity and, consequently, permeability to air and water, which are essential for root life and many processes that take place in soil. A high content of coarse constituents usually increases porosity. But clay and humus bind single soil grains together in increasingly larger aggregates.

Soil has a "structure" on which its porosity-permeability depends greatly. Soil structure is built up by alternate moistening and drying, and plant roots contribute greatly by opening pores between soil aggregates; it is undone by waterlogging. The stability of aggregates increases with humus content, especially humus that originates from grass vegetation (forest humus is less effective).

Calcium and hydrogen soils have structural stability, whereas sodium soils lose stability easily. Soils rich in amorphous clays are very porous and permeable, and their apparent density is exceptionally low. Soils rich in 1:1 clays are also permeable. High content of 2:1 clays, however, is usually associated with low permeability because the soil swells when moistened and the pores close. The clay type is very important; montmorillonite is the worst in this respect. Water movement in 2:1 clayey soils is chiefly through cracks produced when soil dries, and water movement in pores between small aggregates is very slow. Soils rich in coarse constituents are usually permeable. Soils with a high fine-sand content, however, have low permeability, poor aggregation, and small pores between single grains of fine sand; moreover, they are not extensible, and they resist root penetration.

Texture variations in horizons

Because soil horizons vary in texture, humus content, and absorbed ions, soil properties vary from horizon to horizon. This fact should be taken into consideration in evaluating the soil as a whole. Distinction must be made between intrinsic qualities—such as texture, clay mineralogy, the nature of absorbed cations, and their variation along the profile—and the secondary qualities that are consequences of the first and are more or less transitory—such as structure and permeability. Humus content and soluble salts are intermediary; they can be modified by management, added, or leached away.

From an agricultural point of view, the most important soil qualities are cation-exchange and water-holding capacities, the nature of absorbed cations, richness in assimilable plant nutrients and certain other substances (*e.g.,* lime, free aluminum), permeability, structure, humus content, and waterlogging. For engineering uses, texture, clay mineralogy, humus content, swelling capacity, capacity to corrode metals, structure (which can be modified), permeability, and waterlogging (chiefly extrinsic), are important.

Horizon nomenclature. Usually the upper horizon, which is generally richer in humus and may be eluvial, is called A; it may be "humic" (Ah), eluvial (Ae), or "bleached" (A_2). The horizon in which clay has been illuviated is called Bt; that of amorphous clays produced by podsolization ("spodic" horizon) Bir (from iron, which was considered the most important illuviated substance); and that of humus illuviation Bh. Weathered material that has not been sufficiently enriched in humus to be humic and has not had important illuviation of clay or humus is called C. Underlying consolidated rock is called an R horizon.

A horizon of lime accumulation is usually called ca, but the term can be extended to all horizons that give effervescence with an acid. An accumulation of gypsum is indicated as cs; a horizon unusually rich in easily soluble salts by sa; cn designates an accumulation of ferruginous concretions and nodules; mf a ferruginous hardpan; mc a calcic hardpan; x a fragipan (fragil hardpan cemented with iron); G denotes strong gleying; g moderate gleying.

According to clay mineralogy a horizon may be allophanic, illitic, kaolinitic, or superkaolinitic, the diagnosis being chiefly based on the relation between cation-exchange capability (CEC) and clay content. According to base saturation, a horizon may be acid, neutral, natric, magnesic, or calcic. Fifteen percent sodium saturation is sufficient to make a horizon "natric." The terms acid and neutral are replaced by dystrophic and eutrophic (poor or rich in nutrients) in the case of kaolinitic and organic soils in which the relation of absorbed cations to clay content is a better diagnostic indicator than base saturation. Many of these terms are poorly defined in the literature, and definitions vary from author to author.

Mineral composition and rock weathering

Rates of soil formation. Rates of soil formation vary enormously according to the process involved, and for the same process according to conditions. Weathering varies according to the weatherability of the material, and minerals may be classified in this respect. Weathering of a consolidated rock is usually slow at the beginning, accelerates as the rock breaks up and more surface area is exposed to attack, and then becomes increasingly slower as weatherable minerals disappear. It is accelerated by plant roots and by the accumulation of raw humus, and also by a rapid removal of the bases by leaching.

Weathering is more rapid than is usually thought. In experimental leaching of coarse gravel of basalt with pure water at 20° C (68° F), 0.45 percent of the silica was eliminated in 11 months. Leaching with water containing acetic acid (pH 2.5, which is a common acidity value in the vicinity of roots and decaying organic matter) accelerated the process by a factor of 21. Volcanic ashes weather sufficiently to support vigorous vegetation in a few decades.

Tombstones and other man-made structures weather more slowly because they are not in contact with plant roots or decaying organic matter, and, moreover, the surface of attack is very small.

Humification is a very rapid process. Experiments and agricultural practice all over the world have shown that soil organic matter can be doubled or reduced to half in

5–50 years. Podsolization is also rapid, and it has been reported that soils planted with conifers have been podsolized in a century. Salinization and desalination are also rapid because of the higher solubility of salts less soluble than gypsum. Decalcification and acidification with water containing carbon dioxide are intrinsically slower but may be accelerated when produced by water containing organic acids, which is usually the case.

There is less information concerning clay eluviation, but experiments show that this is a relatively rapid process because water often moves through soil cracks, carrying much clay per unit of water. The formation of argillic (clay illuvial) horizons has been estimated to require between 550 and 5,250 years. In many cases the age of soil that has been formed in previously glaciated areas, newly formed terraces, or from recent volcanic material, is known; but it is not always possible to ascertain whether clay eluviation required all this time. In any case, existing evidence and the fact that soils correspond well to present climatic conditions indicate that soil formation processes are relatively rapid.

BASIC FACTORS INVOLVED IN SOIL FORMATION

Climate. The amount and kind of clay that is formed, soil depth, base saturation, and the formation and depth of saline, gypsic, or calcic horizons all depend on leaching intensity. Consequently, the significant factor is leaching rainfall (see Figure 3). The soils of temperate climates are rich in 2:1 clays, whereas those of tropical countries are rich in 1:1 clays. But in both cases when the soil is young, or has been formed from volcanic ashes, it contains considerable amorphous clays, which increase its cation-exchange capacity.

Effects of Ln on weathering depth

Temperature is another influence in soil formation. Because the solubility of silica increases with temperature, high temperatures favour the formation of 1:1 clays, which is another reason for the abundance of such clays in tropical soils. Water penetration and root penetration depend on leaching rainfall, so that the depth of weathering and of soil increase with Ln. Soils of tropical countries are usually very deep; those of cold or dry climates, or of both, shallower. Temperature is also important. At high latitudes and altitudes the lower layers of soil are very cold, even in summer; consequently, weathering and soil depth are limited, which is important in the case of autochthonous soils (*i.e.,* formed in the original location) developed from consolidated rocks. Permafrost limits both soil depth and drainage. In very dry climates no water is lost by drainage. In dry climates substances formed by weathering accumulate in the profile, stratified according to their solubility, and as leaching increases the lower horizons disappear, and the profile becomes deeper. Finally, there are insufficient bases in the soil to neutralize the organic acids produced by decay of roots and organic matter. Solution in the soil then tends toward acidity, the soil is leached with acid water, and hydrogen ions replace other ions. The result of this cycle is an increasingly more acid soil. Acidification, however, also depends greatly on vegetation type, temperature, and waterlogging. In addition to the above direct influences, climate greatly affects vegetation, which is one of the most important factors in soil formation.

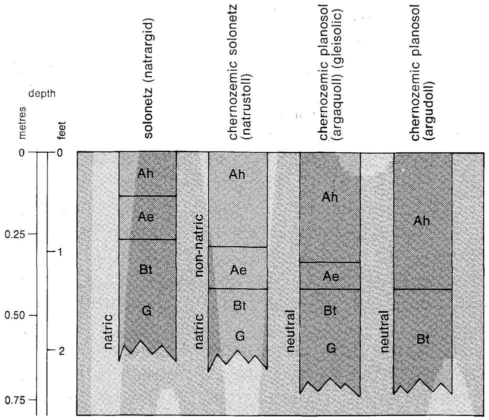


Figure 4: *Influence of drainage on soil formation.* Ah, rich in humus surface horizon (layer); Ae, eluvial surface horizon; Bt, textural B (clay illuviation); G, gley.

Drainage and topography. Leaching is impossible without drainage; thus, impeded drainage has to some extent the same effect as drought. Impeded drainage may result in the formation of soils with 2:1 clays, well saturated with bases, saline, gypsic, or calcareous horizons. It also causes gleization and may result in the formation of a water table, which has various influences, especially salinization and formation of various pans—petrocalcic horizons, duripan, ironstone, fragipan, laterite, etc. (see Figure 4). Topography greatly influences erosion, drainage, flooding, and subirrigation, which have a profound influence on soil formation, as has been discussed. Soils on slopes are usually shallow and stony (lithosols), and those in the valleys are alluvial.

Variation in humus

Vegetation and living organisms. Vegetation largely determines the kind of humus that is formed. Some trees, especially conifers, usually produce raw humus that may cause podsolization. Other trees produce weakly acid humus capable only of producing iron mobilization and clay illuviation. Such soils are brown (brunisolic) and more or less leached. Grasses produce mild humus rich in acids that stabilize clay. The distribution of humus along the profile also depends on the type of vegetation. Grasses have numerous roots that decay rapidly and enrich soil in humus to considerable depth. Tree roots contribute little to soil humus; thus, the humic horizon of forest soils is usually shallow (see Figure 5).

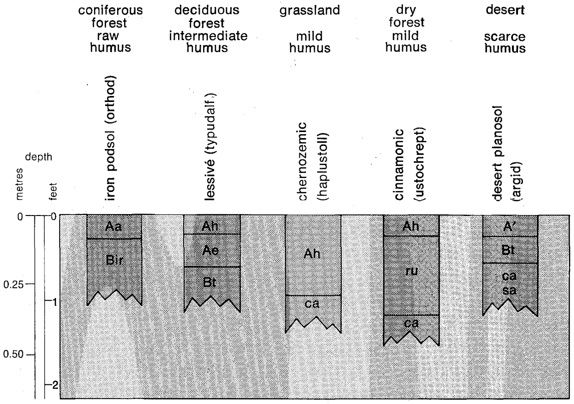


Figure 5: *Influence of vegetation and humus on soil formation.* Aa, albic (ashy) surface horizon (layer); Bir, iron podsolic B (spodic horizon); Ah, humus rich surface horizon; Ae, eluvial surface horizon; Bt, textural B (clay illuvial); ca, lime accumulation; ru, rubified, A′, poor in humus surface horizon; sa, salts accumulation.

Vegetation protects soil from the Sun's rays and retards organic matter decay; it absorbs calcium, potassium, and phosphorus from the lower soil layers and counteracts their leaching; and it protects soil against erosion. Erosion, even hydraulic, is severe in deserts because there is no vegetation to protect the soil. Living creatures within the soil such as earthworms and termites mix the various soil horizons and counteract their differentiation; they also mix soil plant residues on the soil surface and contribute to organic matter decay and humus formation. A certain mixing also occurs when old trees are uprooted by wind.

Parent material. Soil is parent material that has been altered by various processes, and its ultimate nature depends upon its original composition. In soils formed from unconsolidated sedimentary rocks, much of the clay is inherited from parent material. Materials rich in lime are difficult to decalcify, acidify, or podsolize. Materials rich in iron usually give red soils.

Time. Although in geological terms soil formation is rapid, in human terms it requires many decades or centuries. Soil age has a great influence on soil properties. The age of a soil usually is measured from the time the rock was first exposed to the surface or from the time the alluvial material was deposited. The time factor is not always clear because erosion and deposition are continual processes. Moreover, clay is often much older than the deposition, having been formed in one place and transported to another.

Young soils commonly have an undifferentiated profile. Saline, gypsic, and calcic horizons are formed more rapidly than those of clay accumulation, whereas a humic horizon may be formed in a few decades. Young soils are richer in amorphous clays and, under conditions that favour acidification, are less acid than old ones; pH increases with depth in young soils, and it decreases with depth in old ones.

The polygenesis of soils

Interaction of factors and processes. Any effect that each factor has on soil formation depends on the other factors; for example, the effect of climate depends on drainage, parent material, time, etc., and the other factors are similarly interdependent. Different soils are formed under the same climatic conditions according to drainage, parent material, time, vegetation, and other relevant influences; in like manner, different soils may be formed from the same parent material. Because soil is the result of various processes and each factor acts on several of them, it is better to relate soils to processes, and through processes to factors. The environment under which a soil is formed may vary during its formation: drainage may improve or deteriorate; exposure to such factors as erosion, flooding, or subirrigation may be ended or initiated; or there could be a change in vegetation. Thus soil does not always reflect the conditions under which it is encountered.

Climatic changes are very slow, and to affect soil formation appreciably they must be radical. A change in drainage will impede leaching, just as will a very drastic reduction of leaching rainfall; waterlogging may cause the formation of raw humus that produces podsolization in a warm climate, that otherwise favours organic matter decay and excludes podsolization, when drainage is normal. Little leaching rainfall is needed to leach a coarse soil; a great deal of rainfall is needed to leach a clayey soil. Acidification is faster and more extensive when parent material is poor in bases. For all these reasons, caution should be used when attributing the properties of a soil to climatic changes or, still more, when explaining climatic changes on the basis of pedologic evidence.

There are cases, however, in which such evidence is convincing. In Australia and Africa, for example, soils that are rich in 1:1 clays are found in the desert; they have most certainly been formed under a considerably more rainy climate. Because allitic weathering (that producing chiefly 1:1 clays) is deep, subsequent erosion and deposition merely transport such soils from one place to another. In some cases 1:1 clays have been transported great distances from where they were formed: for example, in the delta of the Paraná, near Buenos Aires, there are soils with clays that came from the centre of Brazil, and many soils of northern India have 2:1 clays that have been formed in the Himalayas.

The frequent variation in conditions during soil formation, resulting in polygenetic soils, usually is more closely related to topography than to climate; also, the clay found in a soil may have been formed in past time or at a distant site.

SOIL GROUP CLASSIFICATION AND NOMENCLATURE

Observation of natural categories. The question of soil classification is a controversial one, but there is no doubt that classification should be based on categories that exist in nature, rather than on arbitrary creations of a classification system.

Rationale of group selection

In a natural classification, groups are recognized and subsequently arranged in a system; the system may change, but the groups will pass almost unchanged from one system to another. Pedologists have approached soil classification in such a manner. Observing that some soils have an ashy surface horizon, resting on another darker in colour and richer in fine earth, pedologists called them podsols (ashy soils). Other soils with a deep, dark horizon rich in organic matter and well saturated with calcium were termed chernozems (black soils). Soils of rather undifferentiated profile and of red colour were named krasnozems (red soils). In this way, groups of soils have been recognized and arranged in classification systems; although opinions differ as to how these groups should be arranged within a classification system, essentially the same groups exist for all pedologists.

Fundamental classes of soil

The soil continuum. Because soils form a continuum, it has been difficult to follow the example of other scientists in categorizing on the basis of clear distinctions between groups. The same processes are involved in the formation of various soils, with the result that a soil may have the essential features of ando soils (amorphous clays) and those of chernozemic soils (deep humic horizon, dark and well saturated with bases) and thus be both ando and chernozemic. Soils may reflect many such heterogeneous conditions—for example, chernozemic (mollic) and gleisolic, gleisolic and podsol, and the like. Accordingly, soil classification must be manifold, recognizing that the same soil may belong to various groups. When this necessity is understood, the problem of soil classification is simplified, and the major categories shown in Table 2 may be recognized (see Figure 6).

Table 2: Major Soil Types

Ando	rich in amorphous clays
Brunisolic	2:1 clays; braunified or acid horizons
Chernozemic	2:1 clays; deep neutral dark humic horizons
Cinnamonic	2:1 clays; reddish colour; rich in dehydrated iron oxides
Dark clays	2:1 clays; rich in expanding clays; well saturated with bases
Gleisolic	gray-blue colours or mottles due to waterlogging
Kaolinitic	tropical; rich in 1:1 clays; may have undifferentiated profile, or horizons rich in iron concretions or laterite
Lessivé	clay illuviation produced by organic matter
Organic	extremely rich in organic matter; may have peaty and/or organic horizon
Podsols	eluviation of alumina and iron, usually producing ashy sandy surface horizon and illuvial horizon rich in amorphous clays
Planosols	clay eluviation produced by sodium; illuvial horizon not natric
Rankers	2:1 clays; humic horizon resting on rock or permafrost
Raw	2:1 clays; undifferentiated soil profile
Rendzinas	high lime content
Solonchaks	rich in salts such as chlorides, sulfate, etc.; related to gypsisols, rich in gypsum
Solonetz	clay illuviation produced by sodium; natric illuvial horizon

Supplementary descriptions. Various other terms, denoting special features, also may be precisely and simply defined and included in soil nomenclature. For example: clay, loam, sand (referring to soil texture); lithosol (stony, gravelly); regosol (sandy); aeolian (formed by wind, dune); hammada (stony desert); and rock outcrop. Acid (H s.), natric (Na s.), magnesic (Mg s.), Ca soil, eutrophic, and dystrophic all refer to absorbed cations. Takyr and lunettes are special kinds of saline soils. Roof clays and sulfate saline clays (cat clays) are special kinds of clays. Gray-brown podsolic, noncalcic brown, graywooded, red-yellow podsolic, and lateritic podsolic are different kinds of lessivé soils; the first three have 2:1 clays, the later two are kaolinitic. Sod and humic are used for soils unusually rich in humus for their groups. Prairie, chestnut (kastanozems), and reddish chernozemic are different kinds of chernozemic soils. Podsolized, leached, or degraded chernozems are types of brunisolic or leached soils or both. Claypan planosols are planosolic clays. Serozems are rendzinas lacking humic horizon. Rubrozems are kaolisols with a deep humic horizon. Red desert are mini-planosols (thin horizons). Mini-planosols, mini-solonetz, and mini-podsols are planosols, solonetz, or podsols with thin horizons. Arctic browns are brunisolic permafrost rankers. Terra rossa is a cinnamonic soil from hard limestone. Terra roxa is an ando-kaolisol intergrade soil. Terre de barre is a eutrophic latosolic kaolisol. Low humic gley is an

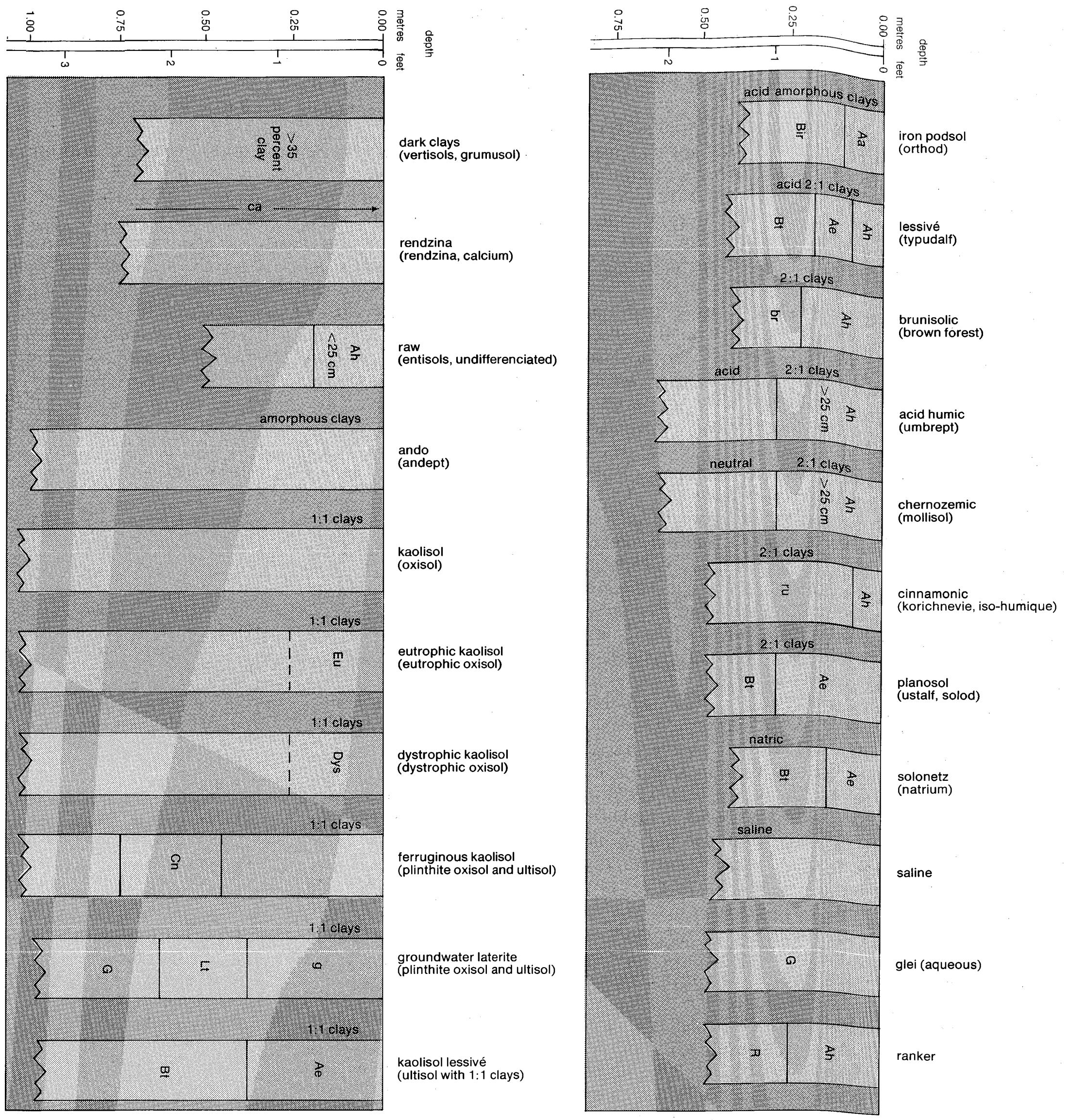


Figure 6: *Fundamental characteristics of some soil groups.*
Specifications referring to reaction (acid, neutral) and to type of clay refer to the whole profile. Aa, albic (ashy) surface; Bir, iron podsolic B (spodic); Ah, humus rich surface h.; Ae, eluvial surface h.; Bt, textural (clay illuvial) B; br, braunified; ru, rubified; G, gley; R, rockbed; Eu, eutrophic (rather well provided with bases) soil; Cn, ferruginous concretions and nodules; Lt, laterite; Dys, dystrophic (poor in bases) upper soil.

acid gleisolic soil. Humic gley is a gleisolic chermozemic soil. Desert crust is an accumulation of products of weathering in virtual absence of leaching. Alluvial, desert, forest, forest-steppe, grassland, groundwater, lowland, mangrove, meadow, mountain, paddy, paramo, prairie, tropical, and tundra are environmental, nontaxonomic terms often used in soil nomenclature.

Precise definition of soil properties

Seventh Approximation System. A system based on measurable soil properties rather than on theories of soil formation, the 7th American Approximation System, was introduced in 1960 and has expanded continuously by supplements. It is the first approach to precise definition of soil groups and pedologic terms, and, moreover, it has implicitly adopted manifold classification: adjectives (mollic, spodic, andic, etc.) are used to show relationships with other groups; names are formed by aggregation of formative elements that denote special characteristics, and consequently groups. Many of the groups named in the preceding section correspond in 7th Approximation to several taxa the nomenclature of which includes the corresponding formative element. For example, andepts and andic correspond to ando; histosols to organic; spodosols to podsols; vertisols to dark clays. The formative element aqu denotes gleisolic; natrarg, solonetz; rend, rendzina; sal, saline.

Indication of soil properties. Some formative elements show soil properties: acr denotes extreme weathering; allic, free aluminum; anthr, man-made soil; arenic, sandy; calci, calcic; chrom, high chroma (red colours); dys or dystr, dystrophic; eu or eutr, eutrophic; gibbs, presence of gibbsite; hydr, saturated with water; lithic, stony; pale, old (soil); pell, dusky colour; psam, sandy; rhod, red; sal, saline; sider, rich in free iron; sombri, dark colour; thapto, a buried soil is included in the profile; vermi, wormholes, worm casts, or filled animal burrows; vitr, ando rich in unweathered volcanic glass.

Identification of horizons. Other formative elements denote presence of certain horizons: abruptic, abrupt textural change; alb, albic horizon; cumulic, thickened humic horizon; dur, duripan; ferr, B horizon rich in free iron, or iron concretions or nodules; frag, fragipan; glossic or gloss, horizon A_2 is tongued; hapl, typical profile; hum, unusually deep humic horizon, or humus podsolic B horizon; ochr, light-coloured A horizon; pachic, thick A; petrocalcic, petrocalcic horizon; plac, thin pan; plinth, laterite; ruptic, broken (discontinuous) horizon; stratic, stratified horizon; superic, laterite in the surface; umbr, deep nonchernozemic humic horizon. Other formative elements of the nomenclature indicate the manner in which the material has been deposited: fluv; alluvial; limnic, lacustrine.

Climatic elements of the system. Although soil temperature and moisture are not intrinsic soil properties, they are used as diagnostic features in the 7th Approximation; *i.e.,* soils are classified according to the climate in which they are encountered. Such formative elements include: bor, boreal; cryo, very cold; pergelic, permafrost; torri, torrid (usually dry); trop, small annual range of temperature; ud, usually humid; ust, dry but not long dry seasons; xer, long dry seasons.

The 7th Approximation System demonstrates that soil description, classification, and nomenclature may be achieved with few symbols that, in combination, can represent the enormous variety of soils existing in nature.

GEOGRAPHIC DISTRIBUTION OF SOILS

Because soil formation depends so much on climate, topography, vegetation, and parent material, soils vary in association with other geographic features, and soil regions may be distinguished in the same manner as climatic and vegetation regions. Such soil regions lend themselves to being further subdivided or combined into higher groups. Each region contains a variety of soils, many of which are also encountered in other regions, but the distribution of soils follows a definite pattern.

The function of soil maps

The concept of soil region is important from both a theoretical and a practical point of view. Theoretically, soil distribution is often the key to answering many questions about soils. From a practical viewpoint, it is difficult and costly to prepare detailed maps of soil locations, which can vary in short distances—sometimes within a few metres.

Polar regions. *Tundra.* Vegetation is scarce in tundra regions (see Figure 7) and consists of herbs and grasses that do not produce very acid humus. Rainfall in the tundra is low and leaching rainfall (Ln) is still lower: 40 millimetres (one inch equals 25 millimetres) in Baker Lake, Northwest Territories; 20 in Eureka, Alaska; 60 in Resolute, Northwest Territories, Canada; 30 in Barrow, Alaska; 40 in Anadyr, 70 in Bulun, 50 in Nizhnekolymsk, and 40 in Russkoye Ustye, Siberia. The bases that are leached from the surface accumulate in the lower part of the profile and circulate in it through plants and by capillarity (the soil surface is often humid and soils are shallow); moreover, drainage is often impeded by permafrost. Acidity is not sufficient to produce podsolization, but it reacts with iron to produce complexes that result in brunisolic soils. The soils formed, arctic brown, are relatively shallow; the pH increase with depth is more rapid than usual; and there is a tendency toward the formation of a humus-rich horizon on the surface of permafrost. In the drier parts of tundra, the lower horizons are often calcareous.

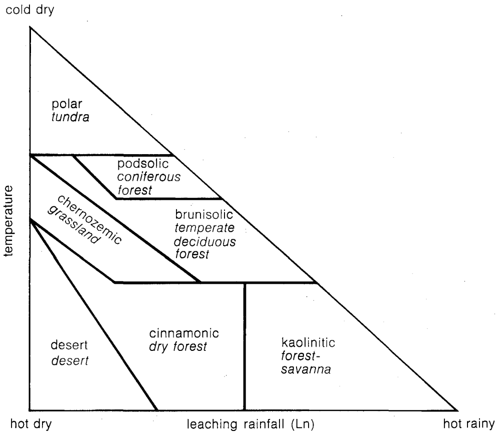


Figure 7: Distribution of soil regions with associated vegetation according to temperature and leaching rainfall (see text).

Waterlogging is frequent in tundra because of the presence of permafrost, resulting in the accumulation of organic matter, a prevalence of reducing conditions and peat formation. Gleisolic and organic (half bog) soils are frequent. The general pattern of distribution is arctic brown in well-drained sites to gleisolic and half-bog mini-podzols in depressions. Alternate freezing and thawing results in a patterned surface (polygonal).

Subglacial desert. In the subglacial desert vegetation is almost absent, weathering is very slow, and the soils formed are very shallow (lithosols, etc.). The limits of tundra and subglacial desert soil regions correspond rather well with the homonymous climatic regions.

Podsolic regions. With a short summer and conditions not too frosty for forest, so that an evergreen coniferous forest prevails, a raw, highly acid humus is produced and podsolization takes place. The soils formed in well-drained sites are iron podsols.

Drained and waterlogged conditions

In waterlogged sites organic matter accumulates, reducing conditions prevail, and gleisolic soils, sometimes organic, are formed. When there is no drainage, bases cannot be eliminated from the profile and soil is less acid, a condition common in the drier podsolic regions. When there is a slow drainage, however, soil is acidified and becomes humus podsol, somewhat gleisolic and peaty. Such soils usually have heath or sphagnum vegetation.

Humus podsols are more common in Atlantic regions because of their higher rainfall, than in continental regions. Hardpans are more common in Atlantic regions. Some regions of Central Siberia have permafrost that interferes with drainage; moreover leaching rainfall is low, with the result that iron leaching is incomplete and the eluvial horizon is yellow rather than ashy. The yellow podsols may be considered as intergrades to arctic brown.

Brunisolic regions. With a longer summer, deciduous forest prevails, which produces milder, less acid humus that forms brunisolic soil, more or less leached. Where leaching rainfall is low and parent materials are calcareous or basalt, soils are neutral (braun erde); in the opposite rainy condition, soil is acid (brun acide).

The most important brunisolic regions are in western Europe, the eastern United States, the state of Washington, and the adjacent Canadian coast. In western Europe, there are many small regions with calcareous materials, basalt, or loess, and low leaching rainfall. Braun erde abound; some of them are chernozemic or have a calcareous horizon at some depth or both. In the transition zone to the chernozemic region of North America, chernozemic braun erde (brown earth) and chernozemic lessivé (prairie) are common, and in the Appalachian Mountains, brun acide.

Chernozemic regions. A grassland vegetation with chernozemic soils occurs when spring is not dry, winter is relatively cold, and leaching rainfall is low. Calcareous, gypsic, or salic horizons are frequent. Solonchaks, solonetz, and planosols—more or less gleisolic and chernozemic—are common in poorly drained depressions.

The three chernozemic regions

The most important chernozemic regions of the world are the Danube Basin and the southern Soviet Union in Eurasia, the Great Plains of the United States and the prairies of Canada, and the Pampas region of Argentina. But there are some differences among them. The parent materials of Russian chernozems are more calcareous and the leaching rainfall is lower (for the same humidity index). Consequently, soils are richer in lime; effervescence with hydrogen chloride begins immediately below the humic horizon or within it; clay eluviation and planosolic chernozems are not common; and the humic horizon is, in general, deeper and richer in organic matter. In the North American Great Plains, parent materials are less calcareous, lime is encountered at greater depth or not at all, clay eluviation is more frequent, planosolic chernozems more common, and the humic horizon is in general shallower and poorer in organic matter. In Argentina, parent materials are chiefly volcanic; cation-exchange capacity is unusually high (andic chernozemic); much sodium is released by weathering; clay eluviation is frequent; and planosolic chernozems, solonetz, and solonchaks are common.

The general pattern of soil distribution in chernozemic regions is chernozemic soils in well-drained areas, with planosols, solonetz, and solonchaks, more or less chernozemic and gleisolic, in areas of deficient drainage.

Cinnamonic regions. In climates with dry seasons where soil is dried thoroughly to considerable depth, fine crystallites or irreversibly dehydrated iron sesquioxides are formed, giving reddish colours and cinnamonic soils. In cinnamonic soils organic matter decays rapidly and does not accumulate on soil surface. Humus originates chiefly in roots and it is well distributed along the profile. Many cinnamonic regions have a Mediterranean climate. Those with a monsoon climate are usually transitional to brunisolic, chernozemic, desertic, or kaolinitic regions. Except for Australia, a Mediterranean climate is associated with mountainous topography and high frequency of limestones and volcanic materials; moreover, leaching rainfall varies considerably from zero to very high figures. Brown cinnamonic (bruns mediterraneans) and red cinnamonic (rouges mediterraneans) are common in areas with high leaching rainfall or ferruginous materials or both. They are sometimes acid and rich in 1:1 clays, being intergrades to kaolisols (krasnozems).

Desert regions. Vegetation is very scarce in the desert, and soil erosion is very severe from wind and occasional rains; weathering is slow and shallow, and leaching is almost absent. Soils are very poor in organic matter and do not have a humic horizon.

Occurrence of mini-horizons

Because sodium elimination is difficult, autochthonous soils are usually planosols and solonetz with very thin horizons (mini-planosols and mini-solonetz); a calcareous horizon usually underlies the B horizon. Soil is often covered by coarse material (desert pavement), which partly results from wind erosion and protects the soil. In such absolute deserts as the Chilean desert, however, products of weathering remain in the place of their formation and a loosely cemented desert crust is formed. Materials eroded from higher land accumulate in depressions, forming dunes and alluvial soils. Where waters accumulate, planosolic, solonetzic, and saline soils abound, and "salares" also are frequent. There are many good alluvial soils, however (rendzinas, chernozemic, etc.), and some of them are gleisolic.

The general pattern of soil distribution may be summarized as follows: high plateaus and plains areas exhibit autochthonous soils, mini-planosols, mini-solonetz, and solonchaks with salares (salt accumulations); slopes exhibit lithosols; and depressions exhibit dunes and alluvial soils, with many solonetz, solonchaks, serozems, and gleisolic soils with salares.

A distinction may be made between American (North and South), Asian, and African deserts. American deserts received much volcanic ash from active volcanoes, and mini-solonetz, mini-planosols, and duripans are plentiful. In Asia, calcareous materials abound, and raw (undifferentiated) soils that produce effervescence with hydrogen chloride are common. In Africa and Australia, palaeosols formed from kaolinitic materials are frequent.

Kaolinitic regions. The high leaching rainfall of tropical countries, even though there may be a long dry season, results in formation of 1:1 clays. Soils are usually acid in the lower horizons, whereas the upper horizon may be eutrophic. The difference between eutrophic and dystrophic soils, which is important from an agricultural point of view, seems well associated with the presence or absence of a dry season. A dry season, even a short one, activates decay of organic matter, favours fires, and interferes with the production of acid humus, which is the principal factor of acidification, several times more effective than water mixed with carbon dioxide.

Water-logging and iron concretions

In soils rich in 1:1 clays, waterlogging produces segregation of iron in the form of concretions or nodules. For that reason laterite is common in waterlogged areas, and, because it is impermeable, aggravates waterlogging and may reach the surface. In granite, gneiss, and similar materials, very common in Africa, the iron released by weathering may adhere to sand or gravel surfaces and cover them with a red coating. When soil is thoroughly dried periodically, such coatings become irreversibly dehydrated, protect the particle or gravel from further weathering, and a ferruginous horizon is formed. Such protection is insufficient in the upper horizon, where organic acids abound, and also in the lower horizons, which are waterlogged continuously; as a result, the ferruginous horizon is encountered at some depth. Some of the iron may come from the waterlogged horizons.

It is of interest that concretions of gravel and laterite contain unweathered material imprisoned in iron sesquioxides. Such material is virtually absent in the overlying and underlying horizons, indicating that both ferruginous horizons and laterite are formed at an early stage of weathering. This also indicates that stone lines of geologic origin may contribute to the formation of ferruginous horizons. It is, however, difficult to believe that there was a stone line parallel to soil surface each time a ferruginous horizon was formed. Ferruginous horizons and laterite are seldom found in soils formed from easily weatherable materials such as basalt, fine sand, and in continuously humid climates.

Young soils from volcanic ash are naturally ando in type. With time amorphous clays crystallize, cation-exchange capacity decreases, and soil becomes first terra roxa (intergrade ando-kaolinitic) and then kaolinitic. Terra roxa, however, is often made up of young soils formed from basalt and abounds in basaltic slopes, where erosion does not permit soil to become old (Misiones, Argentina; coffee region, Brazil).

Mountainous regions. Erosion is the main characteristic of mountainous regions. The soils are usually young, consisting of lithosols and rankers in slopes and alluvial soils in the valleys and surrounding the mountain plains.

Three kinds of mountainous regions may be distinguished: humid-temperate, dry, and humid-tropical. In humid-temperate climates organic matter accumulates and there is an abundance of brunisolic soils (braun erde, brun acide). There are also lessivé and podsols, but many podsols do not have an ashy horizon. The spodic horizon, rich in amorphous clays, is overlain by a gray-brown horizon (brown podsols); and limestones often form lithosolic rendzinas. Acid humus and peaty soils, more or less gleisolic, are common in depressions.

In dry but not desertic mountains, grassland vegetation, sometimes scattered with woody plants, is common. Lithosols and other raw soils and rendzinas, basically cinnamonic and chernozemic, are widespread, with saline, solonetzic, planosolic, and organic, relatively gleisolic soils in depressions.

In humid-tropical mountains the growing season is longer and the growth index much higher; soils are deeper and richer in organic matter (some giant podsols have been found); otherwise they resemble those of humid-temperate climate of the same summer type. At lower altitudes kaolinitic weathering begins and forms kaolinitic soils; erosion, however, results in soils remaining younger than in the lowlands, eutrophic soils and terra roxa are more frequent, and soils are also richer in organic matter. These characteristics partly account for the higher density of population in mountainous regions.

Intermediate and modified regions. Passage from one type of soil region to another is gradual, and intermediate regions can be recognized. Moreover, preponderance of allochthonous alluvial soils, waterlogging, or presence of a particular material justify separation of some regions from the main type.

Nature of transition belts

In the transition belt from podsolic to brunisolic regions, the prevailing soils are podsols and lessivé (gray-brown podsolic), and the distribution is chiefly determined by parent material. In the transition between podsolic and chernozemic regions, Soviet authors recognize a forest-steppe zone with gray-wooded soils (lessivé with somewhat ashy horizon, better saturated with bases, and sometimes calcareous at some depth). Near this forest-steppe region are many sod-podsols (podsols with rather well developed humic horizon), attributed to cropping. An analogous transitional region in Canada occurs between the podsolic and chernozemic region.

The prairie soils in the United States are characteristic of the transition between chernozemic and brunisolic regions and may be considered as chernozemic lessivé or chernozemic braun erde. They are also found in the Danube Basin and other parts of Europe, and combine characteristics of both the chernozemic and brunisolic, or lessivé groups. The region of red-yellow podsolic soils of the southeastern United States may be considered as a transitional kaolinitic-brunisolic region with soils usually lessivé and very acid, usually low in cation-exchange capacity in comparison to their clay content. In other parts of the world, the transition between brunisolic and kaolinitic coincides with mountainous topography or special parent materials. The transition between brunisolic and cinnamonic regions in Europe is marked by cinnamonic-brunisolic intergrades (brun mediterraneans, southern braun erde, etc.); that between chernozemic and cinnamonic regions in the United States by reddish chernozems. The transition between cinnamonic and desertic regions in the United States is marked by reddish brown soils. Comparable soils are encountered in many parts of Asia, Africa, and Australia. Planosolic, solonetzic, and saline soils are common in these transitional regions.

In the transition between cinnamonic and kaolinitic regions, materials rich in easily weatherable minerals give soils with 2:1 clays, whereas 1:1 clays predominate in the opposite case, a condition that prevails in many parts of the Mediterranean Basin with high leaching rainfall and mild winters (*e.g.,* Portugal). In Africa and Australia, however, many kaolisols of such a transitional region are palaeokaolisols.

Humid, and dry volcanic soils

Volcanic regions. Volcanic ashes have such great influence on soil formation that volcanic regions constitute a special group. Ando soils are frequent in volcanic regions. A distinction may be made between humid volcanic regions with acid ando, often deficient in phosphorus, and dry volcanic regions with soils well provided with bases, often chernozemic. Planosolic, solonetzic, and saline soils, and duripans or calcareous horizons, are frequent in dry volcanic regions and still more so in desertic ones, where planosols and solonetz become mini-planosols and mini-solonetz. Ando soils are usually fertile, and as a result settlements grow around volcanoes in spite of dangers. The most important volcanic soil regions of the world are: the Cordillera (western mountain ranges) of North America; the entire Pacific coast of North and South America; the Great Basin of North America located within the Cordillera to the west of the Rocky Mountains; the Andes mountain ranges of South America, the plains to the north and west, and many lands to the east; all of Patagonia, the Pampas region of Argentina, and to a certain extent the Chaco; many parts of the Mediterranean Basin, especially in Italy; Indonesia, the Philippines, Japan, and Kamchatka; and some parts of Ethiopia, Kenya, Uganda, Rwanda, Burundi, and Cameroon.

Alluvial and waterlogged regions. Some regions have alluvial soils of materials that came from regions of different climate. The most conspicuous examples are the alluvial region of northern Italy, one of the better agricultural regions of the world, and those of northern India and Indochina, which nourish a dense population or export much food. In Australia, and to a certain extent in Africa, are regions in which the soils have been formed from kaolinitic materials developed under different conditions; these regions may be called palaeokaolinitic.

In some regions waterlogging is so common that the soils may be called gleisolic. Some of them are podsolic, for example the southwestern coast of Hudson Bay in Canada, most of Finland, and most of northern Soviet Union. Others are chernozemic, as the "Pampa Deprimida" of Buenos Aires, Argentina. Still others are found in the transition region between chernozemic and kaolinitic soils of southern Brazil, Uruguay, and Corrientes, Argentina. Soils vary as a consequence: gleisolic humus podsols and peat in the first case; gleisolic chernozemic planosols and solonetz in the second; chernozemic lessivé (prairie), chernozemic gleisolic (humic gley), dark clays (vertisols), planosols, acid gleisolic (low humic gley), kaolinitic podsolic (red-yellow podsolic), with krasnozems and terra roxa on basaltic hills in the third.

Some tropical regions have relatively young soils, many of which have been formed from calcareous materials. This is the case in the Yucatán Peninsula of Mexico, in the West Indies, and on the coral islands of Oceania. Their soils are predominantly brunisolic-kaolinitic.

BIBLIOGRAPHY. Some classic and general works include V.V. DOKUCHAIEV, "Abridged Historical Account and Critical Examination of the Principal Soil Classifications Existing," *Trans. St. Petersburg Soc. Nat.,* 10:64–67 (1879), in Russian; K.D. GLINKA, *The Great Soil Groups of the World and Their Development,* 2nd ed. (1937, Eng. trans. from the German edition of 1914); EMIL RAMANN, *Bodenbildung und Bodeneinteilung* (1918; Eng. trans., *The Evolution and Classification of Soils,* 1928); C.F. MARBUT, "A Scheme for Soil Classification," in the *Proceedings and Papers of the First International Congress of Soil Science, June 13–22, 1927,* vol. 4, pp. 1–31 (1928); W.L. KUBIENA, *Bestimmungsbuch und Systematik der Böden Europas* (1953; Eng. trans., *The Soils of Europe,* 1953), important for brunisolic, podsolic, and gleisolic soils; A.A. RODE, *Soil Science* (Eng. trans. 1962), an excellent account of Russian pedology; and J. PAPADAKIS, *Soils of the World* (1964), a discussion of their formation, diagnostics, classification, geographic distribution, and agricultural potentialities. Soil genesis is covered in the following works: H. JENNY, *Factors of Soil Formation* (1941); M.L. JACKSON and G.D. SHERMAN, "Chemical Weathering of Minerals in Soils," *Adv. Agron.,* 5:219–318 (1953), excellent account of weathering, especially differences in weatherability between minerals; B.B. POLYNOV, *The Cycle of Weathering* (1937; orig. pub. in Russian, 1934), points out the consequences of differences in solubility between the principal soil constituents; B. BARSHAD, "Chemistry of Soil Development," in F.E. BEAR, (ed.), *Chemistry of the Soil,* 2nd ed., pp. 1–70 (1964); and J. PAPADAKIS, *Geografía agrícola mundial* (1960), introduced the concept of "leaching rainfall" (Ln), and gives a formula for its calculation. Basic works on soil classification include the following; J. THORP and G.D. SMITH, "Higher Categories of Soil Classification: Order, Suborder, and Great Soil Groups," *Soil Sci.,* 67:117–126 (1949), a fundamental work on the subject; UNITED STATES SOIL CONSERVATION SERVICE, *Soil Classification: A Comprehensive System, 7th Approximation* (1960 and suppl.), a classic contribution to soil classification, although it needs to be improved; A. LEAHEY, "The Canadian System of Soil Classification and the Seventh Approximation," *Soil Sci. Soc. Am. Proc.,* 27:224–225 (1963); and J. PAPADAKIS, "Some Considerations on Soil Classification: The 7th Approximation," *Soil Sci.,* 94:115–119 (1962). Basic works on experimental pedology include GEORGES PEDRO, *Contribution à l'étude expérimentale de l'altération géochimique des roches cristallines* (1964); E.G. HALLSWORTH, "An Examination of Some Factors Affecting the Movement of Clay in an Artificial Soil," *J. Soil Sci.,* 14:360–371 (1963); and C. BLOOMFIELD, "A Study of Podzolization," 6 pt., *J. Soil Sci.,* vol. 4–6 (1953–55). MILO I. HARPSTEAD and FRANCIS D. HOLE, *Soil Science Simplified* (1980), is an introduction for the layman.

(J.Pa.)

The Solar System

The solar system consists of the Sun and all matter under the gravitational control of the Sun in its 225,000,000-year period of revolution around its galaxy. Any body not having sufficient velocity to escape from the Sun remains in a closed path (orbit) around it and is a part of its system.

This article surveys the structure and general characteristics of the solar system as a whole and summarizes the key theories concerning its origin. It also describes the components of the solar system and their properties.

For coverage of related topics in the *Macropædia* and *Micropædia,* see the *Propædia,* sections 131, 133, 211, 212, and 213. (Ed.)

The article is divided into the following sections:

Table 1: Comparative Data for Solar System Bodies

	mass (Earth = 1)*	equatorial Earth radius (Earth = 1)†	mean density (g/cm^3)	rotation period‡ (h = hours, d = days) mean solar	sidereal	inclination of axis	mean distance from Sun (a.u.)§	orbital eccentricity	orbital inclination	number of known satellites
Mercury	0.055	0.383	5.44	175.97^d	58.65^d	near 0°	0.387	0.206	7°.004	0
Venus	0.8150	0.9488	5.245	116.75^d	243.02^d	177°2′	0.723	0.007	3°.394	0
Earth	1.0000*	1.0000†	5.517	24^h	23.934^h	23°45′	1.000	0.017	0° by definition	1
Mars	0.1074	0.5326	3.95	24.660^h	24.623^h	25°2′	1.524	0.093	1°.850	2
Jupiter	317.9	11.19	1.33	9.926^h	9.925^h	3°1′	5.203	0.048	1°.305	16
Saturn	95.26	9.46	0.70	10.657^h	same	26°7′	9.539	0.056	2°.489	20+
Uranus	14.6	4.11	1.17	17.3^h	same	97°9′	19.182	0.047	0°.773	15
Neptune	17.2	3.88	1.66	16.3^h	same	28°8′	30.058	0.009	1°.773	8
Pluto	0.0025	~0.18	0.5–0.9	6.387^d(?)	same	unknown	39.785	0.254	17°.137	1
Ceres ‖	0.0002	0.078	~2.3	9.080^h(?)	9.078^h	direct rotation	2.768	0.077	10°.60	0
Halley's Comet	~2×10^{-11}	4.7×10^{-4}	~1(?)	—	~10^h	20–40°(?)	17.94	0.967	162°.24	0

*The mass of Earth is 5.9733×10^{24} kilograms. †The equatorial radius of Earth is 6,378.136±1 kilometres. ‡The mean solar day is the length of the day as measured by the average motion of the Sun, the clock day of Earth. The sidereal period is that required for a 360° rotation relative to the stars. Only if a planet did not move relative to the Sun could the sidereal and mean solar days be exactly equal, although for the outer planets, where the orbital motion is very slow, the differences are extremely small. §One astronomical unit (a.u.) is the average distance of Earth from the Sun, or 149,598,000 kilometres. ‖ Largest of the known asteroids.

Plan of the solar system. More than 99 percent of the mass of the solar system is in the Sun, whose mass is 1.989×10^{30} kilograms (4.385×10^{30} pounds)—333,000 times that of the Earth. The Sun is a gas (predominantly hydrogen) throughout its nearly 700,000-kilometre (435,000-mile) radius because it is heated by continuous nuclear fusion of the hydrogen in its core, where temperatures are near 15,000,000 K (27,000,000° F). As a result, the visible surface of the Sun has a temperature of about 5,780 K and continuously radiates a power of 3.85×10^{26} watts into space. Earth intercepts less than half of one-billionth of that power, but it is sufficient to make Earth a warm, habitable planet.

Of the remaining mass in the solar system more than 99 percent is found in the nine known planets, and more than 90 percent in Jupiter and Saturn alone. Most of the planets follow nearly circular paths around the Sun, with most of the circles nearly in the same plane. Only Mercury and Pluto, the innermost and outermost planets, and also the smallest, have orbital eccentricities greater than 0.2 and inclinations greater than 5°. (Eccentricity measures the amount by which an orbit departs from circularity, being zero for a circle and one for a parabola. Inclination is the angle between the plane of a planet's orbit and that of Earth's orbit.) These and other basic figures for all of the planets are given in Table 1.

Terrestrial and giant planets

The planets can be divided physically into two groups—an inner group of small, rocky, high-density bodies (Mercury, Venus, Earth, and Mars), called the terrestrial planets, and an outer group of large, gassy, low-density objects (Jupiter, Saturn, Uranus, and Neptune), called the giant planets. This leaves Pluto as a curious anomaly—small, icy, of low density, and in its makeup more like a satellite than a planet. The giant planets have occasionally been called the major planets, but the latter term is best applied to all nine planets in juxtaposition to the minor planets, or asteroids. Discussion of a possible 10th planet is heard from time to time. Searches have been conducted, within 30° of the plane of the planets, that would have found any object brighter than about one-tenth the brightness of Pluto. There are very small variations from prediction in the orbits of Uranus and Neptune that could be caused by another planet, but the evidence is not strong.

Planetary surfaces. *The inner planets.* The inner planets are bodies made of refractory materials. They are composed of outer shells of silicates with inner metallic cores, yet their surfaces show differing morphologies because of the vastly different surface environments that evolved over their 4,500,000,000-year histories. Despite similar gross compositions, the inner planets differ in their volatile inventories—*i.e.,* substances such as carbon dioxide (CO_2) and water (H_2O). The relatively slight differences have given Mercury a hellishly hot, crater-pocked surface; have turned the surface of Venus into a gaseous, suffocating inferno; have given Mars a history of megafloods and Arctic-like permafrost; and have allowed Earth's surface to be mainly oceans of liquid water.

The outer planets. Among the outer planets the gas giants—Jupiter, Saturn, Uranus, and Neptune—do not have accessible surfaces. Their rocky cores are submerged beneath massive atmospheres. Pluto is so small and distant that it is equally unknown.

Planetary atmospheres. Atmospheres are the parts of planets and other solar system bodies that are mostly gaseous. Mercury, nearest the Sun, has no appreciable atmosphere. Venus, Earth, and Mars have thin atmospheric layers surrounding the solid portions composing the bulk of their masses. Jupiter, Saturn, Uranus, and Neptune have atmospheres that are so substantial that no detection of solid surfaces had been made by the mid-1980s, though theoretical models predict solid cores deep in their interiors. A thin atmosphere of methane (CH_4) has been detected above Pluto's surface.

Surface temperatures and pressures

For the inner planets with atmospheres, their increasing distance from the Sun results in a progressive drop of surface temperatures and is partially the cause of a like drop in surface atmospheric pressures. Venus' surface temperature is about 730 K and its surface pressure 90 atmospheres (*i.e.,* 90 times the pressure at Earth's sea level). The Earth's average surface temperature is about 289 K. Mars has an average surface temperature of 218 K and an average surface pressure of 0.007 atmosphere. The atmospheric temperatures of Earth and Mars vary with changes in season.

The temperature differences among the giant planets are due to differences in solar heating and to the amount of energy coming independently from the planetary interiors. At the level where the total pressure is equal to one atmosphere, the temperature of Jupiter is about 166 K, and that of Saturn is approximately 91 K. Uranus and Neptune are both roughly 72 K at the same pressure level. The temperatures of both Jupiter and Saturn are known to change with position on the planet and with season. The surface temperature of Pluto is estimated at 55 K, and its surface pressure is at least 0.0001 atmosphere.

Chemical composition

The chemical composition of the atmospheres of Venus and Mars are similar. The atmosphere of Venus is 96 percent CO_2 and 3.5 percent molecular nitrogen (N_2); Mars is 95 percent CO_2 and 2.7 percent N_2 with another 1.6 percent argon (Ar). In contrast, Earth's atmosphere is 77 percent N_2, 21 percent molecular oxygen (O_2), 1 percent H_2O, and 0.93 percent Ar. The remainder of the composition of each atmosphere is shared by several species of chemicals. For Venus the most abundant of these are sulfur dioxide (SO_2), H_2O, Ar, and carbon monoxide (CO); for Earth they are CO_2, neon (Ne), and helium (He); and for Mars they are O_2, CO, and H_2O.

The bulk of the outer-planet atmospheres is molecular hydrogen (H_2), with mixtures of smaller amounts of helium. Jupiter is known to be about 89 percent H_2 and 11 percent He, with minor abundances of CH_4 and ammonia (NH_3). Measurements of Saturn show 94 percent H_2 and 6 percent He, but the He content is expected to be greater in the interior of the planet; its atmosphere contains minor and trace chemicals similar to those of Jupiter. Methane is estimated to make up 0.2 to 4 percent of the total volumes of the atmospheres of Uranus and Neptune, which are otherwise mostly H_2 and He. The only known constituent of the atmosphere of Pluto is CH_4.

All of the atmospheres contain some liquid droplets and solid particles in suspension. In the case of Venus, concentrations of sulfuric acid (H_2SO_4) droplets form the bright clouds that blanket the planet. In ultraviolet light, markings are recognizable that show winds with speeds up to 100 metres per second (about 224 miles per hour). Earth's atmosphere is usually 50 percent covered by liquid water and ice clouds, often clustered in storm patterns. Clouds of water vapour and carbon dioxide are sometimes visible in the Martian atmosphere, which is filled with dust carried aloft from the surface by strong winds.

Visible clouds

The visible clouds of the giant planets generally form bands that surround the planets and are parallel to their equators. Jupiter's bands are the most striking in colour and duration; they are accompanied by smaller regions of local clouds, one of the most observed and longest lasting of which is the Great Red Spot. Saturn's clouds are more subdued in colour and contrast than those of Jupiter. The uppermost clouds of Jupiter and Saturn are probably NH_3 ice particles; at deeper levels clouds of ammonium hydrogen sulfide (NH_4HS) and H_2O may exist. The colours of the clouds may be due mostly to the presence of elemental sulfur or complex hydrocarbon molecules. Cloud patterns have been detected in the atmospheres of Uranus and Neptune; above NH_3 clouds there may be CH_4 ice clouds in the atmospheres of both planets. The absorption of visible red light by gaseous CH_4 above the cloud tops makes both Uranus and Neptune appear bluish green.

Satellites. A satellite is a body in orbit around a planet or other body that, in turn, is in orbit around the Sun or, in principle, another star. When a satellite reaches half the size of the planet it circles, as seems to be the case with Pluto and its satellite, Charon, the system might better be called a double planet. At the other extreme, the planetary rings around Jupiter, Saturn, and Uranus are made up of innumerable tiny satellites, typically a few metres in diameter in the main rings of Saturn and probably much smaller in Jupiter's rings.

The number of verified satellites in the solar system in the late 1980s was at least 57. Their diversity in size, morphology, and composition is at least as great as that among the planets themselves. In size they range from Jupiter's Ganymede and Saturn's Titan—which are larger than the planet Mercury—to tiny, irregular bodies such as Mars's Deimos. As might be expected, the inner planets have silicate satellites. The Moon has a composition very much like that of Earth's own upper layers, mostly iron–magnesium silicates, but has no water. Phobos and Deimos, the small satellites of Mars, have predominantly silicate compositions; they are probably captured asteroids.

Jovian satellites

The satellites of Jupiter present a somewhat more complex situation. Of the planet's Galilean satellites, Io, which is mainly silicate, has been well differentiated—since it shows no spectral signature of water—and is the most geologically active body in the solar system. Io is being constantly pushed and pulled, as it orbits close to Jupiter, by the gravity fields of Jupiter and of Europa, the next satellite out. Europa is also mostly silicate but exhibits the spectral signature of water ice and is very bright. Ganymede has substantially more water than the inner two Galilean satellites and exhibits huge, rifted icy plates that, earlier in its history, were probably mobile over a plastic or partially molten water interior. Callisto has even more water than Ganymede and a relatively stable crust with little mobility over geologic time. Amalthea, a small inner satellite, could well be a captured asteroid and is probably mainly silicate in composition. The outer satellites of Jupiter are all small. Some are in retrograde orbits and may be captured minor planets or condensed remnants of a circum-Jovian cloud.

Except for Titan, the major satellites of Saturn have bulk densities within 40 percent of water, and that substance is undoubtedly a major constituent of all of them. The remarkable trait common to the major Saturnian satellites is that they appear to have had episodes of geological activity, indicating internal heat sources with mobilization of surface materials. How such small icy bodies (all less than 750 kilometres in radius), with little silicate material to generate internal heat through radioactive decay, could exhibit episodes of geological activity is still a mystery.

Titan is the only satellite to possess a substantial atmosphere. Clouds of liquid methane droplets and other unidentified chemicals obscure its surface and give it a reddish orange colour. It is possible that methane plays a role on Titan much like that of water on Earth, with liquid methane seas and solid methane polar caps. Liquid hydrocarbons probably rain down from the upper atmosphere, producing a tarlike layer 100 metres thick.

The major satellites of Uranus, except for Umbriel, are similar to the largest moons of Saturn. They too are icy objects with small amounts of rocky material whose topography bears the effects of internal activity. For example, subsidence features such as fault scarps and trenches are found on Miranda and Ariel, while dark lavalike material covers the floors of various craters on Oberon.

Neptune's largest satellite, Triton, has a variegated surface. A large part of its equatorial region consists of valley-like depressions and polygonal ridge structures. Present, too, are broad plains covered by what is thought to be "lava" flows composed of water ice mixed with methane and possibly ammonia rather than of rock material. Some of these icy flows are extremely old and pitted by numerous impact craters. Plains covered by fresher-looking ice lavas have almost no such craters.

Other components of the solar system. *Asteroids.* The innumerable bodies orbiting the Sun—other than the nine major planets and their satellites—that can in principle be studied individually are asteroids and comets. During the 19th century it seemed there were many operational criteria uniquely defining each class of objects. An asteroid was defined as a small planet without obvious atmosphere in a nearly circular orbit between Mars and Jupiter. A comet was defined as an enormous volume of dust and extremely low-density gas emitted from something in an extremely elongated orbit. Today, only the evolution of a large quantity of gas remains as an observational distinction unique to comets. (See below *Asteroids.*)

Ceres

The first asteroid was discovered in 1801 and named Ceres. Ceres proved to be in a nearly circular orbit 2.77 astronomical units (a.u.) from the Sun, and modern observations give its diameter as 940 kilometres, the largest of the asteroids. (One astronomical unit is the average distance of Earth from the Sun, or 149,598,000 kilometres.) Once an asteroid has a well-determined orbit, it is given a number and a name. By the late 1980s there were about 3,500 numbered asteroids, more than 95 percent of them fitting the original criteria. Statistical studies indicate that in addition to Ceres there are more than 1,000,000 asteroids between Mars and Jupiter with a diameter greater than one kilometre, but the total mass of all these is roughly equal to that of Ceres, and the total mass of all "main-belt" asteroids, including Ceres, is certainly less than a thousandth of Earth.

Most asteroids are small objects of irregular shape. They show properties similar to the Moon—probably cratered and overlain with rocky, dusty debris. Certain differences in colorimetric properties of the asteroids suggest different compositional classes. Those of the C-class, with very low albedo (surface reflective power), are similar to carbonaceous chondrite meteorites, which are brittle, dark bodies with fine opaque material throughout and with hydrated (water-rich) mineral phases. Those of the S-class are thought to be similar to stony-iron meteorites in composition. Another, rarer group is the M-class, which is probably similar to nickel–iron meteorites.

The 5 percent of asteroids that, because of their orbits, do not fit the classic criteria include the Amor-class objects, which come nearer to the Sun than the perihelion distance of Mars (1.38 a.u.); the Apollo-class objects, which come nearer the Sun than Earth's mean distance (1 a.u.); and the Aten-class objects, whose average distance from the Sun is less than Earth's. There are at least 1,000 Trojan minor planets in the same orbit as Jupiter, but preceding or following it by roughly 60°. In 1977 a body a few hundred kilometres in diameter was discovered in an orbit extending inside Saturn's (to 8.5 a.u.) and outward nearly to that of Uranus (to 18.9 a.u.); it has been named Chiron.

Trojan minor planets

Comets. Comets are believed to be "snowballs" of frozen gases and dust a few hundred metres to a few tens of kilometres in diameter. They move in very elongated orbits, and only when they approach near enough to the Sun for the ices to begin vaporizing rapidly do they become obvious and clearly identifiable. Yet the very process that makes them visible destroys them. After 1,000 passages near the Sun, there should be little left except, possibly, a small inert core. A typical short-period comet makes a revolution in 10 years or less, giving it a total lifetime of less than 10,000 years. A large body, especially Jupiter, can make major alterations (perturbations) in the orbit of any comet passing near it, sometimes shortening the period of revolution, sometimes lengthening it. Statistical studies of this process and others affecting cometary orbits and lifetimes indicate that the total number of comets in existence may be as many as 3×10^{12}. Their combined mass would be nearly two Earth masses.

Meteoroids and interplanetary medium. Any body too small to be studied individually and moving in or through the solar system is called a meteoroid. Upon impact with an atmosphere it becomes a meteor. If some part survives and reaches the surface, it is a meteorite.

The smallest asteroids that have been studied individually are a few near one kilometre in diameter that have approached close to Earth, in some cases barely more than the distance of the Moon. There is also direct evidence of meteoroids in the solar system. Meteorites weighing from a few grams up to more than 50 metric tons (55 short tons) and roughly $3 \times 3 \times 1$ metres in size have been found after impact on Earth. Craters, such as Arizona's Meteor Crater, give evidence of yet larger impacts. Every clear night, meteors ("shooting stars") are visible streaking across the sky, most of them particles only a few millimetres in diameter that burn up from atmospheric friction without ever reaching the ground.

The interplanetary medium contains even lighter components than meteoroids, namely individual molecules, atoms, ions, and electrons from many sources. Most neutral molecules and atoms lost from comets or planets are rapidly ionized by solar ultraviolet radiation. The Sun itself is the largest source of interplanetary charged particles, which escape from the solar chromosphere. This "solar wind" has sufficient velocity to escape from the solar system, so although it is always present as a component of the interplanetary medium, individual ions, protons, and electrons are not in closed paths around the Sun.

Solar wind

Theories of origin. It is generally thought that the solar system evolved from a nebular cloud associated with the proto-Sun. As that rotating cloud cooled and condensed, it formed a disk in which compositional and thermal gradients evolved. That early nebular differentiation is reflected in the systematic change in planetary composition, physical properties, correlated surface morphological properties, and composition of planetary atmospheres from the Sun outward through the solar system. The inner planets are silicate and metal-rich, while proportionately more and more volatiles such as water and methane are incorporated in planets and satellites that presumably had a cooler formation environment. The Jupiter system is an interesting analogue to the entire solar system from the standpoint of composition and physical properties—of the Galilean satellites, Io is the closest, and it has the highest density and the least water, while Callisto is the farthest from Jupiter and possesses the most water. Nearly all of the planets and satellites in the solar system whose surfaces have been observed retain evidence of the torrential meteoritic bombardment that accompanied the final states of accretion.

Increasing evidence suggests that most of the original gases in the inner solar system were blown away during its early history. All the lighter gases, such as molecular hydrogen and helium, exist only in the colder and gravitationally more powerful giant planets. The atmospheres of Venus, Earth, and Mars are considered to have evolved substantially over the course of time by extrusion of materials from their interiors in tectonic processes and volcanism. The differences between them are probably due to the effect of different amounts of sunlight and also to the extent to which their interiors were heated. The abundance of O_2 on Earth is almost certainly due to the presence of photosynthetic life, which converts CO_2 to O_2. The massive gravity of Jupiter probably allowed it to retain the same chemical constituents that existed at its distance from the Sun in the early solar nebula. This should also be true to some extent for the other giant planets, and their compositions and those of comets should provide the best clues to the distribution of chemicals in the outer solar system at the time of its formation.

(R.L.Ne./G.S.O./D.C.P./Ed.)

THE SUN

The dominant body of the solar system, the Sun is the Earth's nearest star, eight light-minutes (149,598,000 kilometres) away from it. It is a typical star, approximately midway between the largest and the smallest and the brightest and the faintest stars known. On the basis of its temperature and radius (see below) it is called a G2 type dwarf star.

Energy source

The Sun's energy comes from the conversion of hydrogen to helium at a temperature of about 15,000,000 K, deep in its core where pressures are many thousand million (10^9) times that of the Earth's atmosphere at sea level and the density is 150 times that of water. The energy released in the core by nuclear processes slowly works its way to the surface, where it is radiated into space. The enormous energy output, of which the Earth intercepts only one part in 2,200,000,000, has been sufficiently constant over geologic time spans of hundreds of millions of years for complex life-forms, governed by cycles of day and night, by seasons, or by tides, to develop.

The visible surface of the Sun is the photosphere (Figure 1, top). Above it a layer 5,000 kilometres thick constitutes the inner atmosphere, or chromosphere, while above this lies the exceedingly tenuous high-temperature corona, which extends to the Earth and beyond. The Sun has a magnetic field; it rotates; it generates energy in its core. The outflow of energy from the deep interior through the surface and the surface manifestations of the magnetic field and solar rotation combine to make a complex and fascinating picture of solar phenomena that can be observed in great detail.

History of solar observations

For centuries the Sun was worshipped as a deity and was, therefore, not often the subject of physical study. Anaxagoras, the 5th-century-BC philosopher of Athens, supposed a great meteorite, which fell at Aegospotami in the daytime about 467 BC, to have come from the Sun, and he concluded that the Sun was a mass of red-hot iron larger than the Peloponnesus. With the invention of the telescope, Galileo Galilei, Johannes Fabricius, Christoph Scheiner, and Thomas Harriot almost simultaneously (1610–11) discovered sunspots (Figure 1), but it was the genius of Galileo that recognized their true nature as solar phenomena. Two centuries later, a German amateur astronomer, Samuel Heinrich Schwabe (1843), having diligently observed the Sun for 33 years, announced that the average number of spots varies cyclically with a period of about 10 years. In 1852 the cyclic period was found to be 11.2 years and the possibility of there being still another period of 80 years was recognized. In 1858 it was noted that the sunspots of a new cycle appear at latitudes of ±30° and, as the cycle progresses, the spots break out closer to the Sun's equator so that the last spots of the cycle appear at about ±8°.

Establishment of the first magnetic observatory

A German mathematician and astronomer, Carl Friedrich Gauss, established the first magnetic observatory in 1834 at Göttingen. Others soon followed. Small daily magnetic changes were shown to be related to the sunspot cycle in 1857. In 1904 intense magnetic storms were found to be associated with the passage of large sunspot groups over the central meridian.

Many important advances in solar astronomy have resulted from the construction of new telescopes and ancillary instrumentation of great sophistication. The physical and chemical nature of sunspots became known only after the development of the spectroscope. In 1870 dark bands were discovered in the spectrum of light radiated from sunspots, indicating the presence of molecular compounds; furthermore, it was postulated that the widening and splitting of some lines and the complete reversal of some dark lines to bright ones were caused by gaseous eruptions in the neighbourhood of sunspots—now called flares. The outflow of gases from the cooler dark centre of the spot (the umbra) to the edge of the penumbra, or outer part, at a speed of two kilometres per second was discovered spectroscopically in 1909. A reverse flow, inward, toward the centre of the spot, for the gases at the highest levels of the atmosphere above it, was found in 1913. The invention of the spectroheliograph by the U.S. astronomer George Ellery Hale in 1891 made possible the examination of the Sun in hydrogen light, calcium light, and the light radiated by ions or atoms of other elements, bringing out a wealth of detail in the chromosphere. In 1908 Hale looked for splitting and polarization of some atomic lines (the Zeeman effect) in spot spectra, but it required the large solar tower telescopes and pit spectrograph designed and constructed by Hale at Mount Wilson Observatory, California, to prove the magnetic character of sunspots. Between 1914 and 1924 Hale developed the laws of alternating sunspot polarity (see below). Although he also attempted to measure the Sun's general magnetic field, reliable measures were achieved only after the invention of the photoelectric magnetograph in 1948.

Table 2: Data for the Sun

Distance	149,600,000 km 92,957,000 mi
Mean angular diameter	31′59.3″
Diameter	1,392,000 km (864,950 mi) (109.3 times the Earth's diameter)
Surface area	6.087 × 10^{12} km^2 (2.35 × 10^{12} sq mi) (12,000 times the Earth's area)
Volume	1.412 × 10^{18} km^3 (3.387 × 10^{17} cu mi) (1,306,000 times the Earth's volume)
Mass	1.99 × 10^{30} kilograms (2.19 × 10^{27} tons) (333,400 times the Earth's mass)
Mean density	1.41, *i.e.*, less than 1½ times that of water
Escape velocity to infinity from the Sun's surface	618 km per sec (384 mi per second)
Effective temperature	5,740° K
Total energy output	3.86 × 10^{33} ergs per second
Radiation from the Sun's surface	6.35 × 10^{10} ergs/cm^2/sec
Brightness at the centre of the disk	8.23 × 10^5 lamberts
Candlepower	3.17 × 10^{27} candles
Light flux	141,400 lux, at the Earth's mean distance, outside the atmosphere 103,000 lux, the average illumination from the Sun at the zenith
Period of rotation	
Equator	26.9 days (synodical)
Sunspot zone (16° N)	27.3 days (synodical)
Pole	31.1 days (synodical)
The Sun's interior	
Density	150 gm/cm^3
Temperature	15,000,000° K*
Pressure	4 × 10^{17} dynes/cm^2 (4 × 10^{11} atmospheres)
The Sun as a star	
Spectral type	G2
Colour index	Mpg-Mpv = +0.45
Visual magnitude	Mpv = −26.86
Absolute visual magnitude	Mpv = +4.71
Velocity in space relative to nearby stars	19.4 km per sec in the direction of Lyra (Vega). The motion of the local stars of which the Sun is a member is about the galactic centre with a velocity of 220 km per sec

*Different calculations give results from 13,000,000°–25,000,000°.

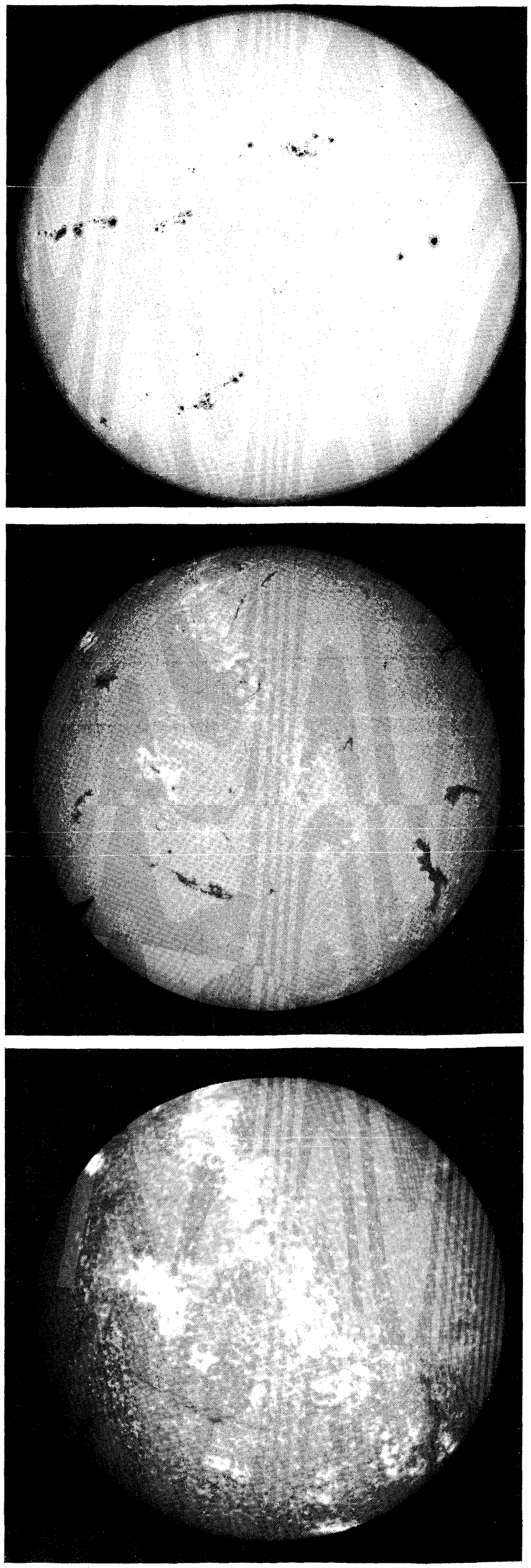


Figure 1: *Photographic studies of the Sun.*
(Top) Taken in ordinary light at sunspot maximum, December 21, 1957. (Centre) Taken in red hydrogen light (H-alpha), September 15, 1949. (Bottom) Calcium spectroheliogram taken September 15, 1949.
By courtesy of Mount Wilson and Palomar Observatories

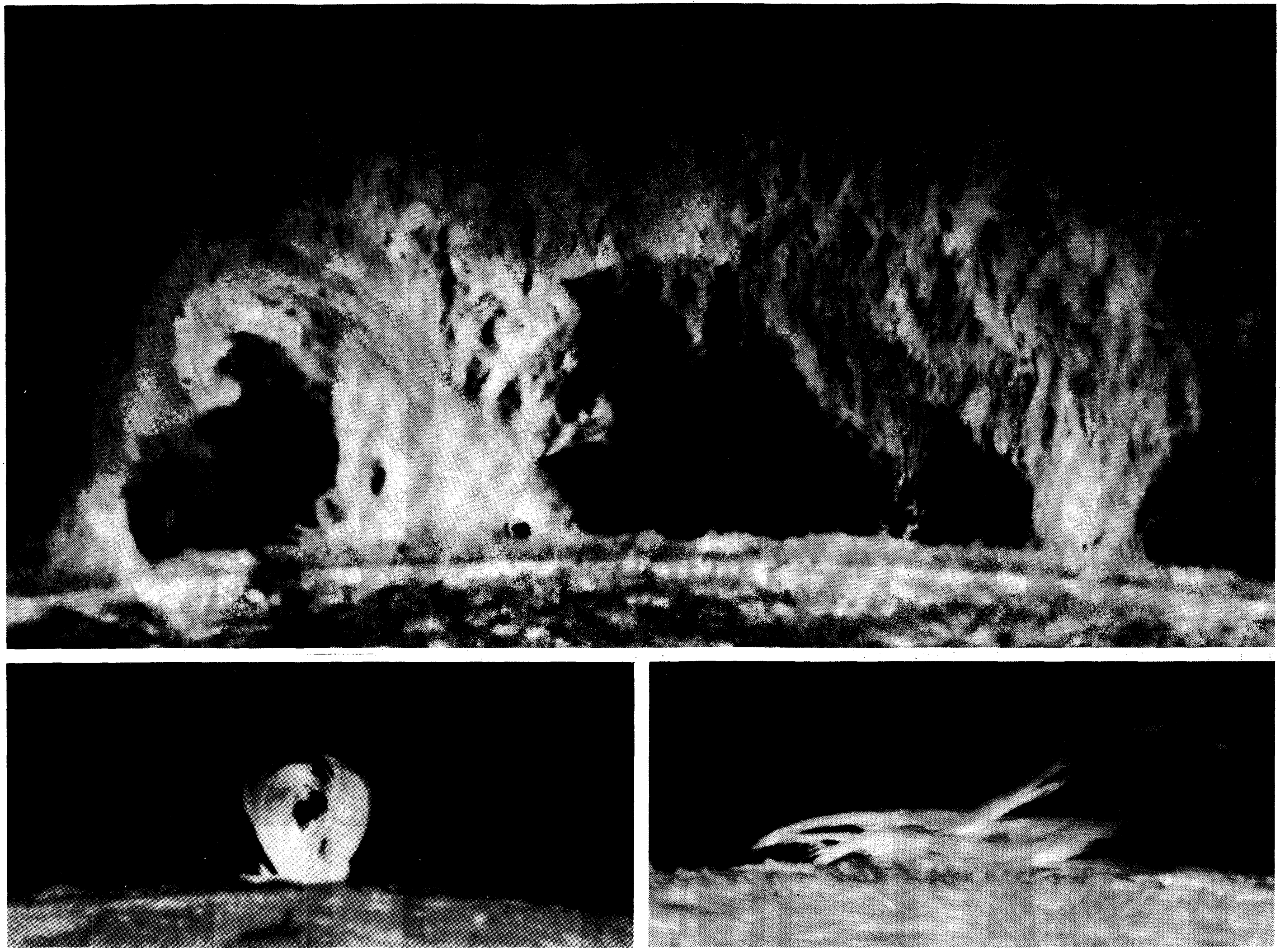


Figure 2: Solar prominences showing streamers, arches, loops, and other phenomena in the light of hydrogen (H-alpha). (Top) March 11, 1971. (Left) August 3, 1970. (Right) September 23, 1971.

By courtesy of Mount Wilson and Palomar Observatories

Joseph von Fraunhofer, a German physicist with greatly improved spectroscopes of his own construction, had visually found and mapped many (574) dark lines in the solar spectrum; to the strongest of these he assigned letters—*e.g.,* A, a, B, C, D—that are still used to identify the lines. The physical significance of the dark lines was explained by another German physicist, Gustav Robert Kirchhoff, in 1859. Kirchhoff conceived the idea of a hot Sun surrounded by cooler layers of vapour, which he called the reversing layers, in which the dark lines were supposed to have been formed. By comparison of laboratory spectra with the solar spectrum, the presence of eight elements in the Sun was later established. Henry Augustus Rowland, a U.S. physicist, published in 1897 a 12-metre photographic map of the solar spectrum that was of the highest quality and that allowed for the presence of 39 chemical elements in the Sun. Rowland's map extended to the ultraviolet limit caused by terrestrial ozone. This limit has only recently been surpassed by observations from space; now the spectrum has been mapped from the visible to less than one angstrom (10^{-10} metre). Toward red, the solar spectrum was recorded to 53,000 angstroms in 1888. Today, with new detectors, the limit is set only by terrestrial water vapour bands that obscure the infrared spectrum beyond 25 micrometres (25×10^{-6} metre). The complete spectrum can be observed only from very high altitudes, above the water-vapour absorption and the ozone absorption.

Attempts to measure the Sun's temperature. The determination of the Sun's temperature has been one of the most difficult problems in solar astronomy. The English astronomer Sir John Herschel at the Cape of Good Hope and Claude-Servais-Mathias Pouillet in France (1837) observed that the vertical rays of the Sun, if totally absorbed, would raise the temperature of a layer of water 1.8 centimetres deep by 1° C per minute. The measurement is in principle easy, but atmospheric absorption has been the uncertain factor. The problem was solved in 1881 and again in 1904, when it was shown that extrapolations must be made for solar radiation of different wavelengths, to determine what the intensity of each would be if there were no intervening air; and the values for all wavelengths must then be summed. A value of 1.96 calories per square centimetre per minute was obtained for the solar constant (the flux of total radiation received outside the Earth's atmosphere per unit area at mean Earth–Sun distance), that only recently has been improved by observations from balloons and high-flying aircraft. Early estimates of the Sun's temperature ranged from 1,461° C to 10,000,000° C. Transforming the solar constant into the energy radiated by the Sun and thence into its effective temperature was impossible until the law that connects the temperature of a surface with the amount of heat radiated per second—the Stefan-Boltzmann law—was established about 1880.

The Moon, with its opaque disk of almost exactly the same apparent size as the Sun, and with a sharply defined edge, has been valuable to the solar astronomer at times of solar eclipse. In the Middle Ages, prominences on the Sun (see Figure 2) that became visible during eclipses were thought to be part of the Moon's atmosphere, mirages, clouds, or holes in the Moon. During an eclipse of

July 1860, which crossed Spain, prominences were successfully photographed, and, by observing the motion of the Moon across them, it was demonstrated that these were solar. During an eclipse of August 1868, observers in India and in England, struck by the brilliance of the prominences, independently showed that they might be seen in broad daylight simply by opening the slit of a spectroscope and aiming it just outside the brilliant disk of the Sun's photosphere. The invention (1930) of the coronagraph made it possible for the first time to study the corona without an eclipse from high-altitude stations and on a nearly continuous basis. With a modified instrument, the K-coronameter (1950), which uses photoelectric detection, it became possible to view the corona through hazy skies at sea level, and measurements at very clear sites were greatly extended, to large distances from the limb of the Sun.

Coronal spectrum

The mysterious coronal spectrum—at first attributed to the mythical element coronium—was discovered in 1941 to be made up of the so-called forbidden lines of highly ionized atoms of iron, nickel, and other elements. The lines are described as forbidden because they are almost impossible to produce in the laboratory.

Detection of radio waves. Radio waves from the Sun were first detected in 1942 on British radar sets. It was suggested that the bursts of radio energy were associated with an active sunspot visible on the surface of the Sun and, in particular, with the large solar flare of February 28. In the same year, a weak background radio emission, or noise, was detected from the "quiet" Sun; *i.e.,* the Sun observed when sunspots and related phenomena are at a minimum. The science of radio astronomy has since become a large and important field of solar research.

The modern concept of energy generation in the Sun's interior follows from Sir Arthur Stanley Eddington's suggestion that only through nuclear processes could sufficient energy be produced to have sustained the output of energy from the Sun throughout the known lifetime of the solar system. More details were provided in 1937 and 1938 when the energy production from nuclear reactions, capable of transforming hydrogen to helium, were calculated. Later experimental and theoretical work has given improved values for all the reactions considered possible for energy production.

The quiet Sun

PHOTOSPHERE

The visible luminous surface of the Sun is the photosphere or light sphere. When carefully examined through a telescope, it is seen to have a mottled appearance called granulation (Figure 3); the most conspicuous objects to be seen, however, are sunspots which form and dissipate and partake in the solar rotation. The photosphere radiates most of the Sun's energy; it comprises a thin but rather opaque layer approximately 400 kilometres thick, the temperature of which varies from 10,000 K at the bottom to 4,200 K at the top. A typical density in the photosphere is around 10^{-3} gram per cubic centimetre, about $^{1}/_{1,000}$ that of air at sea level.

Granulation. About 4,000,000 granules, each one the top of a hot cloud of gas, cover the surface of the Sun. Because of their small size—300 to 1,450 kilometres—they can be seen and photographed from the ground only when the Earth's atmosphere is steady and free of turbulence. Excellent photographs have been made with telescopes carried by balloons to heights of 20 kilometres.

Characteristics of a granule

The best of these show that granules are irregular, often polygonal, and that each is surrounded by a narrow, relatively dark, lane. They last about seven to 10 minutes. A typical bright granule starts as a small grain the diameter of which grows to about 1,500 kilometres before it divides into several parts that spread and fade. Each granule rises with a speed of about 0.5 kilometre per second; as the energy radiates, the gas cools and descends rapidly in the dark intergranular lanes. A remarkable five-minute vertical oscillating motion of the granules was discovered in 1962. It has also been observed that after a bright granule forms, its upward movement sets the chromosphere into oscillation, which then continues for three to five periods after the disappearance of the exciting granule.

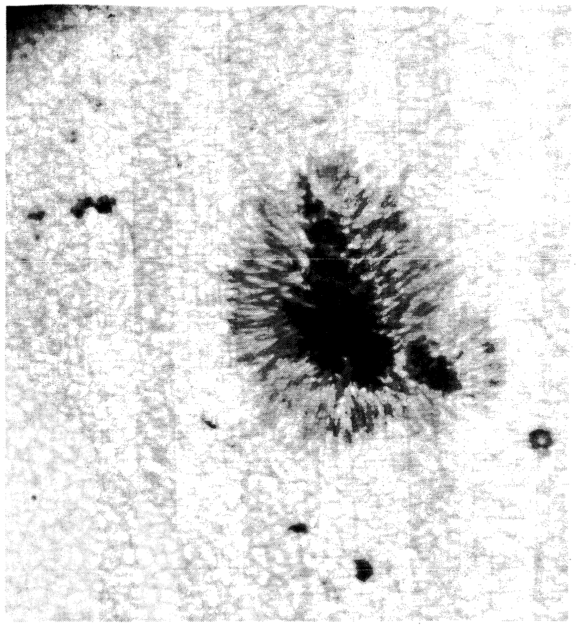


Figure 3: Granulation on the Sun showing a sunspot (centre right), which appears dark by contrast with the hotter regions surrounding it.

By courtesy of Project Stratoscope of Princeton University, sponsored by National Aeronautics and Space Administration, Office of Naval Research, and the National Science Foundation

Limb darkening. The limb of the Sun (the visible edge) appears darker than the centre. The direct view downwards at the centre of the solar disk penetrates, on the average, into deeper, hotter layers of the photosphere; at the limb the higher, cooler layers are seen. Limb darkening (which exists for all stars) is due solely to a temperature gradient in the atmosphere, the deeper layers being hotter. Conversely, measures of the limb darkening lead to a determination of the temperature distribution with depth.

The Sun's temperature. The Sun emits energy over all wavelengths, from X-rays to radio waves. Approximately 40 percent of the emitted energy is in the visible portion of the spectrum, 50 percent is in the infrared, and the remainder is in the ultraviolet. Radiation escaping from the Sun's surface into space comes from different depths of the Sun's atmosphere, each with a different temperature. The boundary temperature is about 4,200 K, but the eye can see to depths where the temperature is 10,000 K; the Sun has, however, a well-defined effective temperature, namely, that of an equal sphere radiating a like amount of total energy in accordance with physical (blackbody) radiation laws. From the total energy radiated a temperature of about 5,700 K is found. Different (colour) temperatures are derived from different spectral regions. A colour temperature of 6,000 K fits the whole visible spectrum reasonably well.

Solar energy

It is not easy to describe the enormous output of solar energy. If the Sun were enveloped in a shell of ice 12 metres thick, it would thaw its way out in one minute of time. If an ice bridge three kilometres in diameter could be built, spanning the immense distance of 149,600,000 kilometres from Sun to Earth, and, if by some means, the whole of the Sun's radiation could be concentrated upon this ice bridge, in one second it would be water and in seven seconds more it would be dissipated into vapour. Though the energy received at the Earth's surface amounts to over 1,500,000 horsepower per square kilometre, to harness it effectively is difficult. To achieve high temperatures requires concentrators or large parabolic collectors, such as those used in solar furnaces.

The Sun's rotation and sphericity. Galileo gave the first proof of the Sun's rotation and shape. He noted the Sun's roundness and saw that the apparent motion of sunspots in parallel lines with diminished movement at the limbs could be explained if the spots were on the surface of a sphere rotating from east to west. Modern spectrographic observations of the Doppler shift (a change of wavelength with motion toward or away from the observer) of Fraunhofer lines show that the equator is rotating faster than the poles. Unlike some stars, which rotate rapidly and are strongly flattened oblate spheroids, the Sun is a slow rotator and is so nearly spherical that it is uncertain whether any flattening can be detected. A polar diameter 70 kilometres less than the equatorial diameter has been suggested by some observations. The ellipticity of the Sun has been of interest to scientists because the gravitational field of a nonspherical body affects planetary orbits in a way different from that of a sphere. It has been suggested that part of the perihelion advance (a gradual rotation of the point closest to the Sun, which is otherwise very satisfactorily explained by the theory of relativity) of Mercury's orbit could be explained in this manner.

The Fraunhofer spectrum. The solar spectrum, within the wavelength interval 2,950–10,000 angstroms, shows about 25,000 so-called Fraunhofer dark lines (see above *History of solar observations* and Figure 1, centre). The identification of elements in the Sun is still carried out by the classical method comparing the solar lines with lines observed in the laboratory for both wavelength and intensity. Care must be taken to distinguish between solar lines and lines produced by water vapour, oxygen, and other constituents of the Earth's atmosphere.

Element identification

About 73 percent of the lines have been identified; 63 elements and 11 molecules are recorded. Iron contributes the greatest number (5,458) of identified lines. There are 1,453 lines from chromium, 1,344 from titanium, 856 from nickel, 388 from zirconium, 1,572 from solar cyanogen, and 1,477 from terrestrial water. For many elements, however, only a few lines are observed, for example: two from beryllium, two from silver, six from platinum, one from gold. The majority of the unidentified lines are faint and probably due to molecules and to the faint lines of the abundant elements in the Sun.

Modern techniques of spectrum analysis make use of the most powerful spectrographs that can be built. This makes it possible to study each spectrum line in detail for position, shape, depth and width.

A Fraunhofer line has considerable width because of the turbulent motions of the solar gases and collisions caused by neighbouring atoms. The resulting line profile, measuring the intensity against wavelength (or frequency), is that of an inverted bell-shaped curve near the line centre with broad, shallow wings, which, for strong lines, may extend many angstroms from the line centre.

The Fraunhofer spectrum has been extended to include lines in the far ultraviolet from rocket observations above the Earth's atmosphere. Within the far ultraviolet region, in the interval 3,000–2,097 angstroms, 7,146 lines are observed. Below 2,100 angstroms, the intensity of the spectrum drops and by 1,500 angstroms the radiation comes from the chromosphere and the absorption spectrum has turned into an emission spectrum. This emission spectrum contains the strong resonance lines of the elements first identified from weaker subordinate lines in the visible spectrum. Also, higher stages of ionization for a given element are observed in the far ultraviolet emission spectrum.

The Sun's infrared spectrum is dissected by numerous terrestrial water vapour absorption bands. Thus, at ground level the solar spectrum must be observed in the "windows" between the bands. To observe the entire infrared spectrum of the Sun, it is necessary to use a vehicle such as a high-flying aircraft, a stratospheric balloon, or a space vehicle that can carry the infrared spectrometer above all, or most of, the Earth's absorbing atmosphere. In the wavelength interval 11,984–25,578 angstroms, 1,786 solar lines and 6,911 telluric lines—*i.e.,* lines caused by substances in Earth's atmosphere—have been listed. Silicon, iron, carbon, carbon monoxide, and titanium contribute most of the solar lines, while water vapour, carbon dioxide, methane, and isotopic bands of these molecules make up most of the telluric lines.

The chemical composition of the Sun. The proportions (abundances) of the elements in the Sun can be determined directly only for the surface layers, the so-called solar atmosphere. In the interior, the abundances of the elements may be different because the elements are transformed there at different rates by nuclear reactions. The chemical composition of the solar atmosphere is thought to be representative of the composition of the Sun at the time of its formation, before such alterations occurred. This initial composition is relevant to the study of the evolution of the Sun and the stars. The chemical composition of the solar atmosphere can be determined in various ways, which depend on the types of observation available.

Solar cosmic rays are emitted from active regions of the Sun. Abundances of the following light elements have been derived from analysis of solar cosmic rays: helium, beryllium jointly with boron, carbon, nitrogen, oxygen, fluorine, and neon. Since the heavier elements are difficult to separate by nuclear counting methods, their abundances are generally given for groups, such as phosphorus to scandium and titanium to nickel.

Coronal abundances

Abundances derived from observations of the solar emission lines that originate in the higher chromosphere and the corona are called coronal abundances. The coronal-emission spectrum may be analyzed in two different ways, depending on whether the observations were secured on the disk or above the limb of the Sun. The far-ultraviolet-emission lines observed on the disk of the Sun are produced by atoms in many different stages of ionization. Analyses have been carried out for the more abundant atoms, such as helium, carbon, nitrogen, oxygen, neon, sodium, magnesium, aluminum, silicon, phosphorus, sulfur, calcium, and iron. The emission lines in the interval 3,000–10,000 angstroms seen above the solar limb are a result of "forbidden" electron transitions within atoms; *i.e.,* transition almost impossible to obtain on Earth. Since the forbidden emission lines originate in the corona, and are thus extremely weak compared to the radiation from the body of the Sun, they can only be observed at the time of a total solar eclipse or by means of a coronagraph. The known forbidden emission lines are due to argon and calcium, and to atoms of the iron group, such as chromium, manganese, iron, cobalt, and nickel, all of them being highly ionized.

The most commonly used method of determining the abundance of elements in the solar atmosphere makes use of the solar absorption (Fraunhofer) spectrum, which extends from the near ultraviolet, through the visible, into the infrared wavelength regions. The majority of Fraunhofer lines are formed in the lowest layers of the solar atmosphere, that is, the photosphere and low chromosphere. The abundances derived from these lines are called photospheric abundances.

By courtesy of Kitt Peak National Observatory, Tucson, Arizona

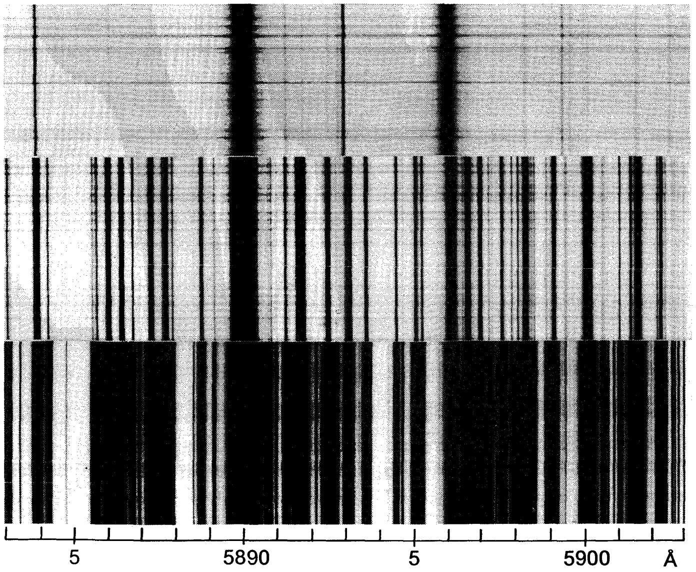


Figure 4: Sodium D lines observed at high Sun (top spectrum) and, through a greater amount of air, at low Sun (middle spectrum). The bottom spectrum shows water vapour bands for comparison. The strengthening of terrestrial water vapour lines at low Sun is most apparent.

Photospheric abundances

The determination of photospheric abundances is based on a measurement of the intensity of the lines. With modern instruments a solar spectrum line can be recorded at very high resolution (Figure 4) and, consequently, the intensity distribution in the line (the line profile) can be obtained with very good precision. The line profile and the measurement of the total intensity absorbed by each line depend not only on the abundance and physical characteristics of the particular atom (or molecule) producing the line but also on the physical conditions (temperature, density) that prevail in the layers of line formation. Thus, in an analysis of the solar spectrum, both the physical state of the solar atmosphere and its chemical composition are determined at once.

In Table 3 the results of abundance determinations for the photospheric and the coronal abundances, both as determined in 1968, are shown. For the elements lithium, beryllium, boron, iron, lead, bismuth, thorium and uranium, more recent and more accurate photospheric abundance results are given here. The differences between photospheric and coronal abundances of certain elements are essentially due to the uncertainties in the abundance determinations, both from the Fraunhofer spectrum and the emission spectrum.

From Table 3 it can be seen that only a minute portion of the solar gas is made up of elements heavier than helium, that about 15 percent is helium, and that the bulk of the solar gas is constituted of hydrogen.

Table 3: Solar Abundances
(log Σ relative to log Σ = 12 for hydrogen)

Z atomic number	element	log Σ photosphere	log Σ corona (forbidden lines)	log Σ corona (UV lines)
1	H	12.00	12.00	12.00
2	He	—	—	11.2
3	Li	1.0	—	—
4	Be	1.2	—	—
5	B	< 2.5	—	—
6	C	8.5	—	8.8
7	N	8.1	—	7.8
8	O	8.8	—	8.7
9	F	4.6	—	—
10	Ne	—	—	7.8
11	Na	6.3	—	—
12	Mg	7.4	—	7.6
13	Ai	6.2	—	6.5
14	Si	7.2	—	7.8
15	P	5.3	—	—
16	S	7.3	7.0	7.3
18	Ar	—	7.3	—
19	K	4.7	5.8	—
20	Ca	6.0	6.8	—
21	Sc	2.9	—	—
22	Ti	4.8	—	—
23	V	4.2	—	—
24	Cr	5.0	6.0	—
25	Mn	4.9	5.8	—
26	Fe	7.6	7.8	7.8
27	Co	4.7	5.5	—
28	Ni	5.8	6.7	6.6
29	Cu	4.5	—	—
30	Zn	3.5	—	—
31	Ga	2.7	—	—
32	Ge	2.5	—	—
37	Rb	2.5	—	—
38	Sr	3.0	—	—
39	Y	3.2	—	—
40	Zr	2.7	—	—
41	Nb	2.3	—	—
42	Mo	2.3	—	—
44	Ru	1.8	—	—
45	Rh	1.4	—	—
46	Pd	1.6	—	—
47	Ag	0.8	—	—
48	Cd	1.5	—	—
49	In	1.5	—	—
50	Sn	1.5	—	—
51	Sb	1.9	—	—
56	Ba	2.1	—	—
57	La	2.0	—	—
58	Ce	1.8	—	—
59	Pr	1.5	—	—
60	Nd	1.9	—	—
62	Sm	1.6	—	—
63	Eu	1.0	—	—
64	Gd	1.1	—	—
66	Dy	1.0	—	—
70	Yb	1.5	—	—
82	Pb	1.8	—	—
83	Bi	< 0.8	—	—
90	Th	< 0.8	—	—
92	U	< 0.6	—	—

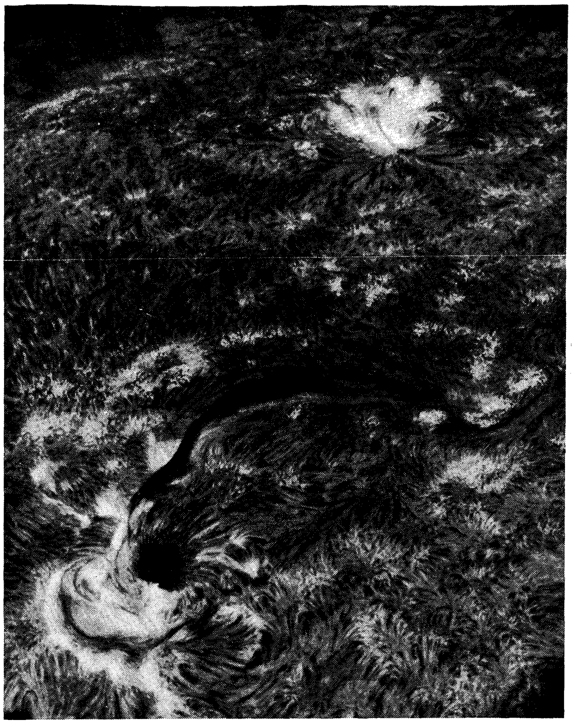


Figure 5: High-resolution photograph of sunspot in hydrogen (H-alpha) light shows bright plages and dark filamentary structures (prominences) on the Sun.

By courtesy of Lockheed Solar Observatory, Palo Alto, California

CHROMOSPHERE

Temperature range

The relatively transparent chromospheric layers form a complex transition zone about 5,000 kilometres thick between the low-temperature photosphere and the high-temperature corona. The chromosphere temperature increases outward from about 4,500 to 1,000,000 K, whereas the density decreases rather rapidly from about 10^{12} particles per cubic centimetre at the base to about 10^{9} particles per cubic centimetre at the top. The chromosphere is a region of great activity, visible at the time of a solar eclipse as a bright-coloured arc from which the "flash spectrum" originates. The low chromosphere is a relatively homogeneous layer of gas, about 1,500 kilometres thick, from which emerges, geyser-like, a multitude of upward-moving jets called spicules. At about 8,000 kilometres the spicules stand apart as separate entities and it is thought that the hot coronal gas reaches down between them. The chromospheric gas is quite transparent to continuous radiation but opaque to light in the cores of strong Fraunhofer lines. As a result, if the Sun is viewed in the monochromatic light of a Fraunhofer line, its substance is visible. By selecting lines of different strengths the astronomer can choose the depth of observation; thus, the chromospheric layers can be figuratively peeled like an onion. Through equipment capable of isolating a spectrum line a fraction of an angstrom in width, the Sun presents a picture of fine and coarse strcutures (Figure 5). Much of the fine structure visible in the cores of metal lines is due to Doppler displacements from ascending and descending volumes of gas. The periodic character of this motion has been discussed in the section on granulation above. Photographs that are taken in the light of one particular wavelength, called the hydrogen-alpha line, show a very coarse mottling, which near the sunspots becomes more or less organized into filamentary whorls.

Flash spectrum. During a total solar eclipse there is a moment when the photosphere is just hidden by the Moon, leaving exposed the chromospheric layers, which appear as a thin, orange-red crescent along the edge of the Moon. The light from this crescent is found to have

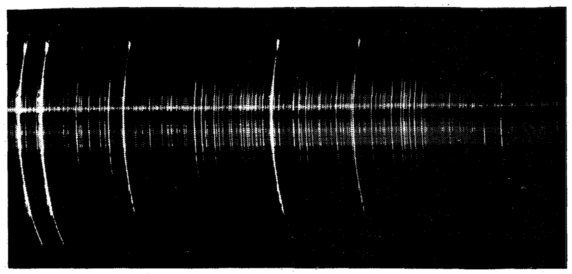


Figure 6: *Flash spectrum of the Sun in ultraviolet light.* The two bright spectral lines at left, showing prominence structure, are the H and K lines of ionized calcium. The bright line in the centre is a line of ionized titanium at 3,685Å. Frequency increases from left to right, wavelength from right to left. Obtained at the solar eclipse of February 1962 in New Guinea.

By courtesy of the Joint eclipse expedition: High Altitude Observatory, Joint Institute for Laboratory Astrophysics, and Air Force Cambridge Research Laboratories

a bright-line spectrum, which, for a few seconds, "flashes" in and out of view as the Moon moves across the Sun's atmosphere, hence the name "flash spectrum" (Figure 6). The brightest lines and those which extend to the greatest heights are lines of hydrogen, helium, and calcium; when examined in detail the spectrum is nearly, but not quite, identical with the Fraunhofer spectrum, the differences reflecting the different physical state of the photospheric and chromospheric layers. Lower pressures and higher temperatures favour greater ionization and excitation. Characteristic flash spectra show an enhancement of high-excitation lines and the appearance of a great number of lines from singly ionized iron, titanium, nickel, chromium, and the rare earths.

Characteristics

Spicules. A 19th-century Italian astronomer, Angelo Secchi, described the chromospheric layers viewed at the limb as like a burning prairie. He had seen through his spectroscope a forest of superimposed jets, "the spicules," each lasting only a few minutes. Spicules appear as tiny, bright surges, less than 500 kilometres in diameter, moving upward with velocities of 20–30 kilometres per second to heights of 8,000 kilometres above the photospheric layers. They are often grouped in rosettes or brushlike shapes. They number about 100,000 and have an average lifetime of five to 15 minutes. On the basis of spectroscopic and radio studies, a spicule column is believed to have temperatures of 10,000 K in the core and 50,000 K on the surface. The denser core material emits lines of calcium and hydrogen and radio waves of millimetre and centimetre wavelengths. The hot sheath is the source of the far-ultraviolet emission of highly ionized atoms and of the lines of neutral and singly ionized helium.

Supergranulation. Spectroheliograms made, for example, with calcium light show a coarse pattern in addition to smaller mottles, which are due to a network of cells about 30,000 kilometres in diameter called supergranules. In these cells an expansion of material takes place from centre to edge with a speed of about 1/2 kilometre per second. These supergranules, which number about 5,000 at any time, have been identified as deep convection cells with a lifetime comparable to a day. It is observed that the flow of material from centre to edge tends to collect the majority of spicules on the cell boundaries and to sweep up the magnetic field domains so that the cellular pattern is visible on magnetic maps as well as on spectroheliograms.

CORONA

The corona is a luminous, high-temperature, rarefied gas envelope of the Sun. Close to the Sun it is about 1/1,000,000 as bright as the solar disk. Its brightness decreases rapidly with distance, so that at two solar radii the brightness is less than one percent of that of its innermost parts; hence, the feeble light is visible only at the time of a total eclipse or through a coronograph. The corona has a kinetic temperature of about 2,000,000 K out to several solar radii. Because of its high temperature it evaporates continuously, producing an outward flow of electrically charged particles called the solar wind, which extends beyond the vicinity of the Earth. The highly ionized coronal gas has a density of the order of 5×10^8 atoms per cubic centimetre at its base; the density decreases with height, diminishing by a factor of 2.7 approximately every 50,000 kilometres.

Form and structure. The corona owes its beauty to its form, which is made up of streamers, filaments and rays of pearly white light. Polar plumes, fanning out like iron filings about the poles of a magnet, suggest north and south magnetic poles for the Sun. At sunspot maximum the corona is globular in form with streamers in all directions, while the minimum phase is characterized by polar plumes and long equatorial streamers extending outward many solar radii.

The *F*-, *K*-, and *L*-coronas

Corona light has three components: the so-called *L*-, *K*-, and *F*-coronas. The *L*-corona, which contributes only about 1 percent of the coronal light, refers to the line spectrum emitted by the highly ionized atoms of the coronal gas, which in such a state is called a plasma. The light of the *K*-corona is caused by scattering of photospheric light by fast-moving free electrons near the Sun. The *K*-component is strongest in the inner corona where the electron density is highest. Since the electrons move at very great speeds (of the order of 6,000 kilometres per second), the photospheric Fraunhofer lines are here completely washed out by the scattering process and, consequently, the *K*-component reveals a continuous spectrum only. The *F*-component is caused by diffracted photospheric light from interplanetary dust particles. Its spectrum is that of the photosphere; *i.e.,* a continuous spectrum intersected by the Fraunhofer lines. The *F*-component is an extension of the zodiacal light into the region close to the Sun. The *L*-component and the *K*-component represent the true corona.

The corona may also be studied on radio heliograms, and on X-ray pictures made from above the Earth's atmosphere. The solar radio emission at wavelengths greater than one metre originates in the corona, and at the other end of the electromagnetic spectrum, any wavelength shorter than 200 angstroms can be emitted only by the corona. X-ray pictures show the regions of strongest coronal condensation. Comparison of observations over different parts of the electromagnetic spectrum show that the corona is very inhomogeneous and that its form, structure, and intensity change considerably with the level of activity on the surface of the Sun.

The quiet Sun emits radio waves. At metre wavelengths the emission comes from great heights where the electron density is less than 3×10^7 electrons per cubic centimetre—half a solar radius from the photosphere. At 10 metres, the corresponding height is 10 solar radii. Thus, the corona can be examined by the radio astronomer, layer by layer, in different radio wavelengths, just as the optical astronomer examines the chromosphere, layer by layer, with the spectroheliograph.

Physical properties. The most remarkable property of the corona is its high kinetic temperature, 1,500,000 to 2,500,000 K. This follows from many types of observation: (1) The coronal spectrum was shown in 1941 to be emitted by atoms of iron, nickel, calcium, and argon that had lost 10 to 14 electrons. Such a degree of ionization can come about only by collisional ionization by energetic electrons driven by high temperatures in a low-density gas. (2) The coronal emission lines are much broader than the Fraunhofer lines and if this is interpreted as Doppler broadening by random motion of the emitting atoms, a temperature of about 2,000,000 K is indicated. (3) The brightness of the Sun in radio wavelengths longer than 100 centimetres is about 1,000,000 K. (4) Finally, the density gradient of the electron gas has been shown to follow an exponential decrease with height and results in a temperature of 1,500,000 K.

It is generally believed that the energy source supplying the heating of the corona lies in the mass motions of granules and spicules in the photosphere and chromosphere. The interaction can take place by direct collision between the atoms of the spicule and those of the corona or by the transfer by shock waves (hydromagnetic, acoustic) of energy from below to the chromosphere and corona.

The solar wind. It was predicted in 1958 that the high temperature of the corona would cause some of it to boil off the Sun, and that the coronal particles would appear in the neighbourhood of the Earth as a solar wind moving with speeds of 300–1,000 kilometres per second. Though it had been suggested as early as 1951 that the behavior of certain comet tails could be explained by the interaction of a solar plasma from the Sun with the comet, not until satellites explored the space around the Earth was direct proof forthcoming. A space probe from the United States to Venus in 1962 established the wind's existence and showed that there is a steady stream of protons and electrons in the vicinity of the Earth with a density of about one to 10 particles per cubic centimetre and carrying with them an associated magnetic field ranging from two to 40 microgauss. Thus interplanetary space is filled with a fully ionized hot plasma of coronal origin. At solar minimum the average flow into space amounts to 10^{33} particles per second; this mass loss is thought to be supplied from below by the dissipation of spicules.

Particle density

A changing pattern of active regions covers the surface of the Sun and proceeding from each is a positive or negative magnetic field, which is carried by the plasma. Because of the Sun's rotation and the steady outflow of material, the magnetic field lines trace curves in space—Archimedean spirals in the equatorial plane.

The active Sun

The active Sun is characterized by the growth and decay of magnetic fields together with all the associated phenomena observed at many different levels of the solar atmosphere: sunspots with associated very bright areas called faculae, plages (other bright areas not necessarily associated with sunspots), flares, surges, prominences, coronal condensations, and the emission of radio, X-ray, and cosmic-ray bursts (see Figure 1 and below).

CENTRES OF ACTIVITY

A centre of activity is an area formed by the arrival from below of a localized strong magnetic field at the Sun's surface. Its description includes all of the transient interrelated phenomena listed above. A few centres of activity may live 200 days or more (eight to 10 solar rotations); most are short-lived, lasting one to two weeks.

Life history of a typical centre of activity

The life history of a typical centre of activity is given below. In the first few days a bipolar magnetic region (one showing both north and south magnetic polarity) develops, without spots, followed later by bright areas known as faculae and coronal extensions projecting from them. The coronagraph shows stable ray and arch structures about 150,000 kilometres long above the region. After two to three days the most important (spot) phase starts with the formation of a small spot and a rapid extension of the fine lineal elements—small prominences—in the neighbourhood of the spot. A second spot forms shortly afterward, followed by the formation of numerous small spots between the larger bipolar pair. A few flares and flare surges take place. After about ten days the group of spots reaches its maximum size. Each spot appears to be embedded in a large, complex penumbral region of converging filaments. Flare activity reaches its maximum along with the associated phenomena of solar radio bursts, enhanced X-ray emission, and, rarely, cosmic rays from superflares. By 27 days—one solar rotation—the activity has decayed and all spots except the preceding spot have disappeared. Flare activity is greatly reduced, but the bright facular areas (extremely bright areas associated with sunspots) continue to increase in size. Next, a stable prominence filament forms on the poleward side of the region, which, after the lapse of 50 days, cuts the facular region in half, separating regions with opposite magnetic polarity. The prominence filament now extends for 100,000 kilometres and is oriented more or less east and west, drawn out by the Sun's differential rotation. At about the 80th day, the facular area breaks up, but the filament continues to lengthen. After 100 days the faculae have disappeared completely and the prominence filament reaches its maximum length. In the succeeding days it shortens and drifts poleward where it still may have an identity after 150–250 days.

SUNSPOTS

Physical nature. In appearance a sunspot consists of two parts: a dark core, the umbra, surrounded by a filamentary border called the penumbra. Single spots comprise about 40 percent of the total. Normally, they are in complex groups exhibiting great variety of form and shape, often covered with brilliant "bridges," giving the impression of great turmoil. The smallest spots of a group are mere pores without penumbra, a few hundreds of kilometres in diameter, whereas the largest spots exceed by several times the diameter of the Earth. Large spot groups may extend over one-sixth of the diameter of the Sun and are easily visible to the naked eye at sunset when the Sun's intensity is much reduced. At the Sun's limb (visible edge) a large spot may appear as a slight depression in the Sun's surface.

Sunspots are dark only in contrast to the surrounding photosphere, for, in fact, they are brighter and hotter than molten tungsten. From their observed brightness a spot temperature of 3,800 K is obtained, as compared to 5,700 K for the photosphere. At spot temperatures, a number of chemical compounds are stable. The spectrum lines of many radicals and molecules, cyanogen, carbon monoxide and hydrides of nitrogen, carbon, oxygen, and magnesium, also known in the disk spectrum, strengthen greatly in the spot, indicating increased abundance; also lines of calcium, silicon and fluorine hydrides, and titanium and zirconium oxides show in the spectrum of a sunspot.

Magnetic property of sunspots

The predominant property of sunspots is their intense magnetic field. In general, the larger the spot, the stronger the magnetic field. A small spot typically has a field of about 500 gauss (a gauss is a unit of magnetic field strength), a large spot, 4,000 gauss. By comparison, Earth's natural magnetic field at the surface is somewhat less than one gauss, and the overall solar magnetic field is now believed to be about the same. In a simple spot the field is nearly axisymmetric; it has its maximum value at the centre of the umbra and decreases to a very small value at the outer edge of the penumbra. On the average, the lines of magnetic force stand vertical to the Sun's surface at the centre of the umbra but diverge outward so that at the edge of the penumbra the direction of the field has an inclination of about 70° to the vertical. An explanation of the relative coolness of sunspots must involve their associated magnetic fields. After the magnetic field is formed the region cools, presumably because convection, bringing energy from below—as exhibited by the surrounding granulation—is stopped. A balance is achieved when the gas pressure, the magnetic pressure, and the mass flow of material in and out of the spot are in equilibrium.

The sunspot cycle. That spots are not always equally numerous on the surface of the Sun is apparent to anyone who examines the Sun telescopically on every clear day. But it was not until 1851 that astronomers were able to find a regular periodicity of about 10 years. At sunspot minimum the Sun may remain spotless for weeks; at maximum several large groups containing dozens of spots are generally visible on its surface. Complete records of the Wolf sunspot numbers (named after the Swiss solar astronomer Rudolf Wolf, whose work on the behaviour of sunspots was fundamental) are known for the past 120 years. Monthly means have been established from 1749 and the epochs of maximum and minimum are known from the year 1610.

Sunspot numbers

Because of the wide fluctuations of the daily number of spots, the level of solar activity is generally expressed as a running mean of a month or a year. The average interval between maximum is 11.2 years but is quite variable, ranging from eight to 16 years. The ascent to maximum, which averages 5.2 years, is steeper than the decline. In addition to the 11-year cycle, a low-amplitude 80-year cycle is recognized, though observations have not yet extended over an interval long enough to define the long wave completely. Sunspot numbers show an 11-year period, but the true length of the cycle is 22 years, when account is taken of the alternation of the sign of spot magnetic fields. The leading spot of a bipolar group in the northern hemisphere has a magnetic polarity opposite that of the leader in the southern hemisphere; and the polarity of the leaders reverse at the beginning of each 11-

year cycle. Thus, the magnetic period of the Sun is twice the 11-year cycle.

The number of spots in each cycle is quite variable. Attempts have been made to fit the observed numbers by superposition of cycles, but, in general, these have failed to predict succeeding cycles. Others have tried to relate the sunspot cycle to planetary influence, particularly that of Jupiter, which revolves around the Sun in a period of 11.8 years. Apparently the true cause of sunspots is deeply seated in the Sun itself.

Each new cycle begins with the appearance of a few spots at about ±30° north and south latitude. Four years later in the cycle, at sunspot maximum, the zone of eruption is at ±15° latitude, while the last few spots of the cycle appear at ±8°. When the latitude of spots, without regard to size, is plotted as a function of time, a so-called butterfly diagram results. This shows that new spots of a new cycle can appear at high latitudes before the disappearance of low-latitude, old-cycle spots, and suggests each new cycle is in some ways independent of the old.

Magnetic field theories

Many of the known features of the sunspot cycle have been explained by a theory of magnetic-field amplification by solar differential rotation. Very much simplified, the theory postulates that the Sun's differential rotation slowly winds up the lines of magnetic force like string on a ball. These lines of force, called flux ropes, which are formed beneath the surface, become buoyant and tend to rise to the surface; thus, bipolar magnetic regions (sunspots) are formed where the flux rope loops through the surface. One worker has separated the field into three components and adopted an interior rotation faster than the exterior with a strong concentration of radial shear toward the equator. With suitable parameters, he was able to reproduce with remarkable exactitude all the phenomena connected with the solar cycle.

OTHER SOLAR PHENOMENA

Distribution

Faculae, or plages. Bright granulated structures, visible in white light, toward the Sun's limb, are called faculae. That they appear only near the limb and not at the centre of the disk means that the higher facular parts emit more radiation than the surrounding high photospheric layer and the lower parts emit less than the local photosphere. The temperature excess amounts to 250° C at the most.

Spectroheliograms of the active Sun reveal large, bright areas closely associated with the photospheric faculae. These chromospheric faculae, or plages, are the extension of the photospheric faculae into the chromosphere. Ultraviolet plages and X-ray plages are observed coming from the high chromosphere and corona. Radio plages are also observed in the centimetre and decimetre radio wavelengths.

Prominences, or filaments. Prominences are among the most beautiful of solar phenomena (see Figure 2). A typical fully developed quiescent prominence is a thin fish-tailed, or ribbon-like, gaseous structure standing above the Sun's surface; it may attain a size of 200,000 kilometres long, 40,000 kilometres high, and 6,000 kilometres thick. As seen on the disk, in hydrogen-alpha spectroheliograms, prominences appear as long, dark structures called filaments. At the Sun's limb projected against the dark sky they are seen as bright clouds composed of a myriad of fine threads. Generally, prominences are connected by appendages to the Sun's surface or rest upon it, but a few stand free. Like clouds, prominences take many fanciful forms. From their appearance and mode of formation they have been classified into the following types: active prominences, eruptive, sunspot, tornado, quiescent, and coronal prominences.

Time-lapse motion pictures reveal the intricate motions of prominences. Thus viewed, they show at least two active groups of the first type: prominences that appear as great funnels with downward-curving streamers feeding knots of material from the prominence into a nearby active area; and the spectacular arched prominences in which material condenses from the corona and moves down along each arm, often for several days, into a spot region. Quiescent hedgerow-type prominences are carried in, and are supported against the pull of gravity by, a trough of magnetic field lines; they separate regions of opposite magnetic polarity on the Sun's surface and may last for months. The large-scale structure of such quiescent prominences is fairly constant, but the fine structure shows much motion, generally downward. On occasion, a nearby solar flare may trigger great activity in the prominence such that it either flows downward into an active region and disappears, or that a portion of it may erupt from the surface of the Sun with velocities of 200–500 kilometres per second. Strangely, the activity appears not to destroy the magnetic-field configuration, for there is a tendency for the prominence to reappear a few days later, phoenix-like, in its former place and form.

Spectroscopy reveals the physical conditions in prominences. There are both hot and cold prominences. Cold prominences with kinetic temperatures of 8,000 to 10,000 K show sharp spectral lines: the Balmer lines of hydrogen, H and K of calcium, and a number of lines of ionized and neutral metals. Active prominences, loops and surges, with temperatures up to 100,000 K show lines broadened by thermal motions of the atoms; the metal lines are weak and ionized helium lines appear.

Flares. In plage areas, sudden and distinct brightenings occur from time to time. These brightenings, called flares, last from a few minutes for small ones, up to several hours for very large ones. Solar flares are among the most violent and dramatic events that occur on the Sun, yet, with few exceptions, they cannot be seen in white light and may well go unnoticed. They are easily seen, however, in the red light of hydrogen and they produce a number of effects observable at the Earth that make them of considerable importance.

Flares are classified by area and brightness on a scale of increasing importance, from −1 to +3, the latter group comprising only 1 percent of the total. As seen in hydrogen light, a typical large flare starts as a brightening of a pre-existing filamentary structure in the neighbourhood of a sunspot. The rise to maximum brightness is sudden, taking place in two to five minutes, with a rapid spreading of the flare area, the brightest parts of which are often elongated. The flare lasts about one hour longer during which a bifurcation and separation of the luminous ribbon and subsequent fading of the flare take place. Frequently, after the onset of a flare, a surge of gas is ejected with a velocity of perhaps 100 kilometres per second. A great flare can create an expanding magnetohydrodynamic wave, a complex effect involving both magnetic and hydrodynamic forces, propagating through the corona. Its passage is seen by the successive activation of disk filaments as it spreads over the surface with velocities of 1,000 kilometres per second and for distances of over 500,000 kilometres.

Geophysical phenomena

Many geophysical phenomena are caused by radiation and particle emission from flares. For example, simultaneous with the observed flare there is often a fade-out of shortwave-radio communication caused by a flash X-ray penetrating to the ionosphere of the Earth and greatly increasing its ionization and absorption. Plasma clouds from the flare region moving out at 2,000 kilometres per second produce auroras and geomagnetic storms a day or so later. Cosmic-ray particles from very large flares have been observed to penetrate to ground level. In the polar regions, measurements of cosmic-radio noise show a marked decrease, caused by bombardment of the upper atmosphere by high-energy protons from the flare.

Observations of the X-ray spectrum during a flare detect a great increase in the emission at the shorter X-ray wavelengths down to below 10 angstroms. This reveals the existence of localized regions with temperatures of the order of 25,000,000 K and more. The formation of such regions is explained as being due to thermal electrons heated by accelerated electrons, which, in turn, are responsible for the microwave emission connected with flares. Usually the enhanced X-ray emission caused by the flare phenomenon appears after the emission at radio frequencies, but the X-ray emission lasts much longer and decays more slowly than the microwave emissions.

In addition to the above-mentioned flare-associated X-ray emission (so-called quasi-thermal due to the heating of localized regions), bursts of very short wavelength (hard)

X-ray and radio phenomena

nonthermal X-rays may occur which are observed in the range between about one and 10 angstroms. They may last from a couple of seconds up to two minutes. Since the energies of these bursts are of the order of 10^4–10^6 electron volts, some of their electrons must be moving with velocities of between 60,000 to about 300,000 kilometres per second. Consequently, the bursts are felt to be due to bremsstrahlung radiation (that emitted by decelerating electrons) or to synchrotron radiation, produced by electrons accelerating in a curved path. The hard X-ray bursts are correlated with various types of radio bursts which also are produced by the acceleration of particles to high speeds. There is, however, not always a one-to-one correspondence between X-ray and radio emission in a flare or any other flare phenomenon.

The interaction of high-energy electrons, protons, and nuclei in the solar atmosphere may also produce gamma rays in a flare region. A few major flares have been observed to eject gamma-ray bursts with energies up to one million electron volts (*i.e.,* the energy that would be gained by an electron in moving through a potential energy difference of 1,000,000 volts). Like X-ray bursts, gamma-ray bursts are short-lived phenomena; their duration, being of the order of a few seconds to a minute, depends on the time the high-energy particles are trapped in the solar plasma.

The radio phenomena associated with a large flare are numerous and complex; different events appear at different frequencies and different times. During the flash phase of a large flare, a stream of high-energy electrons moving with one-third of the speed of light are ejected from the flare into the corona to generate waves moving at speeds faster than the speed of light in the plasma called Cherenkov plasma waves and fast-drift radio bursts, that is, bursts of radiation at rapidly changing radio wavelengths. These are observed in the frequency range 600 to 20 megacycles per second. In five to 10 minutes other fast electrons are trapped in the magnetic field and their oscillation, possibly between magnetic-mirror points, produces a broadband burst over areas of about one-quarter of the Sun's disk. Electrons accelerated downward from the ejection area of the flare plunge into the chromosphere creating X-rays and microwave bursts. In addition to the high-speed electrons, the flare explosively ejects a dense plasma cloud at a velocity of 1,000 kilometres per second. This cloud may interact with the corona at supersonic velocities to form shock waves and radio bursts. The magnetic field embedded in this plasma moving out from the Sun holds electrons moving at a high fraction of the speed of light, which emit synchrotron radiation that can be followed with radio telescopes as a moving burst. From deep in the flare site microwave emissions continue, generated by synchrotron and bremsstrahlung processes (see above). Continuous radio emission continues for hours at metre wavelengths.

Energy released by a flare

The physical causes of flares are still unknown, though it is believed that the energy released by a flare (about 10^{32} ergs in a large flare) must come from the intense magnetic or electric fields associated with the solar active regions. It has been estimated that the energy released by a single flare would be enough to provide electric power for the whole world for 1,000,000 to 100,000,000 years at the current rate of use, depending upon the size of the flare.

Coronal condensations. The corona is far from uniform. Coronal condensations occur over centres of activity. They dominate the scene in X-ray photographs of the Sun and can be observed with radio telescopes. The condensations are shaped in the forms of rays, brushes, loops, or arches, and are often associated with hot prominences. They vary rapidly in brightness (a factor of 2 in two hours is not uncommon) but may remain stable in form for quite some time; in fact, over a large, highly developed spot group or centre of activity, a condensation may remain visible for several months. At the time of a flare the structure of a condensation may suddenly undergo change and show great activity; arch streamers, for example, may expand outward with velocities of 500 kilometres per second. In a coronal condensation, the electron temperature approaches 5,000,000 K and the particle density is five to 10 times greater than that of the surrounding quiet corona. At radio frequencies coronal condensations manifest themselves by enhancement of the radio emission in centimetre and decimetre wavelengths. By far, the greatest enhancement, about 80 times greater than in the quiet corona, can be observed in the X-ray region, in which the emission is proportional to the square of the electron density.

The Sun's general magnetic field. The existence of a weak polar magnetic field for the Sun has long been assumed. The measurement of small magnetic fields by the Zeeman effect is exceedingly difficult; hence, it need not be surprising that, although early measurements of polar photospheric fields gave values of 20–50 gauss, modern measurements have indicated much smaller values, of the order of one gauss, roughly the same as the natural magnetic field at the surface of the Earth. How the observed photospheric field is related to the internal field is not known. The field observed in the photosphere in the polar regions is not the dipole field but results from the slow migration and expansion of magnetic remanents from centres of activity in the sunspot zones to the polar regions. Furthermore, the irregular or random eruption of magnetic flux in the spot zones is eventually reflected in the polar fields and explains why the Sun has, on occasion, had a magnetic south pole simultaneously at both north and south poles and why there is often a lack of perfect symmetry between the two hemispheres. Figure 7 shows a continuous plot of the photospheric magnetic field.

By courtesy of Kitt Peak National Observatory, Tucson, Arizona

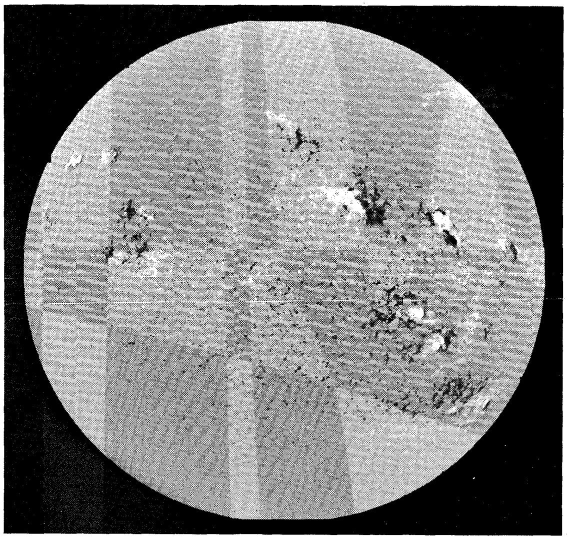


Figure 7: *Photospheric magnetic fields on the Sun.*
The dark and light areas correspond to positive and negative (north and south) polarities. The bipolar nature of most regions is indicated. The figure shows a computer-generated photograph or map built up of direct photoelectric scans covering the entire visible surface.

The solar interior: energy generation and evolution of the Sun

The internal structure of the Sun is determined by its chemical composition, by its mass, and by its previous history. Today, with the development of large, fast computers, theoretical solutions to equations governing the structure of the Sun can be found in a few minutes and the evolutionary track showing the changes in the structure and compositions followed for billions of years.

The development of a mathematical model of the Sun would require that the chemical composition be known throughout its mass, and this is not known. Hence, a sequence of models must be computed and compared to the true Sun. It has been suggested that the flux of neutrinos from the hydrogen-burning fusion reactions in the Sun's interior (see below) may allow direct observations of conditions in the Sun's core, and investigators have set up detectors to capture the neutrinos. The results so far indicate that the central temperature of the Sun is less than 20,000,000 K, but results have not yet fixed the chemical composition.

Energy output

The Sun radiates 3.86×10^{33} ergs of energy every second (from energy = mass times velocity of light squared, $e = mc^2$; this corresponds to a mass loss of 4,700,000 tons per second or 1×10^{-11} solar mass per year). Nuclear reactions are the only possible means that can produce such a great amount of energy. In the hot, dense core of the Sun, within about one-tenth of a solar radius of the centre, nuclear fusions liberate the energy that is emitted millions of years later at the surface of the Sun. The most important fusion process in the Sun is the so-called proton–proton reaction or hydrogen-burning. The fusion of two hydrogen atoms, the nuclei of which are called protons, produces a deuterium atom and releases a positron and a neutrino.

Nuclear reactions

The neutrino escapes with the velocity of light and the positron quickly collides with an electron, whereby they annihilate each other, producing a very high-energy photon of radiation, called a gamma ray. The collision of the deuterium atom with another proton produces an atom of the light helium isotope of atomic mass 3, called helium-3, and the emission of another gamma ray. Eventually, two helium-3 atoms collide and in doing so, produce an ordinary stable helium atom of atomic mass 4 and release two protons and a gamma ray. In total, two electrons and six protons are used up in the production of one stable helium atom; two new protons are rebuilt and two neutrinos and five gamma rays are released in the process, which provide the energy radiated away at the solar surface. About 5 percent of the energy is carried away by the neutrinos (see ATOMS).

The Sun is presently in a state of equilibrium with a balance between its total energy production and its luminosity. The evolution of the Sun has been considered by many theoreticians. The U.S. astrophysicist Icko Iben sees a million years as having been taken for the mass of gas that is now the Sun to evolve from a wholly convective state to a body with a radiative core. The radiative core then grows to encompass nearly all the Sun's mass. After 10^7 years the brightness reaches a minimum and then increases. At 2.3×10^7 years nuclear burning begins. At 2.6×10^7 years a convective core appears and the Sun begins to take on a form much like its present one at 3.7×10^7 years. As the conversion of carbon to nitrogen proceeds to burn carbon rapidly, the convective core is replaced by the condition of radiative equilibrium, in which the amount of energy received is everywhere exactly passed on throughout the mass. In 4.5×10^9 years the brightness increased to the present solar value. At its present stage the Sun consists of about 60 percent hydrogen by weight. After about 6×10^9 years a slight change in mass and luminosity may be recognizable. During the past 4.5×10^9 years, that is, since the formation of the Earth's crust, the Sun has virtually remained unchanged.

Future evolution

The future evolution of the Sun is expected to be similar to that of other normal stars. Eventually all hydrogen will be burned up and nuclear reactions involving helium and heavier atoms will take over. This will change the chemical composition of the Sun; as a result, the Sun will increase in size and luminosity and thus turn into a red giant star. Computations on evolutionary models predict that in a few times 10^9 years the Sun will reach the red giant stage. Finally, when all nuclear energy sources are used up, the Sun will reach its last evolutionary stage, that is, it will become a white dwarf, a star of small radius. Its radius will be about 100 times smaller than it is now. Its internal temperature and its luminosity will gradually decrease on a time scale of 10^9 years and, eventually, the Sun will become a black dwarf, a very dense, nonluminous object of degenerate matter. The total lifetime of the Sun is estimated to be some 10^{10} years. (A.K.P./E.A.M.)

THE MAJOR PLANETS AND THEIR SATELLITES

As previously noted, the known major planets in order of increasing distance from the Sun are Mercury, Venus, Earth, Mars, Jupiter, Saturn, Uranus, Neptune, and Pluto (see Figure 8). The terrestrial planets are Mercury, Venus, Earth, and Mars; the giant planets are Jupiter, Saturn, Uranus, and Neptune. Pluto, the outermost, is believed to be less than one-third the size of Earth, and its orbit crosses the orbit of Neptune. There are reasons for believing that Pluto may once have been a satellite of Neptune that escaped from its gravitational control. The terrestrial planets are of about the same size and mass as the Earth and are of high density, between about three and six times that of water. The giant planets have from 15 to 318 times the volume of Earth but have only about 20 percent the density of the terrestrial planets. Saturn is so light that an average piece of it, if available, would float in water. The high density of the terrestrial planets is attributed to their large metallic content. The giant planets are believed to consist mostly of the lightest elements, hydrogen and helium. They are thought, however, to have a small solid core of iron and silicates.

The inner or terrestrial planets

The Earth and Venus have been described as sister planets because their sizes, masses, and densities are roughly the same. Both have extensive atmospheres, but the composition of these are not the same. Venus has no moon. It is difficult to take the comparison further owing to observational difficulties caused by the extensive cloud cover on Venus. Surface relief and the existence of various landforms such as mountains, volcanoes, craters, and rift valleys have been discovered by radar. Oceans cannot exist on Venus because of the high surface temperature of about 730 K.

Mars, though smaller than Earth, has some similarities, and more extensive photographic coverage in the 1970s revealed the presence of many features on Mars that may have similar origins to those of their counterparts on Earth. Results from space probes sent to Mars and from radar studies have allowed some direct comparisons between planetary details. Many Martian features indicate ancient fluvial erosion, widespread volcanism, and extensive tectonics. Large areas resembling groups of calderas and craters have been photographed. Further, rifts and cracks in long double and single nearly straight lines and areas of intersecting sets of parallel rift systems have been found on Mars. All these can be matched on Earth. Many large height differences (up to several kilometres) have been found to exist on Mars.

Many Martian craters have counterparts on the Moon, which also shows large-scale systems of apparently tectonic lineaments (rills, raised ridges in intersecting sets of parallel systems).

The giant planets

Detailed knowledge of two of the outer planets and cross comparisons of data became possible during the 1980s. The U.S. Voyager 1 and 2 probes passed close to Jupiter in 1979 and then flew by Saturn in the early 1980s, transmitting scientific measurements and high-resolution photographs of both planets.

As pointed out earlier, a satellite is defined as a body that revolves around a planet. It is thus distinguished from a planet by its motion rather than by any physical property. All of the planets except Mercury and Venus are known to have satellites. Some satellites are only a few kilometres in diameter, as in the case of the two tiny moons of Mars and the outer satellites of Jupiter. A few satellites are about the size of Mercury and Mars, as in the case of satellites Ganymede and Callisto of Jupiter, Titan of Saturn, and Triton of Neptune, each of which is about 5,600 kilometres in diameter.

The characteristics of each of the planets and their principal satellites are described below. Considerable attention is given to the Earth's moon to accommodate the wealth of data collected by spaceflights and manned explorations of the lunar surface. (D.G.R./Ed.)

Mercury

Mercury, designated ☿ in astronomy, is the planet closest to the Sun, revolving around it at an average distance

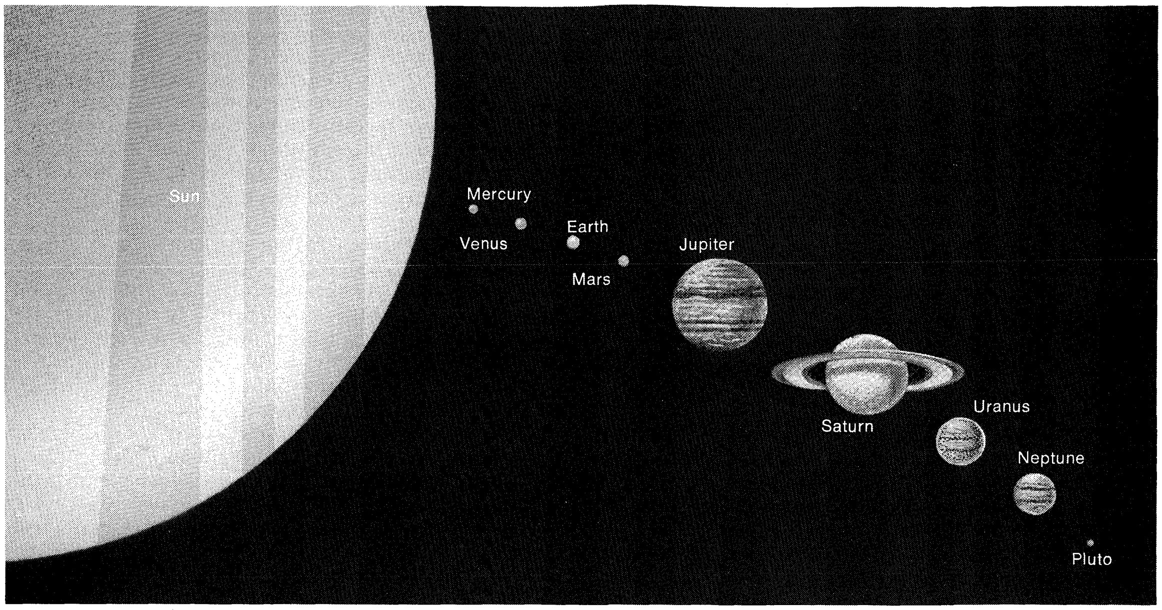


Figure 8: Relative sizes of the planets, arranged in order outward from the Sun (but not according to their respective distances from one another).

of 58,000,000 kilometres (about 2/5 the distance between the Sun and Earth). Mercury's orbit is inside the orbit of Earth, and this creates two important astronomical effects. First, Mercury is never more than 27°45′ of angle away from the Sun and is thus seen with the naked eye only as a "morning" star just before sunrise or an "evening" star just after sunset. Telescopic observations of Mercury are further complicated by mist, haze, and other materials in Earth's atmosphere that scatter and absorb light from Mercury, thereby greatly reducing its visibility. Second, Mercury exhibits phases much like the Moon: when it lies nearly between Earth and the Sun (inferior conjunction), it appears as a thin crescent; when it is at its greatest separation (or elongation) from the Sun, the apparent disk is half-illuminated; and when it is on the opposite side of the Sun from Earth (superior conjunction), its fully illuminated surface is visible. Since these changes in phase occur because of the motion of Mercury in its orbit, its apparent size also varies with the phase. It is largest at inferior conjunction (about 10″ of arc, or 1/180 the apparent size of the Moon) and smallest at superior conjunction (about 4 1/2″ of arc, or 1/380 the apparent size of the Moon).

Phases

Mercury was known to be a planet in Sumerian times, some 5,000 years ago. In classical Greece it was called Apollo when it appeared as a morning star and Hermes, for the Greek equivalent of the Roman god Mercury, when it appeared as an evening star. Hermes was the swift messenger of the gods, and the planet's name is thus a reference to its rapid motions relative to the other objects in the sky.

BASIC ASTRONOMICAL DATA

Mercury's orbit. While orbits of the planets around the Sun are nearly circular, that of Mercury is more elliptical than usual. At its closest approach (perihelion), Mercury is only 46,000,000 kilometres from the Sun, while its greatest distance (aphelion) approaches 70,000,000 kilometres. The eccentricity of this orbital ellipse is 0.206. Mercury orbits the Sun in 88 Earth days at an average speed of 48 kilometres per second, allowing it to overtake and pass Earth every 116 Earth days (synodic period).

Mercury has played an important role in the development and testing of theories of the nature of gravity because its elliptical, inclined orbit—its orbit is 7°004′ out of a plane defined by the total mass of the Sun plus the planets—is perturbed by the gravitational pull of the Sun and the other planets. This small motion (about 10′ of arc per century) has been known for nearly 200 years, and in fact all but about 40″ of arc (about 7 percent) could be explained by the theory of gravity proposed by Newton. This discrepancy was, however, too large to ignore, and explanations were proposed, usually invoking as yet undiscovered planets. In 1915 Albert Einstein showed that his general theory of relativity, which included a treatment of gravity, could explain the small discrepancy. Thus, the motion of Mercury's orbit has been considered an important observational verification of Einstein's theory. Mercury has also been used in additional tests of relativity because radar signals that are bounced off its surface when it is at superior conjunction must pass close to the Sun. The general theory of relativity predicts that such electromagnetic signals, moving in the warped space caused by the immense gravity of the Sun, will follow a slightly different path and take a slightly different time to traverse that space than they would if the Sun were not there at all. By comparing such signals as a function of the position of Mercury with respect to the Sun, a second important confirmation of the general theory of relativity has been acquired.

Models of the composition and evolution of the dust and gas cloud from which the Sun and planets are believed to have accumulated are based on presumptions of the temperature and pressure of the cloud; the composition of a planet at a given distance from the Sun is crucially dependent on these presumptions. Many cosmochemists rely on observations of the density or surface composition of the planets for clues to verify their assumptions. Mercury, being the closest planet to the Sun, is an important "end-member" in these models.

Surface temperatures and seasons. Because of its proximity to the Sun, the surface of Mercury can get extremely hot. High temperatures at "noon" may reach 675 K, while the "pre-dawn" lowest temperatures are 100 K. Thus, portions of Mercury's surface experience temperature variations approaching 575 K.

Seasons on Earth arise from variations in the amount of sunlight reaching its surface. These variations are caused by the tilt of Earth's equator with respect to its orbital plane. Mercury's equator is almost exactly in its orbital plane (its spin axis inclination is nearly zero), and thus Mercury does not have seasons as does Earth. Because of its elliptical orbit and a peculiarity of its rotational period (see below), however, certain longitudes experience cyclical variations in temperatures on a "yearly" as well as on a "diurnal" basis.

Mass, radius, and density. Mercury is about 4,878 kilometres in diameter, smaller than any other planet with the possible exception of Pluto. Mercury is only a bit larger than the Moon and is comparable in size to the largest satellites of Jupiter (Ganymede and Callisto), Sat-

Escape velocity

urn (Titan), and Neptune (Triton). Its mass, measured by the gravitational perturbation of the path of the Mariner 10 spacecraft during close flybys in 1974 and 1975, is 3.30×10^{26} grams, about $^{1}/_{18}$ of the mass of Earth. Escape velocity, that speed needed to escape from a planet's gravitational field, is about 4.25 kilometres per second (compared with 11.2 kilometres per second for Earth). The mean density of Mercury, calculated from its mass and radius, is about 5.44 grams per cubic centimetre, nearly the same as that of Earth (5.5 grams per cubic centimetre). Earth, however, being much more massive than Mercury, owes some of its high density to gravitational compression. Thus, the smaller Mercury must owe its high density to a greater abundance of a naturally denser material, such as iron. Much of Earth's high-density material resides in its core. Mercury probably has a core as well, and it probably is relatively larger than Earth's core. This abundance of high-density material is a constraint on the models of planetary formation. (See Table 4.)

Table 4: Planetary Data for Mercury

Average distance from Sun	58,000,000 km (36,000,000 mi)
Eccentricity of orbit	0.206
Inclination of orbit to invariable plane of solar system	7°004′
Sidereal period of revolution (orbital period = year)	87.9694 Earth days
Rotation period	58.6461 Earth days (= 2/3 of orbital period)
Mean synodic period	116 Earth days
Mean orbital velocity	48 km/sec (30 mi/sec)
Inclination of equator to orbit	<3° (probably 0°)
Mass	3.30×10^{26} gm (0.055 × mass of Earth)
Diameter	4,878 ± 2 km (3,032 mi) (0.38 × diameter of Earth)
Density	5.44 gm/cm^3
Gravitational acceleration	370 cm/sec^2 (0.38 × Earth's gravity)
Satellites	none

Rotation rate. The rates at which planets spin on their axes (rotation rates) have traditionally been determined by visual astronomical observations. Features on a planet are watched as they move across the apparent disk, and a period of rotation is derived from the time they take to traverse the disk. Problems arise with this technique if the features are clouds in a planet's atmosphere, or if the planet is small and its surface features are poorly distinguishable because of intrinsically low contrast, poorly defined margins, or atmospheric conditions at the observatory that cause blurring.

Mercury is so difficult to see from Earth that its rotation rate has been the subject of study for more than 200 years. Initial observations suggested to astronomers in the early 19th century a period of 24 hours, essentially identical to Earth's. This was shown to be erroneous in the late 19th century, when it was concluded that Mercury rotated once every 88 Earth days, exactly equal to its year. This implied that Mercury was in synchronous rotation (always keeping one face toward the Sun), much as the Moon is in synchronous rotation with respect to Earth. Both visual and photographic studies for almost 80 years thereafter seemed to support this view. In the early 1960s, however, observations at radio wavelengths indicated a different story. The amount of radio wavelength energy emitted by a planet is dependent on the temperature of the planet, and observations indicated that the temperature of the "dark side" of Mercury (that side presumed to always face away from the Sun) was much higher than it should have been for a planet in synchronous rotation. Later scientists were able to bounce radar signals off Mercury. These signals showed the effect of a shift in frequency characteristic of moving bodies (the Doppler effect). This effect, familiar as the cause of the change in the pitch of automobile horns as they approach and recede, showed that Mercury's rotation rate was about 59, not 88, Earth days.

Determinations using photographs relayed by the Mariner 10 spacecraft (launched in 1973) led to a much more precise measurement. Mercury rotates once every 58.6461 Earth days, exactly two-thirds of the orbital period of 87.9694 Earth days. This observation confirms theoretical studies, developed after the radio wavelength observations, that suggested that Mercury was in a 3:2 spin-orbit tidal resonance—that is, that tides raised on Mercury by the Sun had forced it into a condition that caused it to rotate three times on its axis in the same time it took to revolve around the Sun twice.

The two-thirds spin orbit

The 3:2 spin-orbit coupling combines with Mercury's eccentric orbit to create very unusual temperature effects. First, although it rotates on its axis once every 59 Earth days, one rotation does not bring the Sun back to the same part of the sky, because during that time Mercury has moved partway around the Sun. A solar day on Mercury (for example, from one sunrise to another, or one noon to another) is 176 Earth days, and equal to two Mercurian years. The eccentricity causes the apparent angular size of the Sun to vary by nearly 50 percent (from 1.1° at aphelion to 1.7° at perihelion, compared to 0.5° as viewed from Earth). Thus, sunlight at perihelion is more than twice as intense as at aphelion. The two longitudes that experience noontime illumination when Mercury is at perihelion are therefore much hotter than the two longitudes that experience noontime illumination at aphelion. The intersections of the equator with these longitudes can be called "poles." Mercury has two "hot" poles, two "warm" poles, and of course, north and south poles defined by the spin axis.

Sunrise and sunset from Mercury

These relationships would create unusual "days" to an observer. At a hot pole a small Sun would rise rapidly, increase in size and decrease in speed as it climbed toward the zenith, actually stop and reverse its course, stop again, and then proceed toward the opposite horizon, decreasing in size while increasing in speed until it set 88 Earth days after it had risen. The stars would be moving through the sky three times as fast as the Sun, and stars that rose with the Sun would set and rise again before the Sun set. An observer at a warm pole would see a large Sun slowly rise, stop, and sink below the horizon, and then rise again, shrinking in size and increasing in speed as it rose toward the zenith. It would then increase in size and slow its pace as it approached the horizon, where it would set, rise again briefly, and finally set again.

ATMOSPHERE AND MAGNETIC FIELD

Tenuous atmosphere

Mercury's low escape velocity and high surface temperatures do not permit it to retain a significant atmosphere. Its total atmospheric pressure, determined by experiments during the Mariner 10 spacecraft mission, is about 2×10^{-9} millibar, or $^{1}/_{500,000,000,000}$ that of Earth, as low as many vacuums created in terrestrial laboratories. The composition of this very tenuous atmosphere is principally hydrogen, helium, and neon, with upper limits placed on the presence of heavier, but much less abundant, gases such as argon, carbon dioxide, krypton, and xenon. Those actually detected may come from the release of primordial gases from deep within the planet, or, more likely, from material flung off the Sun by activity in its high-temperature atmosphere (the solar wind). The abundance of helium is compatible with either a solar wind source or a release of helium from rocks containing the radioactive elements uranium and thorium. Surface features on Mercury show no evidence of atmospheric interaction—no sand dunes, streaks of sand or dust, wind etching of surface forms, or similar phenomena. Because many of Mercury's surface features are very old, it has been concluded that Mercury has not had an appreciable atmosphere since it formed, probably 4,500,000,000 years ago.

Mercury has a small but important magnetic field, with a strength of about 1 percent that of Earth and a moment about 6×10^{-4} that of Earth (Figure 9). This field sets Mercury apart from the Moon, Venus, and probably Mars, which do not display such fields. Mercury's magnetic field can be modelled as a dipole with its axis more or less aligned with the spin axis, thus making the resemblance to Earth even closer. The origin of the magnetic field is unknown but of great scientific importance because it may help to understand the origin and nature of Earth's field and also the evolution of Mercury's interior. Theoretical studies suggest that Earth's magnetic field is actively

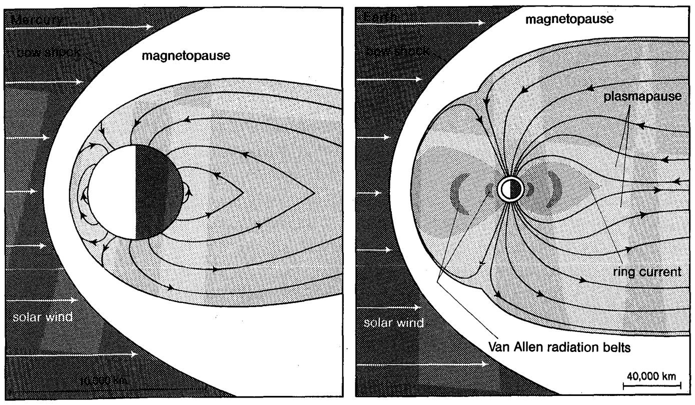


Figure 9: Comparison of the magnetic fields of Mercury and Earth. Mercury's field is very similar to Earth's, but smaller.

generated by the motion of hot, electrically conductive materials in its core (a magnetohydrodynamical dynamo). If Mercury's field is similarly created, then Mercury's interior is still hot and active, even though little evidence of this is seen on its surface. Alternatively, Mercury's magnetic field may be a remnant of an earlier epoch, when a dynamo was active and capable of magnetizing rocks that could retain the imprint of the ancient field after the dynamo waned. A third alternative is that the field is not indigenous to Mercury, but rather is induced by the immensely stronger solar magnetic field. It appears that such induction cannot create today a field as strong and steady as that observed, but in the distant past, when the Sun's field may have been even stronger, it may have been able to induce a sufficiently large field in Mercury to create a remnant field preserved in magnetized rocks.

THE SURFACE

Surface properties. The surface properties of Mercury (the composition of its rocky material and its small-scale structure) can be only inferred from optical and thermal- and radio-wavelength observations. Because such techniques have been applied to the Moon and Earth, as well as to samples of lunar and terrestrial rocks and soils, observations of Mercury can be compared to known materials and its properties more accurately inferred.

Similarities to the Moon's surface

The reflection of sunlight off Mercury's surface as viewed from different angles suggests a surface layer with a rough texture, composed of fine, dark grains of rock. Temperature measurements at different wavelengths in the infrared and microwave portion of the spectrum, and at different viewing angles, support this interpretation. Multispectral observations at visible and near visible infrared wavelengths indicate a general similarity to mature lunar soils, with a subtle absorption of radiation at about one micrometre wavelength (visible light extends from about 0.4 to 0.7 micrometre) suggesting about as much iron-rich material as is found in the lunar highlands (about 5 percent ferrous oxide by weight).

By courtesy of (top) Jet Propulsion Laboratory/National Aeronautics and Space Administration, (bottom) U.S. Geological Survey/National Aeronautics and Space Administration

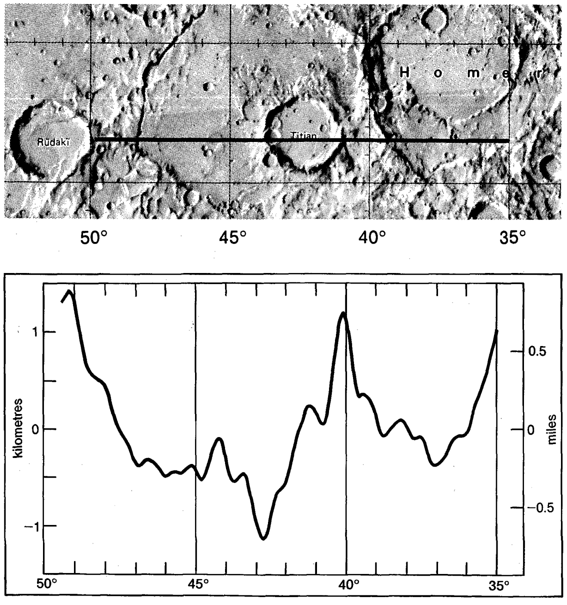


Figure 10: Earth-based topographic profile along the line indicated on the map shows large craters to be between one and two kilometres deep.

Earth-based radar observations show a rough topography much like that of the Moon, including depressions with raised rims and blocky surfaces that clearly portray their impact origin (Figure 10). These observations, acquired before the Mariner spacecraft mission, were the first indication of lunar-like surface features, as well as surface physical properties and composition.

Spacecraft observations contribute additional information on physical and compositional properties of the Mercurian surface. The darkest features on Mercury are not as dark as the darkest features on the Moon, but the overall brightness (albedo) of many Mercurian surfaces is comparable to that of the heavily cratered, relatively bright lunar highlands, which reflect about 15 percent of the sunlight striking them. The brightest areas seen on Mercury are found in the floors of a few unusual craters, where as much as 50 percent of the incident sunlight is reflected. The average brightness of Mercurian surfaces suggests a composition deficient in titanium- and iron-bearing minerals such as are found in the lunar maria (low-lying dark areas), and perhaps more like the calcium-bearing minerals of the lunar highlands. The few extremely bright crater floors require compositions or surface textures unlike any seen on the Moon.

The Mercurian regolith

Thermal infrared observations from Mariner 10 show results similar to Earth-based studies; namely, that the Mercurian surface is covered by a layer of fine silicate dust and sand at least tens of centimetres thick. Such a regolith, or rock mantle, is also found on the Moon, where it is formed by the repeated pulverizing of the surface rocks by the impact of microscopic particles moving through space at many tens of kilometres per second. Infrared observations also show, however, areas of compacted soil or boulders or bedrock, indicating a few places of thin regolith or fresh, unweathered surfaces.

Observed features. Mercury is a difficult planet to observe from Earth, and this led to several erroneous conclusions about it. Such observations could at best attempt to resolve features several hundreds of kilometres across, but usually this was not possible and only regions 1,000 kilometres on a side could be seen. It was not until the Mariner 10 spacecraft flew past Mercury that high-resolution views of its surface were obtained (Figure 11) and surface features unambiguously described. At their best resolution the Mariner 10 photographic data are nearly 5,000 times better than Earth-based views.

Six types of features are seen on the Mercurian surface: impact-generated craters, usually named for men and women who have made contributions to the arts and humanities; scarps, named for ships used in exploration on Earth; ridges and valleys, named for astronomers and radio observatories, respectively; plains, named for the planet or god Mercury in various languages; and mountains, usually the rims of extremely large craters, named for the nearest plains.

Impact craters. Impact craters are the most abundant landform on Mercury, an attribute shared with many other solid bodies in the solar system. Craters come in many sizes and shapes, though they usually appear nearly circular. They are studied intensively for two reasons. First, their abundances and variations in size as seen on several planets permit investigation of similarities and differences between planets. Second, the number of craters on a given surface is most likely an indication of that surface's age—the older the surface, the more craters that have formed on it.

The craters on Mercury are very much like those on the Moon. The smallest visible craters are about one to two kilometres across; they are bowl-shaped, with raised rims

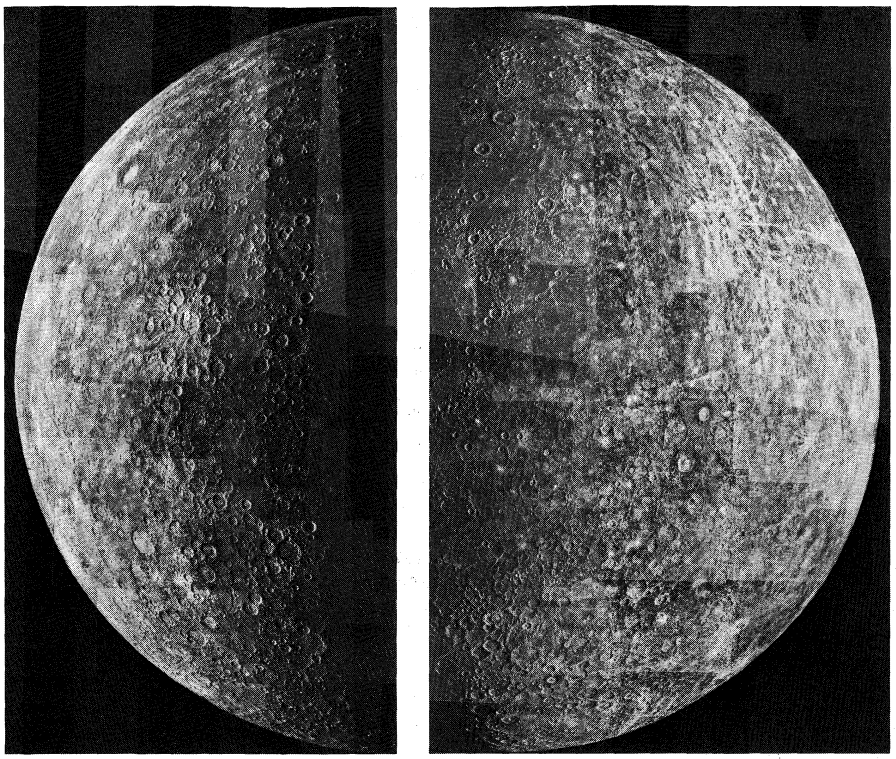


Figure 11: Two mosaic views of Mercury, each about one-half in shadow. The view on the left, taken as Mariner 10 approached Mercury, shows impact craters, intercrater plains, scarps, and hilly and lineated terrain. The view on the right, taken after Mariner 10 had flown past Mercury, shows the enormous Caloris impact basin, smooth plains, bright rays associated with very fresh craters, and many older craters.

By courtesy of the Jet Propulsion Laboratory/National Aeronautics and Space Administration

and blankets of ejecta surrounding these rims and thinning rapidly away from the crater. Slightly larger craters, about 10 kilometres in diameter, begin to show interior complexities (slumping, terraced walls, central mounds or peaks, and flat floors), along with more extensive ejecta blankets. Secondary craters, formed by large blocks ejected by an impact, form chains and clusters beyond the zone of continuous blanketing. The largest craters often have a ring of isolated or continuous peaks inside and concentric to their rim.

Influence of gravity

Subtle variations in the morphology of fresh craters may show the influence of gravity and material properties on impact phenomena. One dramatic influence of gravity on the cratering process is its effect on ejecta. Mercury's gravity is 2.3 times greater than that of the Moon, and this permits ejecta to reach only 65 percent of the distance it would reach on the Moon. By studying craters of the same size on the two bodies, it can be shown that the amount ejected is roughly the same, so the restriction of Mercurian ejecta causes a higher concentration of secondary craters nearer the crater and an increase in the thickness of the continuous ejecta blanket. These restrictions also reduce the effect on Mercury of proximity erosion, in which ejecta from one crater lands on and covers or degrades an adjacent crater.

The number of craters on various surfaces on Mercury also shows remarkable similarities to lunar cratering. The age of some lunar surfaces is known because Apollo and Luna space missions brought to Earth samples of rocks that have been dated by radioactive age-dating techniques. Although it is not certain that the same number of craters were formed at the same rate on Mercury as on the Moon, their histories have probably not been widely different. Thus, Mercurian surfaces are thought to be very old—perhaps 3,000,000,000–4,000,000,000 years old. Variations in the details of the number of craters at different sizes, and differences in these from comparable variations in the lunar craters, suggest a more complex picture, combining variations in the number of crater-forming impacts (reflecting different sources of interplanetary debris) and variations in crater preservation after formation (reflecting planetary processes such as global melting and volcanism).

Escarpments. Among the unique features of the surface of Mercury are enormous escarpments up to hundreds of kilometres in length and as much as three kilometres high. Many of these scarps cut through surfaces and topographic features (Figure 12), while in other areas they are themselves covered. These relationships permit both the relative age and possible origins of the scarps to be studied. The morphology of the scarps and their transectional relationships show that most occur along steep, nearly vertical faults. By close inspection of several craters cut by scarps, it appears that portions of these craters have been shoved on top of other parts (Figure 12). Thus, the scarps are inferred to have formed by compression. The number and length of scarps imply a shrinking of the surface area of Mercury by about 0.1 percent, or a decrease in the planet's radius of between one and two kilometres.

The shrinking of Mercury

The orientation of the scarps and other linear features further constrains the mechanisms of global crustal fracturing and movement (tectonic deformation). These suggest that tidal stresses induced in Mercury by the Sun (probably those that established the spin-orbit resonance) deformed the crust in a specific manner, depending on latitude, and that concurrent or subsequent to this tidal deformation the interior of Mercury cooled, causing the planet to shrink. Most scarps are relatively old; a few small, young scarps probably formed later by local deformation (associated with, for example, impact cratering or volcanism) and usually follow the trends established in the crust by the earlier global deformation.

Ridges and valleys. Three types of ridges are found on Mercury. Relatively narrow, discontinuous ridges are usually found radial to extremely large impact craters (called basins) and are often accompanied by intervening valleys; they represent impact-induced radial faulting and radial deposition and grooving during ejecta emplacement. Many well-defined ridges are found in areas of smooth, relatively lightly cratered flat surfaces (smooth plains) and probably were formed through local adjustment of the underlying rock to the weight of the plains-forming material. Finally, large ridges (up to several hundred kilometres long), found

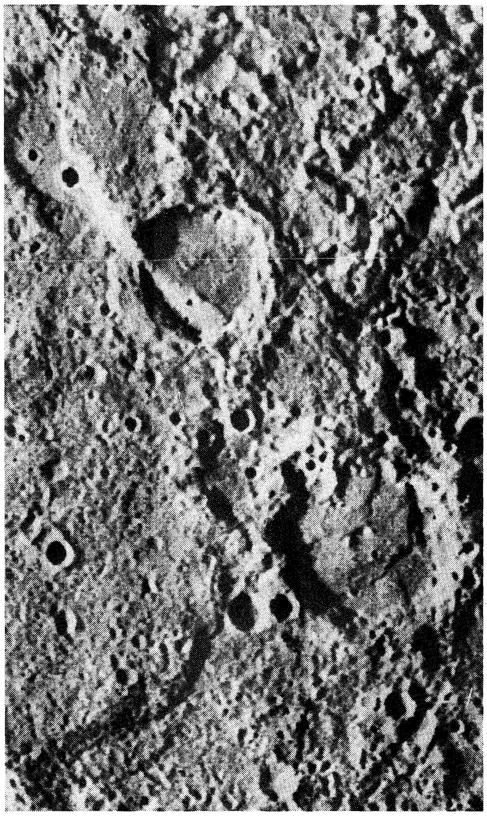


Figure 12: The Vostok scarp cuts two craters, as seen in the upper left portion of this Mariner 10 photograph. The lower of the two craters is called Guido d'Arezzo and is 65 kilometres across. The upper right portion of the crater appears to be thrust over the lower left.

By courtesy of the Jet Propulsion Laboratory/National Aeronautics and Space Administration

in the most heavily cratered areas and in the moderately cratered zones between these areas, follow trends similar to scarps and are inferred to be sources of volcanic material erupted early in Mercurian history.

Valleys are relatively rare on Mercury. They were formed mostly in association with impact craters, either as an integrated chain of secondary craters, as radially grooved ejecta, or as radially faulted, uplifted bedrock. A few valleys may also have formed as fault-bounded depressions (grabens) of tectonic origin.

Plains. Plains, or relatively flat or smoothly undulating surfaces, are ubiquitous on Mercury and the other planets. They represent a canvas on which other landforms develop. The covering or destruction of a rough topography and the creation of a smoother surface is called resurfacing, and plains are evidence of this process. There are at least three ways that planets are resurfaced, and all three may have had a role in creating Mercurian plains. One mechanism is to reduce the strength of the crust to retain high relief by raising its temperature and, over millions of years, allowing the mountains to sink and craters to rise. A second mechanism involves the horizontal flow of material along surface topographic gradients, eventually collecting in low areas and filling up to higher levels, as more volume is added. Lava flows behave in this manner. The third mechanism is for fragmented material to be deposited on a surface from above, first mantling and eventually covering the rough topography. Blanketing by impact crater ejecta and volcanic ash are examples of this mechanism.

Mountains. Mercury can be said to be mountainous only to the extent that the largest impact craters create zones of blocks of rock that are regionally extensive and of great local relief. The mountain-building processes common on Earth (involving the collision of large regional plates of Earth's crust moving under the influence of internal heat) do not occur on Mercury. The mountains on Mercury are mostly the exaggerated rims and rim deposits of very large craters.

The physiography of the planet. The features described above combine in regional and global patterns to establish physiographic provinces that reflect the geological evolution of Mercury (Figure 11). Three principal types of terrain are found globally, and a few others are found only in certain regions. The three global types are heavily cratered terrain, intercrater plains, and smooth plains; and the regional terrains include those associated with an enormous crater called the Caloris Basin and an area of highly disrupted hills and depressions that occurs on the other side of the planet, exactly opposite Caloris.

Heavily cratered terrain. The term heavily cratered terrain is applied to areas of closely grouped and overlapping craters with diameters greater than about 30 kilometres (Figure 11). The small distances between craters inhibit the detection of ejecta blankets and fields of secondary craters about individuals within the group. Many of the craters appear degraded (*i.e.,* they have smoother walls, more rounded rims, and are pocked by additional small craters) when compared to more isolated, fresher looking craters. This degradation probably reflects in part their greater age and in part the effectiveness of erosion and deposition by ejecta from adjacent craters (proximity erosion). The interiors of many craters in heavily cratered terrain are occupied by relatively smooth plains.

Degradation of craters

Intercrater and smooth plains. Intercrater plains, the term applied to the less cratered zones between the areas of heavily cratered terrain, are the most widespread physiographic unit on Mercury. They may be relatively smooth or have a low, irregular, undulating relief. They are characteristically covered by many small craters (mostly less than 10 kilometres in diameter), thought to be secondary craters but which cannot be traced back to a specific primary source. The time relationships between the formation of heavily cratered terrain and intercrater plains are difficult to discern. In some areas secondary craters and ejecta from craters clustered in an area of heavily cratered terrain can be traced out onto intercrater plains. Thus, in these areas the plains are older than the cratered terrain. In other areas, however, the clustered craters do not have discernible secondary craters or ejecta, and the relative ages are thus indeterminate. In still other areas ejecta relations indicate that the craters came after the plains, yet under close stereoscopic scrutiny the intercrater plains areas are seen to be themselves clusters of subtle, circular depressions of low relief. It appears from all of these cases that heavily cratered terrain and intercrater plains were formed during the same interval, although the rate of formation of plains appears to have differed from the rate of production of craters.

Smooth plains are relatively deficient in craters and are often found inside young, large craters. They are thus considered to be among the youngest features on Mercury. The largest observed occurrence of smooth plains fills and surrounds the Caloris Basin. Ridges and scarps are abundant on smooth plains. The plains resemble the lunar maria in pattern, location of occurrence, surface texture, and crater population, but they do not differ in albedo and colour from their surroundings, as do the maria.

The mechanisms by which the intercrater and smooth plains were produced are of great importance because their age differences permit investigation of the evolution in time of the resurfacing on Mercury. Initial thoughts were that both plains were formed by volcanic flows, and that the lesser abundance of younger, smooth plains indicated that Mercury was initially hot and volcanically active but had cooled as it lost its primordial heat and become less active with time. An opposing view suggested that both kinds of plains were formed by enormous amounts of ejecta derived from craters, that the greater abundance of older intercrater plains was the result of the large number of impacts occurring early in Mercurian history, and that the smooth plains represented the ejecta from the last, great impacts, such as that which formed the Caloris Basin. An additional view suggested that the areas of intercrater plains, seen to be subdued depressions, were in fact craters degraded by the inability of a warm crust to support their relief (the result of crustal plasticity). It now seems likely that all three mechanisms operated on Mercury. Although scientists still debate the relative proportions, it is most likely that volcanism is responsible for most of the resurfacing, with crustal plasticity responsible for early shaping of most of the heavily cratered regions into intercrater plains between areas of heavily cratered terrain. That most smooth plains were produced by volcanic activity rather than by impact ejecta is suggested by the areal restriction of ejecta owing to gravitational effects, the absence of specific source craters, the wide range in the areas of smooth plains compared to the areas of craters in sites where they occur together, and the occurrence of plains inside craters without additional, exterior craters as sources. Some impact ejecta effects are seen, however, notably in proximity erosion, and plains formed by such ballistic processes may be locally important.

Evidence of volcanism

The Caloris Basin. The Caloris Basin—so named because it lies on the longitude of one of the hot (Latin: *calor*) poles—is about 1,300 kilometres in diameter (Figure 13). The interior of Caloris is occupied by smooth plains that are extensively ridged and fractured in a crudely radial and concentric pattern. The largest ridges are a few hundred kilometres long, about three kilometres wide, and less than 300 metres high. Fractures are comparable in size, and may be flat-floored. Some resemble the fault-bounded depressions that are called grabens. Where they cross, the fractures transect the ridges. Two terrains surround the Caloris Basin—the basin rim and basin ejecta terrains. The basin rim consists of a ring of irregular mountains, up to two kilometres high, bounded on the interior by a relatively steep escarpment and on the outside by a smaller, inward-facing scarp. The intervening 100 to 150 kilometres consists of mountain blocks up to 50 kilometres square, with plains within the depressions between blocks. Beyond the rim is a zone of linear, radial ridges and valleys (see above). These ridges and valleys are embayed and partially filled by two types of plains, some smooth and some with many knobs and hills only a few tenths of a kilometre across. These plains may represent basin ejecta or volcanic material, but their relationship to the large structures of the Caloris Basin indicate that they formed after, though not necessarily long after, the basin.

By courtesy of the Jet Propulsion Laboratory/National Aeronautics and Space Administration

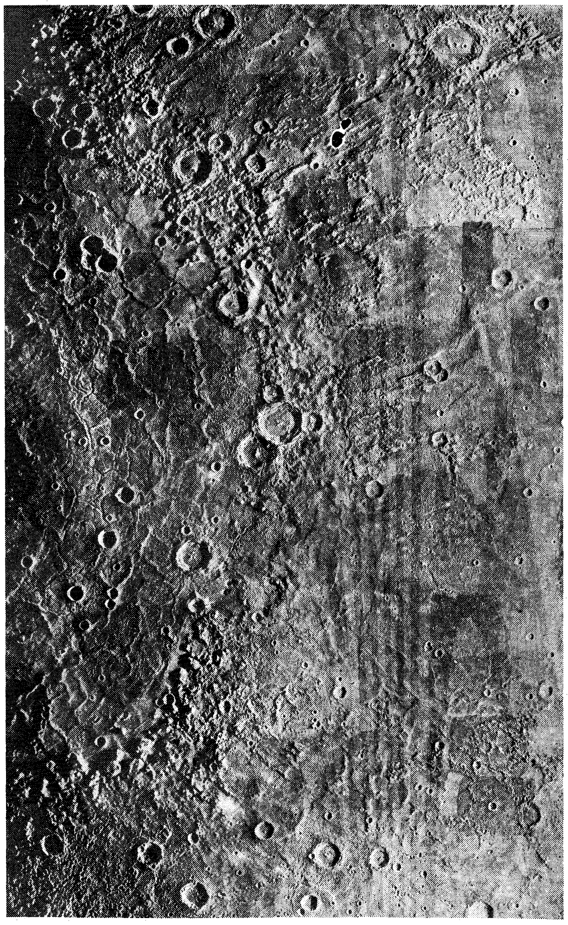


Figure 13: The Caloris Basin, as seen in a mosaic of Mariner 10 photographs.

Region opposite Caloris. Antipodal to Caloris is an extensive area of hills and depressions termed hilly and lineated terrain (Figure 14). The hills are 5 to 10 kilometres wide and up to 1.5 kilometres high, and they have crudely polygonal outlines. Crater rims have been disrupted into hills and fractures by the process that created the terrain. Some of these craters have smooth floors that have not been disrupted, suggesting a later infilling of smooth plains material. The location of this terrain exactly opposite the Caloris Basin, and the similarity in part to features seen antipodal to some lunar basins, suggests to most investigators that some link exists, either by ejecta travelling preferentially halfway around the planet or by the focussing of seismic energy. The second alternative has come to be most favoured.

By courtesy of the Jet Propulsion Laboratory/National Aeronautics and Space Administration.

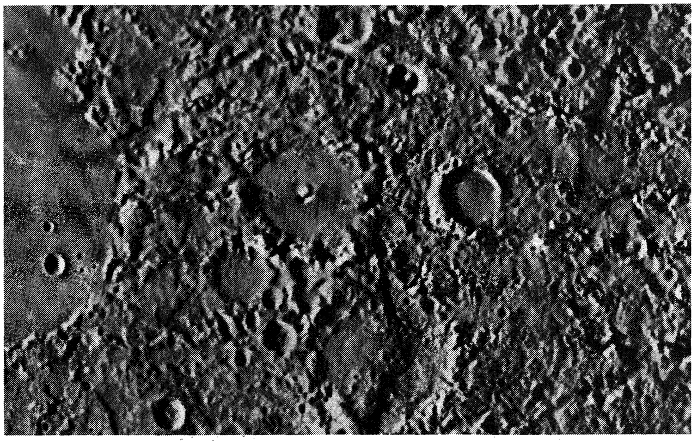


Figure 14: Hilly and lineated terrain antipodal to the Caloris Basin. The area is about 400 kilometres across.

GEOLOGICAL HISTORY

The geological evolution of Mercury can be divided into five stages: initial formation (accretion) and segregation of materials (differentiation); early thermal and dynamical evolution; decreasing bombardment and planetary cooling; reduced crater and plains formation; and relative quiescence.

Mercury, like the other planets, is believed to have formed from materials within a gas and dust cloud associated with the early Sun. This cloud (the solar nebula) condensed to form solid materials, and through processes of gas drag and gravitational interaction eventually formed small bodies that collided to form the planets. Because of its proximity to the Sun, Mercury was probably made of materials stable at relatively high temperatures. It was also probably spinning faster than it is today, because as it accumulated mass it also accumulated spin (angular momentum), much as an ice skater speeds up by drawing in his arms. Sometime during or shortly after this process of accretion, Mercury's interior materials segregated to form an outer shell of iron-poor silicates and an inner zone of iron. Though there is no direct evidence of this differentiation, the high density of Mercury cannot be reconciled with its silicate outer surface without a core of iron or nickel–iron. Mercury probably has an outer zone, perhaps 500 kilometres thick, of silicates much like the lunar highlands, over an iron core amounting to 75 to 85 percent of the radius and 50 percent of the volume of the planet. At its formation this core was probably liquid, but it now may not be.

Early evolution of Mercury

The early evolution of Mercury was next dominated by its early thermal history, the dynamical evolution of its spin and orbit, and the intense bombardment of its surface by interplanetary debris. Little direct evidence of this period is preserved on Mercury, although the end results are clear. The gravitational energy released by core formation would have greatly heated Mercury if it were not already hot. Craters formed by continuing accretion of the few bodies left near Mercury would surely be destroyed or highly degraded as the surface lost its capacity to support relief (crustal plasticity) and as volcanism occurred. Such processes probably created the intercrater plains, and the preserved heavily cratered terrains represent those areas formed late in this epoch, when both crater destruction and plains-forming volcanism were already waning. During this entire period, however, Mercury was being affected by solar tides, slowed in its spin, and distorted in shape. These distortions were preserved in the cooling upper silicate layer as fractures that guided volcanic extrusions and crater excavation.

Toward the end of the period of heavy bombardment, not only had the silicate shell cooled, but most likely the core had begun to solidify as well. This caused the planet to shrink slightly, creating the large scarps. In order for the core to "freeze," however, even more heat had to be deposited at the interface between the iron core and silicate shell. This may have provided material for smooth-plains-forming volcanism. The end of this period was marked by the enormous Caloris impact.

The seismic energy and fracturing of the crust that attended the Caloris impact permitted a final surge of plains formation, one occurring after the cratering had fallen off to a very low and nearly constant rate. Response of the surface to the weight of these plains caused fractures and ridges to develop, usually along patterns inherited from earlier deformation. Finally, Mercury's interior was sufficiently cool that volcanic activity ceased altogether. With the exception of occasional impacts, little else has occurred there that is visible in spacecraft images or discernible in Earth-based telescopic observations of composition and physical properties.

As inferred from relative crater abundances on the different cratered and plains surfaces, from theoretical models of heat generation and loss, and from dynamical models of spin evolution—and by analogy to lunar evolution known from radioactive age dating of samples—most of this sequence occurred relatively early in Mercury's history. If accretion occurred 4,500,000,000 years ago, core formation and despinning probably was complete by 4,300,000,000 years ago, the crust was rigid and volcanism on the wane by 4,100,000,000 years ago, and the surface and interior were essentially quiescent for well over 3,500,000,000 years. Thus, despite an active early life, Mercury is now, like the Moon, probably a dead world. (M.C.Ma.)

Venus

The second planet from the Sun, Venus is, after the Moon, the brightest natural object in the night sky of the Earth. Venus, symbol ♀ in astronomy, revolves about the Sun in a nearly circular orbit (eccentricity 0.007), which is inclined to the ecliptic plane (the plane in which the combined centre of gravity of the Earth and Moon revolves around the Sun) by 3° 24′. It is the planet that comes closest to the Earth, being less than 42,000,000 kilometres away at inferior conjunction (when Venus is between the Earth and the Sun). At superior conjunction (when Venus lies almost behind the Sun), the disk of Venus is fully illuminated as seen from the Earth; it is then about 256,000,000 kilometres from the Earth. Its mean distance from the Sun is 107,500,000 kilometres (Table 5).

BASIC ASTRONOMICAL DATA

Venus' orbit. As it revolves around the Sun, Venus undergoes phase changes similar to those of the Moon. Its sidereal period (the time needed to complete one revolution around the Sun) is 225 Earth days, but its synodic period (the time in which it goes through one cycle of phases) is 584 Earth days. Galileo's historic early observation of the gibbous phase of Venus (seen with more than half but not all of the apparent disk illuminated) gave decisive evidence against the Ptolemaic, or geocentric, theory of the arrangement of the planets. It also solved a problem of the heliocentric theory; Venus is not much brighter when it is nearest the Earth, because we then see only a thin crescent. When it is farther away we see a larger fraction of the illuminated hemisphere. Galileo cautiously announced his observation in an anagram in Latin, then the international language of science. A rearrangement of the letters of Galileo's message gives *Cynthiae figuras aemulatur mater amorum,* or "The mother of Love [*i.e.,* Venus] imitates the form of Cynthia [*i.e.,* the phases of the Moon]."

Table 5: Planetary Data for Venus

Mean distance from Sun	107,500,000 kilometres 67,200,000 miles
Eccentricity of orbit	0.007
Inclination of orbit to ecliptic plane	3° 24′
Sidereal period of revolution (year)	225 Earth days
Rotation period	−243.0 ± 0.1 Earth days
Mean synodic period	584 Earth days
Mean orbital velocity	about 21.8 miles (35 kilometres) per second
Inclination of equator to orbit	very small
Mass (Earth = 1)	0.81
Diameter (atmosphere not included)	12,103 ± 0.2 kilometres
Density	about 5.1 grams per cubic centimetre
Satellites	none known
Acceleration due to gravity (Earth = 1)	0.904
Atmospheric composition	0.96 CO_2, 0.035 N_2, 0.0002 H_2O; trace quantities of CO, O_2, SO_2, HCl, and other gases
Surface atmospheric pressure (Earth = 1 bar)	94 ± 1 bars
Mean surface temperature	735 ± 5 K
Mean visible cloud temperature	230 ± 10 K

Venus as morning or evening star

In middle latitudes Venus can rise or set more than three hours before or after the Sun. When seen in the western sky as the evening star, it was called Hesperus by the Greeks, and when visible in the east as the morning star it was called Phosphorus. Probably as early as 3100 BC it was associated with the goddess of love.

Gross characteristics. At superior conjunction, when it is farthest from the Earth and beyond the Sun, the angular diameter of Venus is about 10 seconds of arc; but at inferior conjunction, when the Earth and Venus are practically in line with the Sun and on the same side of it, it exceeds 60 seconds. The mean radius of the solid body of Venus is 6,051.5 ± 0.2 kilometres, compared to 6,371 kilometres for the Earth; and its mass, as determined by its gravitational influence on other astronomical bodies and on such spacecraft as Mariner 5, is about 0.81 that of the Earth.

The acceleration due to gravity on the surface of Venus is about 90 percent that on Earth. For these reasons Venus has been thought of as "Earth's sister planet." Recent information, however, indicates that this phrase is not altogether appropriate.

Retrograde rotation

Period of rotation. Since, from the vantage point of the Earth, Venus is completely covered by clouds, it is not possible to determine its period of rotation by optical examination of its surface features. From observations of faint and evanescent markings in the clouds, various observers have deduced periods of rotation ranging from 24 hours to 225 Earth days. The development of radar astronomy, however, has provided more precise information. The period of rotation can be determined by measuring the Doppler effect at the limbs of the planet—that is, the increase or decrease in the frequency of a radar wave after reflection from the approaching or receding limbs, or edges. Individual surface features that reflect radar as they move across the planet's rotating disk are now known also, and these can be followed and a period of rotation determined. Accordingly, it is known from the work of R.F. Jurgens and others that the planet rotates in 243.0 ± 0.1 Earth days in a retrograde sense; *i.e.,* in a direction contrary to that of the motion of the other planets. The retrograde rotation of Venus is quite remarkable. With the marginal exception of Uranus, whose rotational axis is nearly in the plane of the ecliptic, all the other planets in the solar system rotate in a direct sense; that is, if the solar system is viewed from above (to the north), where the planets appear to revolve around the Sun in a counterclockwise direction, then all the planets except Venus and Uranus will also rotate about their axes in a counterclockwise direction. Venus, however, rotates clock-wise, in the direction opposite to its revolution. Radar and passive radio determination of the position of the axis of rotation show it to be very nearly perpendicular to the plane of its orbit. Another remarkable fact about the rotation of Venus is that with a rotation period of nearly 243.16 days, a value close to the observed period, Venus presents the same face to the Earth at every inferior conjunction. The questions as to why Venus should be rotating "backward" and why it should be influenced by the presence of the Earth are unsolved celestial mechanical problems connected with the origins and early histories of the planets. One suggestion is that Venus has a pronounced bulge or lumpiness in its equatorial plane, which permits the Earth's gravitational force to exert a tidal couple on Venus. Another suggestion is that the Earth's gravitational influence operates not only on the solid body of Venus but also on its massive atmosphere (see below).

EARLY SCIENTIFIC VIEWS

Perhaps the earliest scientific speculation on the environment of Venus was that of Christiaan Huygens, who, in the 17th century, engaged in prolonged scrutiny of Venus with one of the earliest astronomical telescopes: "I have often wonder'd that when I have view'd Venus . . . she always appeared to me all-over equally lucid, that I can't say I observed so much as one Spot in her . . . is not all that Light we see reflected from an Atmosphere surrounding Venus?"

The appearance of Venus during transit, as it passed in front of the Sun in the year 1761, suggested to Mikhail Vasilyevich Lomonosov that Venus has an atmosphere at least as dense as the Earth's. The visible face of the planet shows, in the experience of almost all visual observers, an unbroken appearance, which suggested to many a very thick cloud layer. By analogy with the Earth, such clouds were imagined to be composed of water. But the Earth is about 50 percent cloud covered and Venus is about 100 percent cloud covered. The idea then evolved that Venus was a wetter place than the Earth, suggesting in 1918 to Svante August Arrhenius that "everything on Venus is dripping wet . . . a very great part of the surface of Venus is no doubt covered with swamps . . . the constantly uniform climatic conditions which exist everywhere result in an entire absence of adaptation to changing exterior conditions. Only low forms of life are therefore represented, mostly no doubt, belonging to the vegetable kingdom; and the organisms are nearly of the same kind all over the planet."

If the clouds were composed of water, it should be possible to detect water vapour above the clouds by spectroscopic means. An unsuccessful attempt to make such a water vapour detection by Charles Edward St. John and Seth Barnes Nicholson in the 1920s led them to deduce instead a dry desert planet with clouds made of dust stirred up from the surface by winds. Subsequent speculation about Venus proposed a surface covered with carbonated water or with petroleum-like hydrocarbons. More definitive information is now available about the surface conditions on Venus (see below).

THE CLOUDS

The clouds of Venus have a high reflectivity. At some visible and near-infrared wavelengths, they reflect 85 percent or more of the sunlight falling on them. The clouds are not perfectly white but exhibit a pale lemon-yellow hue, due to the presence of some material that absorbs blue light. Although they are featureless when viewed in visible light, when photographed using ultraviolet wavelengths the clouds display variable, low-contrast features (see Figure 15). When the reflectivity at all wavelengths is considered, it is possible to calculate the clouds' effective temperature, *i.e.*, the temperature at which they radiate back to space the same amount of energy that they absorb from the Sun. This temperature is about 230 K, exactly the value given by direct infrared measurements of the clouds by William Sinton and John Strong, among others.

During the 1970s, temperatures within the clouds of Venus were directly sensed by Soviet and U.S. space probes in their descent through the planet's atmosphere. The U.S. Pioneer Venus mission of 1978 sent four probes into widely separated regions of the planet. These probes measured the detailed structure of temperature within the clouds and found that the temperature of 230 K occurred about 70 kilometres above the planet's surface—where the cloud tops were observed to be by other probe experiments—thereby confirming the temperatures estimated from reflectivity data.

The cloud layers

The Soviet Venera probes and the U.S. Pioneer Venus probes and orbiter also determined many of the physical and chemical characteristics of the clouds, as well as their altitudes and depths. The clouds are in three closely spaced layers between about 46 and 70 kilometres above the

By courtesy of L.D. Travis and W.B. Rossow

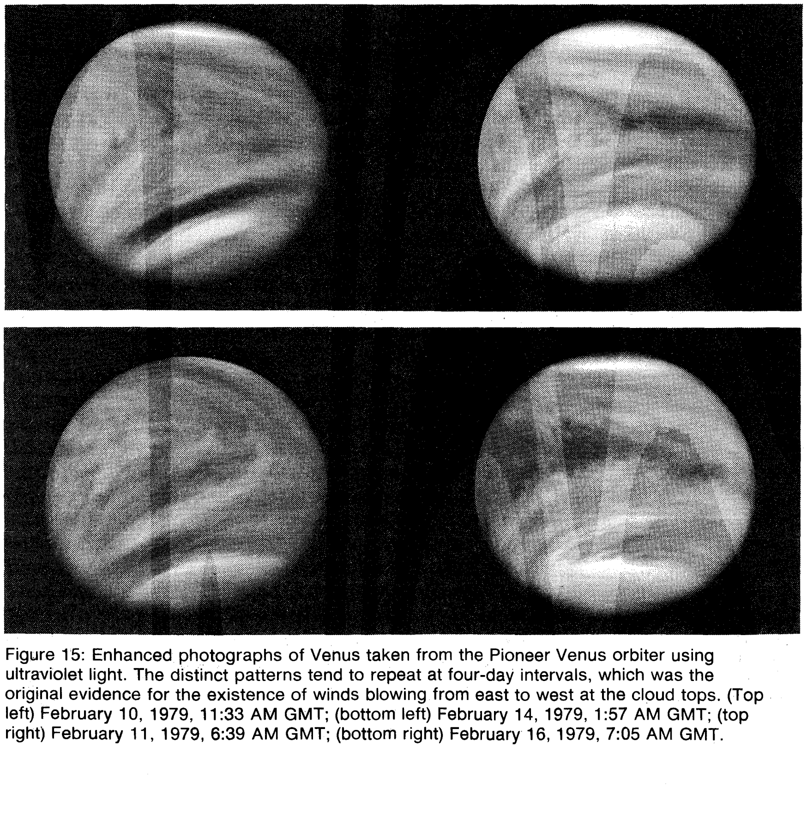


Figure 15: Enhanced photographs of Venus taken from the Pioneer Venus orbiter using ultraviolet light. The distinct patterns tend to repeat at four-day intervals, which was the original evidence for the existence of winds blowing from east to west at the cloud tops. (Top left) February 10, 1979, 11:33 AM GMT; (bottom left) February 14, 1979, 1:57 AM GMT; (top right) February 11, 1979, 6:39 AM GMT; (bottom right) February 16, 1979, 7:05 AM GMT.

Table 6: Composition of the Atmosphere of Venus

gas	mixing ratio	gas	mixing ratio
CO_2	0.96	He	10 ppm
N_2	.035	Ne	10 ppm
H_2O	<.001	Kr	<0.2 ppm
Ar	70 ppm	COS	<2 ppm
SO_2	600 ppm*	HCl	0.4 ppm
CO	20 ppm	HF	0.01 ppm
O_2	40 ppm		

*At cloud base.

surface. Outside of these layers, the atmosphere is almost free of particles, except for a haze of tiny particles about 0.002 millimetre in diameter that extends a few kilometres below and about 20 kilometres above the main clouds.

The clouds completely envelop Venus from the equator to the poles, and from the dayside to the nightside. Moreover, they are generally similar at all locations except that the overlying haze is deeper and more dense near the poles. The dense and ubiquitous cloud cover has very important consequences; it is one of two factors responsible for the exceedingly high temperatures in the deep atmosphere of Venus near the planet's surface. The clouds are termed "optically thick," which means that they are opaque. Very little of the light that is incident on their upper surface ultimately emerges from the lower surface. This fact notwithstanding, there is enough light at ground level to permit photography, as demonstrated by the Venera 9 and 10 landers.

Three distinct families of cloud particles have been identified. Their chemical compositions are very different from that of Earth's clouds, which are largely liquid water and water ice. That the major material of Venus' clouds is concentrated (80 percent) sulfuric acid, H_2SO_4, was confirmed by in situ observations of sulfur dioxide and sulfur trioxide vapours from the evaporation of cloud droplets. (Their composition had been originally inferred from the optical properties of the liquid droplets as measured from Earth.) The larger acid droplets have a representative diameter of 0.008 millimetre. The haze also is mainly sulfuric acid. Another type of particle is up to 0.030 millimetre long; as measured by the Pioneer Venus particle-size spectrometer, this type of particle may be a tiny, solid crystal, probably in the form of a thin plate structure. The composition of these particles is not as well determined, but it may be a metallic chloride, perhaps iron chloride ($FeCl_3$), that was boiled out of the hot planetary surface and diffused up to higher and cooler altitudes, there condensing into crystals. The water vapour in Venus' atmosphere appears to be limited to that which is in equilibrium with the sulfuric acid droplets, which have a high affinity for water and scavenge it from the atmosphere.

By courtesy of A. Seiff

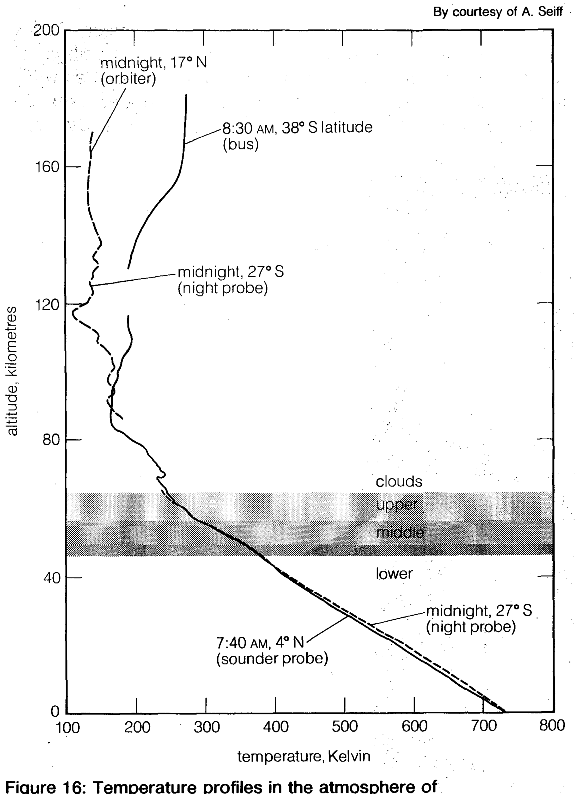


Figure 16: Temperature profiles in the atmosphere of Venus measured in 1978 by the Pioneer Venus mission. Below 100 kilometres, the atmospheric temperatures are amazingly uniform from day to night and over latitudes up to 60°. Above 100 kilometres, the night side becomes much cooler than the day and is not far above the condensation temperature of CO_2.

THE ATMOSPHERE

Carbon dioxide

Composition. The Venera 11 and 12 and Pioneer Venus probes of 1978 each carried two instruments to measure composition of the atmosphere. They directly measured nitrogen (N_2), which, at 3.5 percent, is a minor constituent of the atmosphere, and carbon dioxide (CO_2), which, at more than 96 percent, comprises the bulk of the atmosphere. (Although the percentage of nitrogen is small, its total amount in the massive atmosphere of Venus is equivalent to four times the nitrogen present in the Earth's atmosphere, where it is the major constituent.) Oxygen (O_2), the second most abundant gas on Earth, is a trace constituent of Venus' atmosphere and about one-fifth as plentiful as it is on Earth. Other gases detected by the sensitive instruments on the entry probes are present in trace quantities (see Table 6).

Water vapour measurements by different techniques have produced somewhat varied results but lie in the range of 0.1 to 0.5 percent, depending on altitude. The highest value was obtained a few kilometres below the sulfuric acid cloud, and the lowest values were recorded near the planet's surface. Above the clouds, the atmosphere is exceedingly dry because temperatures there are below the freezing point of water. If there were 0.1 percent of water throughout the atmosphere, and if the water were condensed into liquid on the surface, it would form a layer 10 centimetres deep over the entire planet. (Of course, liquid water cannot exist on the surface at the high temperatures found there.) Compared to the quantities of water in the Earth's vast oceans, the amount found on Venus is negligible. Investigators have proposed two possible explanations for the lack of water. The first states that Venus was formed with lesser amounts of volatile materials, including water, because of its position nearer the Sun, where temperatures during the formation of the planets were high. According to the second theory, Venus originally had much larger quantities of water; this water was vaporized by high surface temperatures and then dissociated into hydrogen and oxygen by sunlight at upper levels, the hydrogen escaping into space and the oxygen diffusing downward and reacting chemically.

Sulfur dioxide

The sulfur dioxide content of the atmosphere varies with altitude. It is greatest near the cloud base, where the cloud droplets vaporize and decompose into water and sulfur oxides, which are then carried upward where lower temperatures permit them to recombine and condense into droplets that grow to the maximum size found in the middle cloud. The droplets then settle to lower altitudes, where they vaporize and repeat the process. This cycle is somewhat analogous to the water-vapour condensation cycles in cloud formation on the Earth.

Hydrochloric and hydrofluoric acid vapours, measured spectroscopically from the Earth, are present above the clouds in trace amounts that may offer additional evidence of chlorides in the large particles of the clouds.

An important finding of the Pioneer Venus mass analyzers was that the ratio of abundance on Venus of the isotopes argon-36 to argon-40 is 300 times greater than it is on the Earth. This implies a significant difference in the chemistry of planetary formation because of Venus' location nearer the Sun.

Structure. Temperatures in the lower atmosphere of Venus increase continually from the cloud tops to the surface, where it is a very hot 735 K (see Figure 16). This is well above the melting temperatures of lead and tin, and above the boiling point of mercury (although it has not been detected, mercury could exist as a trace constituent vapour in the deep atmosphere). The first indications of high atmospheric temperatures on Venus came in 1956 from Earth-based measurements of intense microwave radiation. Interpretation of this radiation was at first varied, including explanations other than high temperatures. Spacecraft measurements, beginning with Venera 5 and 6 in 1969, however, have shown conclusively that the deep atmosphere is indeed very hot.

Temperature structure of the atmosphere was measured with great precision by the Pioneer Venus probes. They found remarkably similar temperature profiles at the four probe sites, at latitudes from the equator to 60°, and from the morning sector to the midnight meridian (see Figure 16). Temperature differences at a given altitude were much smaller than they are at such widely different locations on Earth. This is understood to be a consequence of the great thermal inertia of the massive deep atmosphere, coupled with a rapid east-to-west wind that moves the atmosphere at higher levels rapidly around the planet. This motion tends to equalize temperatures on the nightside and dayside.

High atmospheric pressures

High temperatures are accompanied by equally impressive pressures, about 94 atmospheres at the planet's mean surface. This pressure signifies a huge mass of gas, principally carbon dioxide, in the atmosphere. The mass ratio of Venus' atmosphere to Earth's atmosphere is the ratio of surface pressures divided by the ratio of gravitational accelerations, or $94/0.904 = 104$. To find a pressure on Earth equal to the pressure at the surface of Venus, it is necessary to descend into the oceans to a depth of one kilometre.

Earth-like conditions in the atmosphere of Venus are found in the clouds. For example, at 50 kilometres altitude above the planet's surface, the pressure is one atmosphere and the temperature is 350 K. Temperatures comfortable for human beings, around 295 K, are found near the top of the middle cloud at about 55 kilometres altitude, where the pressure is about half an atmosphere.

The thermosphere

Above the clouds, temperatures of the atmosphere continue to decrease to a minimum level of about 165 K in the Venus stratosphere. A nearly isothermal (constant temperature) region then extends from about 85 to 110 kilometres. Above this level lies the Venus thermosphere, the region of heating by absorption of very shortwave ultraviolet radiation from the Sun. On Earth, the thermosphere attains temperatures of 1,000 to 1,500 K. Venus is surprisingly different. Because its principal gas is carbon dioxide, an efficient radiator, it re-emits the absorbed energy at infrared wavelengths, so that the peak midday temperature in the thermosphere is about 300 K, or about room temperature on Earth. On the nightside, where solar heating is absent, the temperature drops to as low as 130 K, not far above the solidification point of carbon dioxide at these altitudes. Reasons for the cold nightside upper atmosphere are not yet understood in full detail.

Evolution. It is interesting to consider the interrelated questions of how the atmosphere of Venus evolved to its present state; why it is so different from the atmosphere of Venus' "sister planet," Earth; and why the deep atmosphere of Venus is so hot. It is believed by most astronomers that the planet's massive carbon dioxide atmosphere was released from carbonate rocks under the action of high temperatures. Conversely, the maintenance of high temperatures depends on the mass of the atmosphere and the presence of large quantities of carbon dioxide gas, which, in the "greenhouse effect," absorb and trap heat radiated up from below. Hence, an evolutionary progression is suggested in which pressure and temperature each act to increase the other over a long time period comparable to the age of the solar system.

The conspicuous external difference between the Earth and Venus is that Venus is 0.72 times closer to the Sun. This implies that 1.91 times as much solar energy is received by Venus and, for equal reflectivity, also implies surface temperatures greater by about 345 K. At the present time, the reflectivity of Venus' clouds is very high—such that 80 percent of the Sun's energy is reflected back into space—and Venus therefore absorbs less energy from the Sun than does the Earth. Before the cloud cover formed, however, Venus was probably about 345 K warmer than the Earth is now. Whether this comparatively small difference in temperature was enough to set in motion the chain of events that led to the current massive atmosphere of Venus is uncertain. It is interesting, however, that Venus, Earth, and Mars exhibit a progression in which the amount of atmosphere decreases as the distance from the Sun increases.

The greenhouse effect

It does seem clear that the present high temperatures in Venus' deep atmosphere are the result of the radiative equilibrium usually termed the greenhouse effect. The small amount of solar heat that penetrates to the deep atmosphere is conserved by the high infrared absorption of the carbon dioxide atmosphere; water vapour and the clouds and haze particles close the transparent "windows" in the carbon dioxide spectrum through which heat could escape. There is an analogue to the greenhouse effect in everyday experience on Earth—the warming that accompanies a cloudy night sky, when the clouds trap the Earth's radiant heat and send it back down, and the association between clear skies and cold nights. More precisely, a radiative balance occurs at every level when the amount of downward-directed solar radiation that is absorbed is equal to the amount of infrared radiation that is emitted upward. When local temperatures satisfy this balance, the atmospheric temperature is maintained. Radiative balance calculations using the data on chemical composition, haze particles, clouds, and water vapour as measured by the Pioneer Venus and Venera space probe experiments can generate a temperature profile that closely corresponds with the temperatures directly measured by the probes.

Atmospheric circulation and winds. Another method by which heat can move upward in the atmosphere is by convection, in which warm gas rises from below and cool gas descends to a nearby location. The Pioneer Venus temperature data showed that convection occurs far less generally on Venus than had been supposed. In fact, only a shallow layer two or three kilometres in depth in the clouds and another layer between 20 and 30 kilometres from the surface overturn in the way the Earth's atmosphere overturns beneath cumulus clouds. A third convective layer may exist near the surface. The remainder of the atmosphere, including one deep layer from 28 to 52 kilometres from the surface, is stable; *i.e.,* it carries heat upward only by radiation and not by motion of the atmosphere.

The four-day circulation

There are, however, general motions of the atmosphere of global scale, in which the main winds blow from east to west at velocities near the cloud tops of about 100 metres per second. These winds, which transport the clouds and circle the planet in about four days, are commonly referred to as the four-day circulation. This circulation was discovered by observation of a four-day repetition of patterns in the clouds that are visible in ultraviolet photographs. The first such photographs, taken through telescopes from Earth, only dimly show the clouds' features, but many such photographs taken from the Mariner 10 and Pioneer Venus spacecraft show the patterns more clearly (see Figure 15). The rapid rotation of the atmosphere at cloud altitudes efficiently transports solar heat from the dayside, where it is absorbed, to the nightside, so that, below 100 kilometres, temperature differences from day to night are no greater than a few degrees Celsius. Because the winds are ultimately driven by solar heat absorption (which occurs mainly in and above the clouds), it is not surprising to find the highest wind velocities in and just above the clouds. Below the clouds, wind velocities decrease progressively and steadily to a final level of one or two metres per second near the surface.

There are also relatively low-velocity north–south winds in the clouds. These winds perform the important function of transporting solar heat from the equatorial region toward the poles and returning the cold polar air to the equatorial region, where it is warmed. In the case of Venus, the vigorous equatorial–polar pattern of circulation is located in the clouds, whereas on Earth such circulation patterns are driven by surface heating and begin at the surface. The north–south winds of Venus limit latitudinal

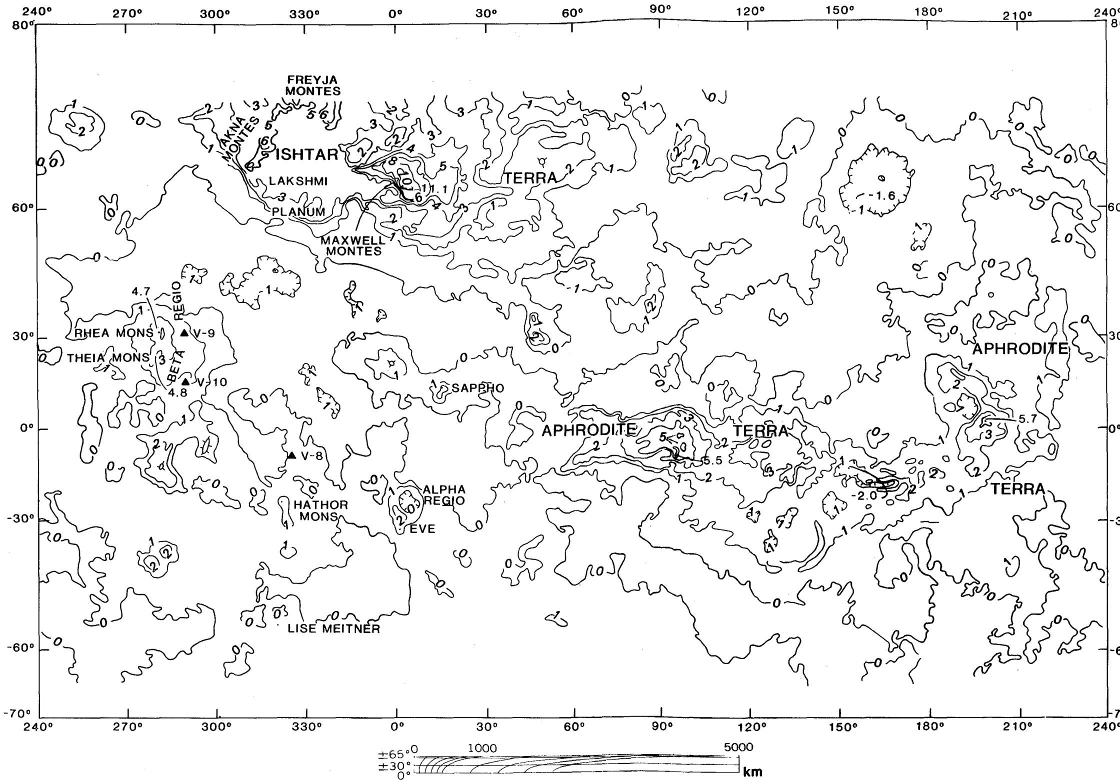


Figure 17: Surface contour map of the planet Venus, compiled in June 1980. Contour lines are described at intervals of 0.5 and 1 kilometre.

By courtesy of H. Masursky, U.S. Geological Survey, Ames Research Center, NASA; Massachusetts Institute of Technology

temperature differences, just as the four-day circulation limits the day-night differences. It is believed that other north-south circulation patterns that efficiently distribute heat from equator to the poles are present in Venus' deep atmosphere, possibly supplemented by large-scale horizontal eddies that are similar to cyclones on Earth.

THE SURFACE

Topographical features. Although clouds continually enshroud Venus and shield its surface from telescopic observation, the planet's surface has been examined by the longer waves of radio telescopes and of the Pioneer Venus radar altimeter. A radar map of almost the entire surface has been produced, showing the elevations of the surface; the presence of mountains, volcanoes, highlands (continents), and lowlands; large circular features similar to impact craters; and rift-valley features, one of which is 9,000 kilometres long, that are similar to the comparably immense Valles Marineris on Mars. The range of surface elevation extends from a high peak 11 kilometres above the mean planetary radius (6,051.5 kilometres) to low points 3.5 kilometres below the mean surface. The acceleration of gravity at the mean surface of Venus is 8.869 metres per second per second, or 0.90 times that on Earth.

Radar also measured the roughness of the terrain and found that the mountains are rough (signifying large rocks, steep surfaces, and rapid change in slope), while the plains are relatively smooth. Venera 9 and 10, after landing on the surface, returned pictures of their surroundings. The field around the Venera 9 landing spot is strewn with large rocks, while that around the Venera 10 landing appears to be dominated by an immense boulder outcropping, approximately flush with the surrounding surface. Venera 8 measured the composition of surface materials and found them to be granite-like in a region where the planetary crust was undisturbed. Venera 9 and 10 found basaltic composition in a region near Beta Regio, which is believed to be volcanic and to have experienced lava flows that left materials on the surface similar to those around volcanoes on Earth and the Moon.

Surface conditions. A human being standing on the surface of Venus during the daytime, if he could survive, would notice that the high clouds admit some scattered sunlight to the surface, much as do clouds on a heavily overcast day on Earth. The time from one sunrise to the next would be equivalent to about 118 Earth days. Because of the extremely high surface pressures on Venus, light from the bright side of the planet would be refracted through the lower atmosphere around toward the dark side, and some scattered light would be available even at night. The refraction effect is so strong in such a dense atmosphere that during the daytime it would be possible to see beyond the true horizon, and as a level surface on Venus would, for the same reason appear concave, the viewer would seem to be in a depression that remained around him, no matter how far he moved across it.

Effects of refraction on Venus

Radar and microwave observations imply that the surface of Venus is granular on a large scale, suggesting a rock-strewn surface. Dust storms may occur because of the high atmospheric density, even with the low velocities found near the surface. Erosion by windblown dust could modify the terrain features on Venus. Such erosion, however, does not affect the large-scale surface features seen on radar maps of the planet. In addition to apparent mountain ranges, large circular features that may be of impact origin and others analogous to the lunar maria have been detected by radar techniques. Fine dust should tend to accumulate in the lowlands, and many of the large craters may be filled in with dust.

The surface temperatures are so high that liquid water is excluded everywhere. In shadows, objects probably could be seen to glow by their own red heat. Temperatures everywhere on Venus are much too high to permit terrestrial forms of life. The temperatures are, in fact, so high that most organic molecules would denature rapidly.

The existence of some forms of life, however, for example in the vicinity of the clouds, is not impossible, although at the present time it can only be conjectured.

Astonishingly hot, with an oppressively dense atmosphere containing corrosive gases, with a surface glowing dimly by its own red heat and characterized by bizarre optical refraction effects, Venus (curiously identified in ancient literature with Lucifer) seems very much like the classical view of hell. (C.Sn./A.S./Ed.)

MAGNETIC FIELD, IONOSPHERE, AND SOLAR WIND INTERACTION

Lack of a magnetic field

The earliest spacecraft to fly by Venus, Mariner 2 and 5 and Venera 4, found no measurable intensity in the planet's magnetic field. Data from the Pioneer Venus orbiter place the strength of the magnetic field at less than 0.00001 gauss. The data do not, in fact, find any real evidence for an intrinsic magnetic field; *i.e.,* one associated with the solid planet. There are very weak fields associated with the ionosphere and with the solar wind that are generally less than 0.001 times the strength of the Earth's surface magnetic field. The absence of a planetary magnetic field on Venus is consistent with the internal dynamo theory of field generation, which ascribes the existence of planetary magnetic fields to internal current systems driven by planet rotation, generally similar to electric generators. Since Venus rotates slowly, once in 243 days, its dynamo is weak.

The ionosphere of Venus, as on other planets, is created primarily by the absorption of solar shortwave (extreme ultraviolet) radiation. On Venus, the impact of solar wind protons and electrons on neutral molecules in the upper atmosphere may also contribute. The absorbed energy can result in an electron escaping from the molecule, leaving a positively charged ion and a free electron. Because of Venus' proximity to the Sun, these processes occur somewhat more frequently than on more distant planets. The number of free electrons in Venus' ionosphere, however, is fairly typical of that found in the ionospheres of other planets out to Jupiter—about 100,000 to 300,000 per cubic centimetre, depending on distance from the subsolar point and on altitude. The number of ions per cubic centimetre increases with decreasing altitude down to 160 kilometres above the surface, but theory indicates that it should peak at about 130 kilometres. The major ions are O_2^+ and CO_2^+. The density of ions and electrons is much lower on the nightside than on the dayside, generally less than 10,000 per cubic centimetre. The upper limit of ionosphere, called the ionopause, is sharply defined, but it is variable from about 300 kilometres altitude on the dayside to about 1,000 kilometres on the nightside. The dayside boundary is imposed by magnetic pressure from the solar wind that impinges on the upper atmosphere.

Bow shock wave

Venus—like Earth, Mars, Jupiter, and Saturn—is enveloped by a bow shock wave, which forms in the supersonic flow of the solar wind about the planet. Because Venus lacks a magnetic field, its bow wave stands closer to the planet, only about one-third of the planet radius away at the subsolar point. This is because the solar-wind particles are reflected directly by the ionosphere, rather than by the planetary magnetic field, which extends much farther out in space for Earth, Jupiter, and Saturn. The

By courtesy of the National Aeronautics and Space Administration

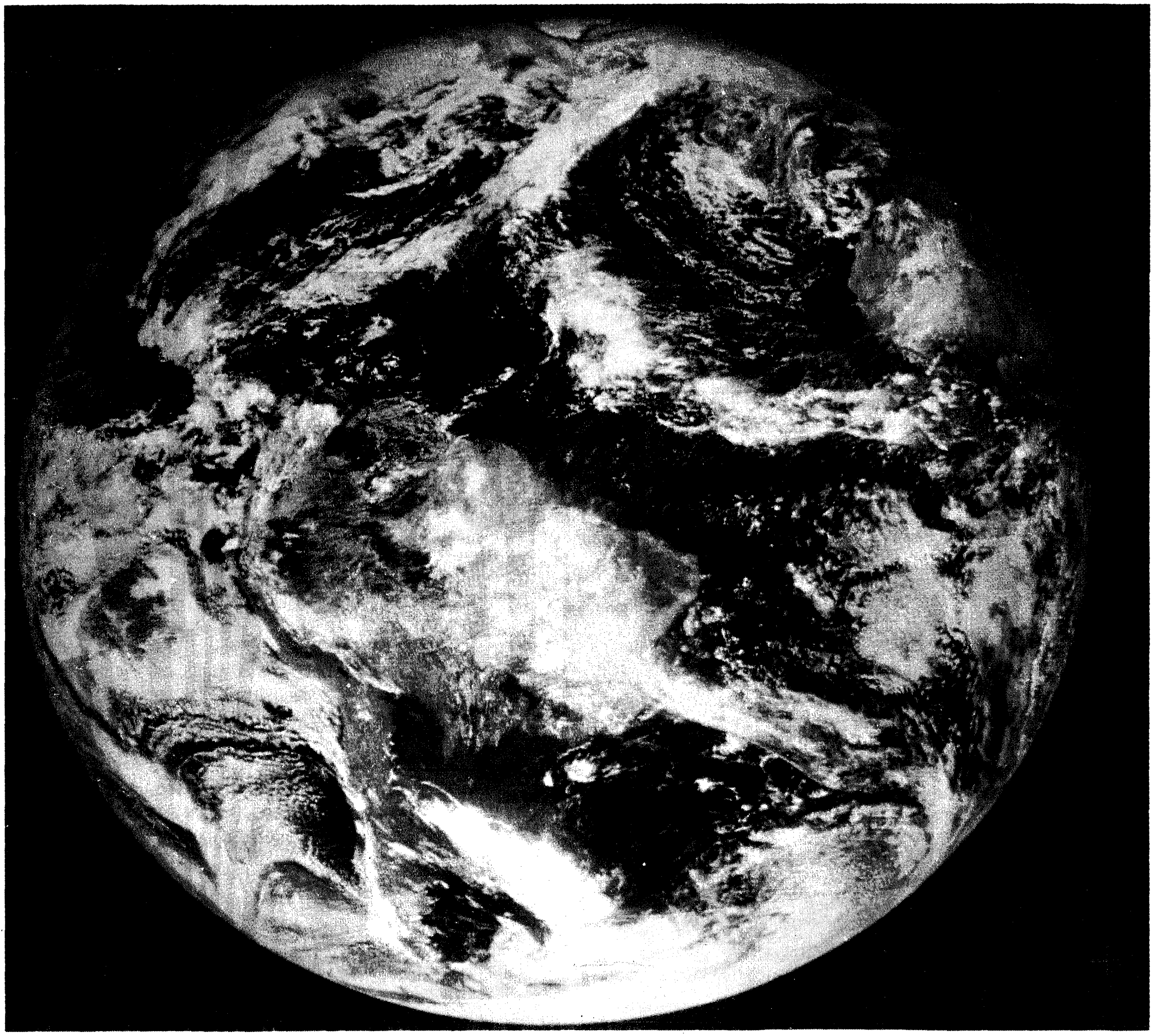


Figure 18: *The Earth from 35,900 kilometres in space.*
The National Aeronautics and Space Administration's ATS III satellite transmitted this colour photograph to a ground station at Rosman, North Carolina. It was received at 10:30 AM EST on November 10, 1967, from its on-station position at 47° west longitude on the Equator over Brazil. Four continents (portions of North and South America, Africa, and Europe and the Greenland ice cap) can readily be seen, while the Antarctic continent is blanketed under cloud cover. Major weather over the central United States, stretching from the Great Lakes to Mexico, represents a cold front moving eastward. A tropical storm (bottom centre) with a cold front extending into Argentina can be seen.

bow shock wave is basically similar to the wave formed about projectiles and rockets in supersonic flight. In fact, a sphere fired from a gun at a Mach number of 2 or 3 has a shock wave closely similar in appearance to the bow wave surrounding Venus. (A.S.)

Earth

Comparison with other planets

The Earth, symbol ⊕ in astronomy, is the third planet in distance outward from the Sun. Many consider the Earth–Moon system to be a unique double planet because the Moon is much larger in comparison with its primary, the Earth, than most other planetary satellites. In the more usual view, the Earth has a single natural satellite, the Moon. Like some other planets of the solar system, the Earth has radiation belts, an ionosphere (part of the atmosphere containing electrically charged particles, or ions), an atmosphere, and a lithosphere, or crust (outer surface); beneath the crust, the deep interior is divided into layers called the mantle and the inner and outer core (see Figure 19). Much more is known about details of the Earth's structure and composition than about those of the other planets because of its accessibility. The Earth has, however, been much changed and differentiated locally because of its active and turbulent history, and, thus, care is needed in making interpretations. Processes such as volcanism, mountain building, and seafloor spreading are still going on.

The curvature of the Earth can be demonstrated from ground level in many simple ways—for example, by measurements of Eratosthenes of Cyrene in antiquity (see below). Photographs from rockets in the 1950s clearly showed the curvature of the edge of the Earth (at the horizon), as did later photographs taken from the space vehicle Gemini 4. In the 1960s, for the first time, the Earth was photographed from sufficient distance to show it as a planet, more beautifully coloured and more varied than any other. (See Figure 18.) In this picture, cloud patterns show clearly the scale of atmospheric circulation. Below the clouds, the oceans appear blue, and the outlines of continents can be seen. The Earth's oceans are probably unique among the planets because of the small range of temperatures and pressures within which water exists as a liquid.

Like other planets, the Earth is approximately a sphere. It spins on its axis (the line joining the points at the surface called the North Pole and South Pole) once a day relative to the Sun. This is equivalent to once in about 23 hours 56 minutes 4 seconds relative to the stars (the sidereal day). The sidereal day is shorter than the solar day (24 hours 3 minutes 57 seconds) because the Earth, revolving at the same time about the Sun, must rotate farther, after completing the sidereal day, before the solar day is completed.

This section presents an overview of the Earth as a whole, ignoring the details of its surface features, the oceans, and the atmosphere. Those topics are treated in the articles CONTINENTAL LANDFORMS; OCEANS; and ATMOSPHERE. For additional information on the Earth's interior, surface configuration, physical properties, and chemical composition, see EARTH, THE and CHEMICAL ELEMENTS. The history of the Earth is summarized in the article GEOCHRONOLOGY, and the origin and nature of the Earth's major surface features are treated in GEOMORPHIC PROCESSES; RIVERS; and PLATE TECTONICS.

BASIC PLANETARY DATA

Figure of the Earth. The radius from the centre to the poles is not quite the same as that from the centre to a point on the Equator (see Table 7), but the concept of the Earth as a sphere is a good approximation. The distance a traveller has to proceed northward in order to make the elevation of the Pole Star increase by one degree is the length of one degree of latitude. The length of a degree of latitude depends slightly on the latitude at which it is measured. This dependence indicates that there is a deviation of the Earth from spherical shape. Sir Isaac Newton was the first to demonstrate that the centrifugal force of the Earth's rotation would tend to make the Earth bulge out at the Equator, giving it the shape of an oblate spheroid (the surface formed by an ellipse when it is rotated about its short axis). Such a surface is flatter near the pole than near the Equator, and one degree of latitude is therefore longer in high latitudes than it is in low ones. Gravity also varies over the Earth's surface, primarily with latitude, being somewhat higher by about 0.5 percent at the poles than at the Equator.

Table 7: Earth Data

Equatorial radius	6,378.136±1 km (3,963.374 mi)
Polar radius	6,356.784 km (3,949.921 mi)
Mean radius	6,371 km (3,959 mi)
Flattening	1/298.257
Eccentricity	0.017
Circumference, equatorial	40,075.51 km (24,902.45 mi)
M (mass)	5.976×10^{27} g
V (volume)	1.083×10^{27} cu cm
Mean density (*M/V*)	5.517 g per cu cm
Mass ratios: Sun to Earth to Moon	333,400 : 1 : 0.01228
Mean distance from Sun	1.496×10^{8} km (92,900,000 mi)
Earth's speed in its orbit	29.8 km/sec (18.5 mi/sec)
Revolution around Sun	365.25 days
Rotation	23 hr 56 min 4.09 sec
Area	5.096×10^{8} sq km (197,000,000 sq mi)
Land	1.48×10^{8} sq km (57,000,000 sq mi; 29% of Earth)
Highest point	8,848 m (29,028 ft; Mt. Everest)
Lowest point	−11,034 m (−36,201 ft; Marianas Trench)
Lowest point on land	−397 m (−1,302 ft; Dead Sea)

Factors in the changing length of the day

Rotation and precession. Because the Earth revolves around the Sun in 365.25 days, the sidereal day (the exact length of which is 23 hours, 56 minutes, 4.099 seconds of mean solar time) is shorter than the conventional mean solar day by about one part in 365. Astronomical observations show that the Earth is not an ideal timekeeper. The length of the day is subject to certain tiny changes. The most important are (1) a gradual slowing down of the Earth's rotation—*i.e.,* lengthening of the day; (2) small annual or seasonal variations; (3) small irregular variations. The magnitude of the gradual slowing down found from ancient astronomical observations, especially from ancient eclipses and from the more accurate astronomical data available since about 1600, amounts to an aggregated time change of about 30 seconds per century. If this phenomenon is assumed to have existed throughout the Earth's lifetime, the Earth must once have rotated much more rapidly than it does now. The cause of the slowing down has not been established with complete certainty, but most probably it is due to the effect of tidal friction in the oceans. The loss of angular momentum (that is, momentum acquired through rotation) by the Earth is compensated by that gained by the Moon's orbit, which then moves farther away from Earth. The seasonal fluctuations in the length of the day are tiny, on the order of some thousandths of a second per year; they can be explained by changes in the size of the polar ice caps and changes in atmospheric circulation. The irregular changes are equally tiny and can apparently be traced to motions in the Earth's core.

Lacking any external influences, the direction of the Earth's axis of rotation in space would remain the same for all time. Gravitational forces of other members of the solar system, in particular the Moon and the Sun, however, influence the Earth. Because the Earth is flattened, the Moon's gravity tends to tilt the Earth's axis, so that it becomes perpendicular to the Moon's orbit, and to a lesser extent the same is true for the Sun. A spinning body, such as the Earth or a gyroscope, subject to such a force will not, however, yield to it, but the axis of rotation will precess—*i.e.,* move around the surface of a cone in space. This motion of the Earth's axis is known as the precession of the equinoxes (times of equal day and night in spring and autumn). It was discovered by the astronomer Hipparchus in 120 BC and first theoretically explained by Sir Isaac Newton in the 17th century. Precession implies that the pole is not always in the same place among the constellations but moves about; for instance, 12,000 years hence it will coincide with the star Vega. The length of one period of the precession is 26,000 years. The precession makes the Sun's orbit in the heavens (the ecliptic) and its intersection with the celestial equator (the equinoctial points) move in such a direction as always to meet the Sun, at a rate of 50.26″ per year. This change amounts to about 29° since Hipparchus' time; hence the ancient Greek's description of the Sun's yearly path across the constellations no longer fits. The force tending to alter the tilt of the axis is not constant; when the Sun or Moon is crossing the Equator, for instance, it has no effect. The path of the pole is therefore not exactly a circle but contains a superimposed smaller oscillation, the nutation, with a period of 18.6 years and an amplitude of 9.2 seconds of arc. The Chandler wobble, another gyroscopic effect of extremely small amplitude and a period of 436 days, is accompanied by a slight deformation of the Earth's body. The solid body of the Earth undergoes a slight daily tidal deformation caused by the gravitational action of the Moon and Sun, about 25 centimetres at most.

The Chandler wobble

Numerical data. Some quantities relating to the Earth are known with a precision exceeding one part in 1,000,000. Others are known only to within a much lower error limit, and those derived from the latter have corresponding uncertainties. The one most important quantity of the former type is the constant of universal gravitation, $G = 6.67259 \times 10^{-11}$ m^3 kg^{-1} s^{-2}. The numbers with many digits given in Table 7 might someday have to undergo revision in the last one or two decimal places, but they are internationally agreed upon as the most likely values.

STRUCTURE OF THE INTERIOR

Data from seismic waves. Two phenomena provide information about the deeper layers of the Earth's interior: seismic waves generated by earthquakes and, to a lesser degree, the Earth's magnetic field. There are two main types of seismic waves, the P-waves (primary, because they travel fastest) and the S-waves (secondary, because they travel more slowly and, hence, arrive later at the station than the P-waves). The P-waves are compressional (longitudinal) in nature, whereas the S-waves are shear (transverse) waves. The existence of two types of waves greatly increases the amount of information about the Earth's interior.

Types of seismic waves

The most conspicuous feature of the Earth's interior is the existence of a sharp boundary at which reflection and refraction occur and which encloses the part known as the core; it represents an extremely well-defined discontinuity. The mean radius of the core (ignoring flattening) is 3,470 kilometres, or nearly 55 percent of the Earth's radius. The most outstanding feature of the core is that it never transmits S-waves. Because the failure to transmit transverse waves is a distinctive characteristic of liquids, it is generally accepted that the core material is in a liquid state, probably mainly of molten iron. Existence of the core was first quantitatively established in 1913. Later studies have shown an additional boundary, somewhat diffuse, which encloses the so-called inner core. Its radius, not precisely known, is about 1,300 kilometres. Evidence indicates that the inner core is solid iron (solidified under extreme pressures).

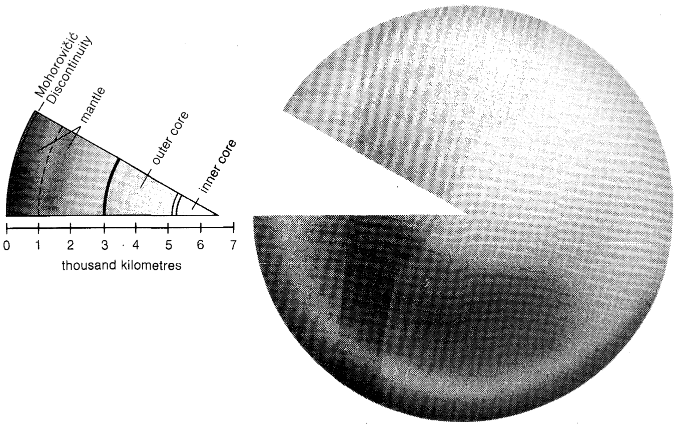


Figure 19: Main layers of the Earth.

Seismic velocities, density, pressure. In the mantle and the core a progressive stiffening of the material with increasing pressure makes the velocity of seismic waves increase at greater depths. The density distribution cannot be uniquely determined from such data, but certain rather restrictive conditions are imposed on it thereby. A further condition results from a knowledge of the moment of inertia of the Earth about its axis, known with great accuracy from a combination of astronomical and geodetic observations. The moment of inertia about an axis through the centre of a homogeneous sphere, for which *m* is the total mass and *a* is the radius, is 0.4 ma^2. Observations show that the moment of inertia of the Earth is 0.334 ma^2, which indicates a rather pronounced increase of density toward the central parts.

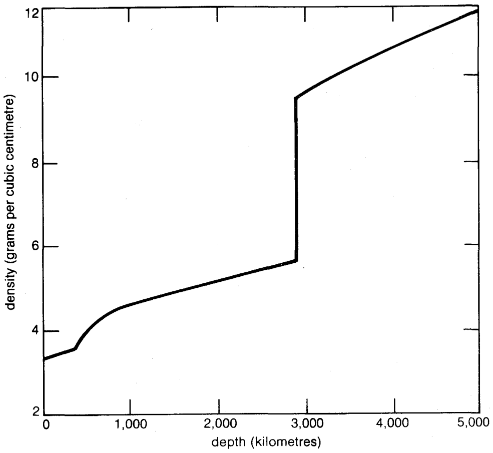


Figure 20: Known density distribution in the Earth down to the limit of the inner core.

The pressure inside the Earth rises rather regularly with increasing depth. In the mantle the rise is at a rate of about 470 atmospheres per kilometre (one atmosphere of pressure equals 14.7 pounds per square inch). The pressure at the boundary of the core is 1,370,000 atmospheres; at the centre of the Earth it is 3,700,000 atmospheres. The strength of gravity is also a function of the depth; it clearly must vanish at the centre. Gravity is remarkably constant throughout the mantle, to within 1 or 2 percent; but, beginning at about the boundary of the core, it decreases steadily to zero.

ORIGIN, COMPOSITION, AND AGE

Formation of the Earth. Astronomers have speculated about the origin of the solar system ever since it was suggested in the 18th century that the stars and planets arose from the condensation of vast gaseous masses filling cosmic space. The formation of the Sun may have occurred much earlier than that of the planets. It is now generally accepted that the planetary system developed out of a disk-shaped mass of gas surrounding the Sun. The original cosmic matter consisted mainly of hydrogen (and some helium), and most of these gaseous components must have been lost to space early from the disk. The less volatile matter condensed, mainly as silicates and metallic iron, these being the prime constituents of present-day meteorites as well as of the interior of the Earth. At the time the gas escaped, the solid matter probably consisted of bodies much smaller than the present planets, because the larger bodies would hold on to a gaseous envelope by gravitation, at least to some of the gases heavier than hydrogen. Most probably, the original solid components were dust grains.

Origin of the planetary system

While this picture is crude and speculative in its details, it brings out one crucial point common to all modern theories as distinct from the views held in the 19th century: the original matter from which the planets condensed must have been relatively cold. The temperature in this primary lens-shaped body of gas and dust was probably quite low, because it should readily have been cooled by radiation that could escape in a direction perpendicular to the plane of the disk. When dust congregated to form the Earth, gravitational energy was released and could be transformed into heat, but it is not known what fraction was available for actual heating, because this amount depends critically on the speed of aggregation. If the aggregation was slow, most heat should have escaped by radiation, and this is the most likely case.

Chemical composition. The composition of the Earth's crust can be investigated directly by quantitative chemical analysis of samplings from many rocks. The pioneer in this field was Victor M. Goldschmidt, who in the early decades of the 20th century undertook the painstaking task of determining the average proportions of the chemical elements in the Earth's crust. To appreciate the significance of the results it is necessary to compare them with what is known about the chemical composition of cosmic objects other than the Earth. Among the objects that fall upon the Earth from outer space are meteorites. Meteorites are generally considered to be fragments of a larger solar system body or bodies. Many chemical analyses of meteorites have been made. Next, quantitative determinations of the relative proportions of the chemical elements can be made for the Sun by spectroscopic methods. Finally, though with greater limitations, the same can be done for other stars.

One conclusion long drawn from the vast mass of observations is that the gross chemical composition of matter in the universe is relatively uniform. There might be small variations from one star to the next, but the impression of overall uniformity is strong. Scientists are thus led to think of a definite gross composition of primordial matter from which the Earth and the other planets were formed. The figures given below for the relative proportions (known as cosmic abundances) of the most important elements are from a classical compilation by H. Suess and Harold C. Urey (1956). Since the abundances are only relative, they are expressed as the numbers of atoms existing per one atom of a reference substance. The latter is taken to be the element silicon, since rocks consist largely of compounds of silicon. These abundances are as follows:

hydrogen	40,000.0	aluminum	0.09
helium	3,100.0	silicon	1.00
carbon	3.5	phosphorus	0.01
nitrogen	6.6	sulfur	0.37
oxygen	21.5	argon	0.15
neon	8.6	calcium	0.05
sodium	0.04	iron	0.60
magnesium	0.91	nickel	0.03

Remarkably enough, only six of the remaining elements have abundances that lie between 0.001 and 0.0001; all other elements are so rare that they exist only in minutest traces on the cosmic scale.

Location of concentrations of elements in the Earth's layers

On the whole, the lighter compounds are more concentrated in the crust, including oxides and more complex compounds of sodium, magnesium, aluminum, and calcium that form many important constituents of common rocks. Geologists are able to discriminate between rocks of the surface layers and others that have been pushed up from lower layers and should therefore resemble more closely the chemical composition of the Earth's mantle. On the basis of such data it is commonly believed that the Earth's mantle consists primarily of certain compounds (silicates) containing oxygen combined with silicon, magnesium, and iron, three of the most common nonvolatile elements. Sulfur most probably exists in the interior of the Earth in the form of iron sulfide. Geochemical evidence, however, indicates that sulfur is far less abundant on the Earth than the overall cosmic abundance figure given above.

Composition of the core

It has long been held that the Earth's core is composed of iron. Iron is by far the heaviest of the common nonvolatile elements; hence, it would tend to sink toward the more central parts of the Earth. Recent studies indicate, however, that, while the inner core consists of solid iron in pure form, the outer core is probably composed of molten iron alloy. Profiles of seismic-wave velocities and densities through the core strongly suggest the presence of certain contaminants or alloying elements such as oxygen and sulfur. The cosmic abundances of these particular elements, along with other geochemical considerations, all tend to point to the occurrence of iron oxides and sulfides in the outer core.

Radioactive age determinations. Geologists early arrived at minimum figures of well over 100,000,000 years for the main part of geological history. The discovery of radioactive decay around 1900 provided an entirely new and much more reliable method for dating events in the Earth's history. Radioactivity consists of the transformation of one species of atomic nucleus into another. In the case of naturally radioactive substances, this transformation occurs spontaneously, and no chemical or physical condition of the environment has any effect on the rate of radioactive decay. In any given interval of time a fixed fraction of the atoms of the mother substance decays into atoms of the daughter substance. Radioactive decay, therefore, furnishes a natural and trustworthy type of clock wherewith to measure the distant past. Only the most long-lived of radioactive substances are useful for dating earlier periods in the history of the Earth.

The age of the Earth

Attempts have been made to estimate the total age of the Earth under certain plausible assumptions, using data on the relative proportions of lead isotopes in very old rocks and in meteorites. The age arrived at is about 4,500,000,000 years, a figure now well established and believed to be accurate to within a very small percentage of possible error.

MAGNETISM AND TEMPERATURE

Characteristics of the Earth's magnetic field

Earth's magnetic field. The Earth's magnetic field has two outstanding characteristics. The first is its irregularity, as seen for instance in the variation of the magnetic declination from place to place, the overall pattern of which is quite complicated. The second is its gradual change in time, a phenomenon well-known to all navigators because the declination of the needle changes continuously. Magnetic maps indicating the declination are seriously in error after a few years and must be redrawn periodically from new observations. This change is known as the secular variation of the Earth's magnetic field. If a variation of 10′ per year continues over, say, 60 years, it corresponds to the considerable change of 10° in declination. The pattern of change is, however, itself variable.

These rapid changes make it clear that the field does not originate in the solid mantle. The magnetic field is caused by electric currents that flow in the fluid, metallic outer core. The secular variation indicates that the fluid of the outer core is in motion. This motion carries the electric currents about and thus produces the variation of the field observed at the Earth's surface. The details of this variation, rather irregular, may be interpreted as representing so many large-scale eddies in the fluid. From the secular magnetic variation may be obtained a value for the average speed of the fluid iron: 0.03 centimetre per second or about 10 kilometres per year. Further development of these ideas leads to the conclusion that the fluid motion not only modifies the electric currents but also can produce and maintain them. This process resembles closely the one that takes place in a rotating generator of electric currents as found in power stations, in which current is produced by the motion of metallic wires past each other. In the core, streams of molten metal are the equivalent of the moving wires. This explanation of the Earth's magnetic field, known as the dynamo theory, has become accepted generally.

Rock magnetism. When a lava melt is solidified in a magnetic field and the field is removed later, the solid rock retains a slight amount of so-called remanent magnetism along the direction in which the original magnetic force pointed. The remanent magnetism is due to small grains of magnetic material (iron oxide) embedded in the rock. It becomes possible to infer from this the direction of the Earth's magnetic field at the time an ancient lava flow occurred. Similarly, sandstones have been shown to retain a small remanent magnetization along the direction of the field in which the material was deposited. The study of rock magnetism has been systematically developed, giving information about the character of the magnetic field at earlier times in the Earth's history. There seems to be little doubt that, during the last 500,000,000 years or so that are accessible to detailed geological study, the Earth has had a magnetic field similar to its present one.

Past magnetic reversals

Work on rock magnetism has shown, surprisingly, that apparently the magnetic field has reversed its polarity many times, perhaps hundreds of times during the past history of the Earth—that is, the magnetic south pole became the north pole, and conversely. A detailed explanation of this phenomenon is lacking; it can be said merely that its occurrence is conceivable in terms of the dynamo theory, because a self-excited generator can produce electric current in one direction as well as in the other, depending on its starting conditions. This phenomenon is also seen in some stars.

Temperature and heat transport. There is no direct way of determining the temperature distribution inside the Earth, but it is well-known that the temperature rises in the lower levels of deep mines. The rate at which heat is delivered to the surface is given by the rate at which the temperature rises (the thermal gradient), multiplied by the thermal conductivity, which is a characteristic constant of the material. Thus, by measuring the two last-named quantities, the heat flowing out of the Earth can be determined. The results of heat-flow measurements on the ocean bottom indicate that, even where the oceans are very deep, the heat flow is comparable to that flowing out on the continents. The average thermal gradient in the outermost layers of the Earth is about 25°–30° C per kilometre of depth. The average heat flow near the surface of the Earth is about 1.2×10^{-6} calories per square centimetre per second.

Assuming, say, average thermal gradients of 14° per kilometre at some greater depth, the temperature at a depth of 300 kilometres would be about 4,200° C, far above the melting point of any silicate, even at the pressures prevailing at such depth. Hence the temperature must level off; that is, the thermal gradient must radically diminish even at moderate depths inside the Earth. It long has been concluded that most of the heat flowing out of the Earth is produced by radioactivity (mostly from the decay of the uranium and thorium families) and that the radioactive material must now be strongly concentrated in the top layers of the Earth.

Source of heat

It was long believed that most of the radioactive material is in the crust, and because the crust is thick under the continents but thin or almost absent under the oceans, this occurrence would mean that most of the heat flow would occur in continental areas. Heat-flow measurements on the ocean bottom have shown that this view is wrong; it is now usually assumed that heat, especially under the oceans, is also transported by mechanical, convective motions.

In the deeper layers of the Earth it is necessary to rely largely on guesswork, but any assumed distribution is subject to certain conditions. It must remain below the melting point of the simple silicates throughout the mantle, it must be above the melting point of iron in the outer core, and if the inner core consists again of iron the temperature must rise to the melting point of iron at the boundary of the inner core and exceed it inside the latter. The best way to obtain the temperature inside the Earth is, therefore, to estimate the increase of melting point with increasing pressure. The existing estimates of the temperature in the Earth's interior tend toward relatively low values, perhaps in the neighbourhood of 3,000° C at the boundary of the core and near 4,000° C at the centre of the Earth. These estimates, however, may be in error by many hundreds of degrees.

ORIGIN OF THE SURFACE FEATURES

The surface of the Earth shows a remarkable degree of irregular structure. The surface levels cluster around two preferred heights: (1) sea level—most of the areas of the continents and of the offshore continental shelves lie near it—and (2) ocean bottoms, with an average depth of about five kilometres. Levels between these are rare. Another remarkable regularity has to do with the asymmetrical distribution of land and water masses. If a diameter is drawn arbitrarily across the Earth's centre, in more than 90 percent of all cases, one end is on land, the other on the sea. The fact that conclusive explanations, going beyond mere speculation, of such simple regularities cannot be given is indicative of lack of understanding of most of the basic physical and chemical processes that have shaped the surface of the Earth into its present form.

Isostasy. The crust considered as a load on the mantle is less dense; it is specifically lighter than the substratum. Such a load will sink only to a depth at which it is floating freely with respect to vertical displacements; it resembles an iceberg floating in the water with the larger part submerged beneath the surface. According to the principle of isostasy there is a tendency toward vertical equilibrium of the crust. Quantitatively speaking, the total mass encountered on going from the surface down toward the Earth's centre tends to be the same for any point of the surface. Where there is a mountain there seems to be more mass, but isostasy indicates that there must be a deficiency of mass farther down; the mountain has a root consisting of matter less dense than the matter lying under lowlands and the oceans. The root is bigger than the mountain itself; the lighter material extends down to about six times the mean elevation of the mountain.

This idea of isostatic (vertical) compensation was first developed on the basis of gravity observations. If ideal isostatic balance prevailed, gravity would be the same over the oceans and continents (apart from variation with geographical latitude). Actually, most mountainous elevations and most oceanic depths are fairly well compensated isostatically. Deviations, known as gravity anomalies, occur in the regions in which active mountain building occurs, such as the Pacific island areas (*e.g.,* the Japanese islands). Volcanic activity is concentrated in these same regions, as are earthquakes. Many earthquakes originate fairly near the Earth's surface, but in active regions earthquakes are observed to originate at depths down to 700 kilometres. Thus the upper mantle seems to participate to at least this depth in mountain-building.

Volcanism and earthquakes

Deformation of the crust: mountain building. Four agencies have been held responsible for mountain building, either individually or jointly, though it is probable that several (especially the last three) have been acting together.

Principal agencies of orogeny

1. *Contraction*—The Earth resembles a shrunken apple, the mountains and valleys being likened to the wrinkles of its skin. This is the oldest hypothesis, but since the discovery of radioactive heating of the Earth it has been difficult to believe that the Earth has shrunk appreciably in geological times.

2. *Convection*—If the mantle were a viscous fluid heated from below, convective upwelling would occur at a slow rate. If the thermal conditions are right, there must certainly be a tendency toward convection, though the mantle is a plastic solid and not a viscous fluid, and this might make actual patterns look quite different from those of a fluid.

3. *Phase change*—A change in the crystalline structure of minerals, which may occur as a result of either temperature or pressure changes is a phase change. Much evidence shows that at high pressures certain more compact forms of minerals that do not exist at the surface of the Earth are stable. It has been claimed that the Mohorovičić Discontinuity at the bottom of the crust represents such a phase change. Temperature might rise by radioactive heating and this rise could shift the discontinuity downward. Temperature falls again and the discontinuity rises if the heat is carried off by volcanism and other agencies. The tendency toward isostatic equilibrium might then produce fairly large vertical displacements.

4. *Accretion*—Because at increasing depths in the Earth the temperature first rises rapidly and then more slowly, the silicates of the mantle might be closer to melting at a depth of a few hundred kilometres than elsewhere. If melting occurs locally, the material will separate into its lighter components on top and its heavier below. Lighter material might gradually make its way through the overlying layers to the top, and in this manner the amount of crustal material can grow. Volcanism appears to be a process of this kind. The continental blocks of the crust may thus have only gradually accumulated through ge-

ological ages. Strong evidence exists that the ocean and the atmosphere are secondary products of accretion—that is, they have been slowly formed in the course of the Earth's history from material coming out of the rocks.

A more definitive theory of mountain building arose in the 1960s and 1970s. The discovery of the extent and nature of the Earth's mid-oceanic ridges and the fact that seafloor materials became older along paths normal to the ridges led to the concept of seafloor spreading—that new ocean floor is generated by the upwelling of the mantle along the ridge crests, that it subsequently moves laterally away from the ridges and toward the continents, and that it is eventually subducted or dragged downward back into the mantle along continental margins. These processes are thought to result in both continental drift and mountain building and are further discussed in the articles EARTH, THE; and PLATE TECTONICS.

(W.M.E./Ed.)

The Moon

The Moon, designated ☾ in astronomy, is a natural satellite of the Earth, but it is sometimes described as a planet because it is larger than most known natural planetary satellites and, because of its size, has many of the properties of a small planet such as Mercury. The Moon, like the Earth and the other terrestrial planets, has a solid body, nearly spherical in shape. The Moon's radius is about 1,738 kilometres, only 0.2725 that of the equatorial radius of the Earth; this difference, with the fact that the Moon's mass is only $^1/_{81.3}$ of the Earth's mass, is largely responsible for the difference between the atmospheres of the two bodies. The escape velocity of a particle from the surface of the Moon is about 0.213 that from the Earth, and, if past temperature conditions were much like present ones, the gas molecules and atoms of any early lunar atmosphere would have escaped in a time that is short when compared with the Moon's lifetime.

The Moon's mean density is about 3.34 grams per cubic centimetre, close to the density of the Earth's mantle; but studies of rocks returned from the Moon's surface indicate that the Moon's interior composition is not the same as that of the Earth's mantle and that in the Moon there is only a very small (possibly no) core. Many of the rock samples retrieved from the large, relatively smooth, dark areas of the face of the Moon called the maria are best described as basalts, but they are quite uncharacteristic of any known terrestrial rock or meteorite. They are about the colour of graphite—charcoal gray.

The whole Moon reflects less than one-tenth of the light that falls on it. The low reflecting power of the Moon had been known for a long time from estimates of its overall albedo. "Albedo" is a Latin word meaning "whiteness." While it is relatively easy to measure the relative brightness, or reflecting power (normal albedo), of flat surfaces illuminated and viewed normally in the laboratory, it is more difficult to assess the overall albedo of a spherical planet. The brightness of the Moon varies through its phases because of the roughness of its surface and the resulting variable amount of shadow. The Full Moon, when the Sun, Earth, and Moon are in the same line, is 10 times as bright as the Moon at First Quarter, the position of the satellite when it has completed one-quarter of its orbit around the Earth. When three-quarters of the orbit is completed, the Moon is said to be in the Last Quarter.

Variable brightness

(G.Fi.)

THE MOTION OF THE MOON

The apparent motion of the Moon. The lunar orbit is nearly circular, but it is more accurately described as an ellipse with the Earth at one focus, its eccentricity being about 0.0549. As seen from the Sun, however, the Moon sweeps out a sinuous concave path because its elliptic geocentric motion is superimposed on the elliptical orbit motion of the centre of gravity of the Earth–Moon system about the Sun.

The month. The Moon's motion around the Earth may be judged by reference to the Sun, the stars, the Earth's equator, or the Earth's orbit. To an Earth-bound observer, it is the fastest moving permanent natural object in the sky, making a complete circuit of the Earth with respect to the stellar background in only 27.322 days. This is the sidereal month.

By the time the Moon has completed one revolution of the Earth relative to the stars, the Earth–Moon system has itself moved through 7.5 percent of its orbit about the Sun. This causes the time between successive Full Moons or successive New Moons (called syzygies) to be longer than a sidereal month. This period of the lunar phases of 29.531 days is called the synodic, or lunar, month. The synodic month is determined by using the Earth–Sun direction as a reference.

If the orbital plane of the Moon were in the ecliptic (the plane of the Earth's orbit around the Sun), there would be a lunar eclipse at every Full Moon and a solar eclipse at every New Moon. Eclipses do not happen so frequently, however, because the lunar orbit is inclined to the ecliptic by 5°.15. The points at which the Moon crosses the ecliptic into the northern and southern celestial hemispheres are called the ascending and descending nodes, respectively. The interval between successive crossings of a given node, called the draconic, or nodical, month, is 27.212 days; the term draconic refers to the dragon that the ancients believed swallowed the Moon and Sun to cause eclipses. The draconic month differs from the sidereal month because the nodes regress (move opposite to the orbital motion), completing a full revolution every 18.6 years. This regression is caused principally by the gravitational pull of the Earth's equatorial bulge. The cycle of lunar and solar eclipses repeats every 18.6 years, a period that is sometimes called the saros.

Distance from the Earth

The mean distance between the Earth and the Moon is 384,400 kilometres, but, because of the ellipticity of the lunar orbit, the actual distance varies about 11 percent, between 363,000 and 406,000 kilometres. At its closest approach to the Earth, a satellite is said to be at perigee, and it is at apogee at its most distant separation; the time interval between successive perigee passages by the Moon of 27.555 days is called the anomalistic month.

Viewed from the celestial north pole, the Moon's geocentric orbital motion, the Earth–Moon's heliocentric motion, and the axial rotations of both Earth and Moon are all counterclockwise. Because the Earth spins west to east in a time that is short compared to the month, the Moon, like the Sun and the stars, appears to rise in the east and set in the west, even though its true motion relative to the celestial north pole is from west to east. Because of its orbital motion, the Moon rises, on the average, about 50 minutes later each day.

Optical librations. Galileo discovered that the Moon always turns the same face to the Earth. This means that the Moon rotates on its axis with exactly the same period as its orbital motion, a common feature among natural satellites. While its axial rotation is nearly uniform, the orbital speed of the Moon varies by 11 percent—the same amount of variation as in the Earth–Moon distance. The Moon moves faster near perigee and slower near apogee, with an average speed of 1.02 kilometres per second. Thus, from the Earth, the Moon seems to oscillate as the orbital speed loses or gains on the constant rotation. This causes features in the region of the limb, the visible edge of the Moon, to slip alternately into and out of view. A patient observer on Earth can eventually see about 57 percent of the total lunar surface. These apparent oscillations of the Moon are known as optical librations and can amount to Moon-centred angular displacements of 8° in longitude and 6°.8 in latitude. The laws governing such oscillations were discovered empirically by Gian Domenico Cassini in 1693.

The actual motion of the Moon. The mathematical treatment of the Moon's actual geocentric motion is called the lunar theory. Customary astronomical terminology distinguishes between an analytic procedure, called a theory of general perturbations, and a numerical method, called a theory of special perturbations. Because of the Moon's nearness to Earth and of its rapid angular motion, the lunar theory has long played a major role in formulating and testing dynamical theory.

Pre-Newtonian history. In ancient times, the Moon's phenomena were of great religious importance. Thus, considerable attention was paid to its motions; and the observations made were of an exactitude unsought in recording terrestrial phenomena. The Babylonians devised numerical methods for predicting eclipses and heliacal risings. The Greek astronomer Hipparchus in the 2nd century BC discovered the eccentricity of the lunar orbit, the motions of perigee and node, and the inclination of the orbital plane to the ecliptic. He explained the motion as being uniform in a circular orbit with the Earth placed eccentrically, and this was adequate to represent the observations then available. The numerical values of the parameters were derived from eclipse observations, some dating to Chaldean times. Although his results were superior to previous attempts, the restriction to eclipse data led Hipparchus to an incorrect value of eccentricity, through inability to distinguish the major inequality (departure from uniform motion, true value 6°.29), due to eccentricity, from the greatest solar perturbation. The latter inequality, called evection (1°.27), was discovered by the Greek astronomer Ptolemy in the 2nd century AD. Little significant improvement in the knowledge of the lunar motion was secured until the 16th-century Danish astronomer Tycho Brahe discovered the other great inequality caused by solar attraction, called the variation. This phenomenon, whose amplitude is 0°.66, vanishes when the Moon is new, full, or in quadrature. The physical explanation of these effects required the discovery of universal gravitation.

The discoveries of Hipparchus

Analytic theory. The analytic treatment of the lunar motion is carried out in two stages. In what is termed the "main problem," it is assumed that the Sun, Earth, and Moon behave as point masses and that the barycentre, or centre of gravity, of the Earth–Moon system describes a fixed ellipse around the Sun. After the solution of this problem, the remaining effects are introduced as small corrections to the main solution. These effects include the gravitational attractions of the planets acting on both the Earth and Moon, the forces caused by the irregular shapes of the Earth and Moon, corrections because of the use of a non-inertial coordinate system, and relativistic corrections.

The solution of the main problem gives the coordinates as harmonic series in which the arguments are linear combinations of the time-variant angles l, l', F, and D. The periods of revolution of l, F, and D are the anomalistic, draconic, and synodic months. That for l' is the anomalistic year in the Earth's heliocentric motion. The series coefficients are power series in these parameters: e ($= 0.055$) the eccentricity and γ ($= 0.045$) the sine of half the inclination of the lunar orbit; m ($= 0.075$) the ratio of average angular speeds of Earth and Moon; a ($= 0.0025$) the ratio of average geocentric distances of Moon and Sun; and e' ($= 0.017$) the eccentricity of Earth's orbit. The values of these parameters are determined from observation. The perturbations developed in the second stage of the analysis introduce the planetary orbit periods and the motion of the Moon's node into the arguments, as well as secular effects that grow without bound with time.

Isaac Newton based his theory of gravitation largely on the comparison of the motion of the Moon with that of bodies falling freely near Earth. He succeeded in proving that the principal periodic inequalities, as well as the motions of node and perigee, were caused by solar attraction. He deduced many new inequalities not previously known, but his motion of perigee was only about half the observed value; this disparity was later reduced to about 8 percent. Newton's theory was published in geometrical form in his *Principia,* but the method of fluxions (developed later into the integral calculus) was probably used in the actual derivation. No substantial advance was made beyond Newton's work until the mid-18th century, when the lunar motion was studied by many mathematicians.

Outstanding lunar theories were those obtained in the 19th century by Peter Andreas Hansen, Charles Eugène Delaunay, and George William Hill and Ernest William Brown. Hansen's theory was the most nearly numerical of the analytic theories, and it was of greater accuracy than any other before Brown; it was used worldwide from 1862 to 1922. Corrections by Simon Newcomb were applied after 1883. Hansen's work has served as the basis for orbit computations for artificial satellites.

The most complete algebraic development of the lunar theory until the present time was that of Delaunay. It was never adequate for comparison with observations, but it has been valuable in theoretical studies, the extension of other lunar theories, and the analysis of the orbits of natural satellites of other planets.

Hill developed methods that allowed the achievement of results of given accuracy with much reduced labour. These methods have been effective in studies of the general problem of three bodies (see MECHANICS) as well as in the practical solution of the lunar problem. The two principal features were (1) the introduction of the value of m from the outset to avoid convergence difficulties and (2) the use of a rectangular coordinate system rotating at the mean angular rate of the Moon. Hill's elegant method for determining the motion of perigee was later applied by John Couch Adams to the motion of the node. Based largely on Hill's work, Brown then developed the lunar theory that has been generally used to compute the coordinates of the Moon since 1923, although it was still necessary to replace the theoretical motions of node and perigee with the observed values, as the former were unsatisfactory.

Brown's series contain about 1,500 terms, five times as many as in Hansen's theory. Reduced to tables, it filled 650 quarto pages, and one man working alone with a desk calculator could extract the coordinates just fast enough to keep up with the real Moon. The electronic computer relieved this burden by permitting, since 1952, the direct calculation of the series. At the same time, corrections were introduced to compensate for the variations in the Earth's rotation rate and for the effect of tidal friction.

Brown's theory remained adequate for even the most precise astronomical uses for nearly half a century, and was still used for national almanacs in 1981. During the 1960s, the technologies of space exploration and radio astronomy, however, introduced the need for an enormous improvement in both the accuracy and internal coherence of the theory. At the same time, the rapid development of electronic computer systems for algebraic manipulation brought a resurgence of interest in the construction of analytic theories. The powerful new mathematical technique of Lie series, applied by computer, has permitted a return to the purely algebraic approach abandoned after Delaunay. Of the efforts to produce radically new theories, none had been sufficiently completed for general use by 1981. Along with the lunar coordinates, some of these efforts provide the necessary functions for correcting the lunar orbit parameters, an important innovation.

Numerical lunar theory. Although strictly numerical orbit theories have long been constructed for some solar system objects, this was impractical for the lunar motion before the development of the electronic computer. The nearest approach to a special theory was the attempt by Sir George Biddell Airy in 1883. He was unsuccessful and the idea was abandoned.

Space exploration required more accurate lunar positions than could be obtained from extant analytic theories; the solution was to compute the ephemerides by numerical integration of the equations of lunar orbital motion. This technique had been used for several decades for planetary orbits, but its application to the lunar problem was more difficult because of the Moon's high angular speed and the accuracy with which small perturbations can be observed. Numerical ephemerides first came into general use in 1969 and are now widely used for the most exacting applications, such as spacecraft navigation, lunar laser ranging, and radio interferometry. The method of special perturbations has the advantages of arbitrarily high internal precision and simultaneous inclusion of all desired perturbing forces in one basic calculation. Since both laser and radio observations can detect motions of a few centimetres, the calculations must account for such effects as planetary attractions, the Earth's oblateness, lunar libration, tidal friction in the Earth, and very complicated relativistic effects. The coupling between orbital and lunar rotation is now observable, so simultaneous numerical integration of these two motions is more commonplace.

Use of numerical ephemerides

The major disadvantage of the numerical approach is the uncorrectibility of a special theory; improvements to the parameter values require a completely new integration.

Cassini's laws for lunar rotation

The rotation of the Moon. Cassini's laws for the rotation of the Moon go beyond the fact of synchronous spin. They state, in addition, that the equator of the Moon is inclined by about 1°.5 to the ecliptic (6°.7 to the lunar orbit), and that the rotation axis always lies in a plane perpendicular to the line of nodes of the lunar orbit on the ecliptic. Thus, the axis precesses in space with the same period of 18.6 years as the nodes. The synchronism and the axial precession are evidence of a gravitational resonance in which the Earth controls the Moon's spin. The apparent monthly oscillations that result are due primarily to the parallax of an Earth-bound observer. The Moon does oscillate in its rotation, but these wobbles, called librations, are so small as to be undetectable except by precise observation.

Physical librations. The Moon is not spherically symmetric (see below, *The mass and gravity of the Moon*). The gravitational attraction of the Earth on the excess mass in the irregularities of the Moon's shape produces torques that influence rotation. Even the Cassini motion is caused by a bulge on the Moon along the principal axis pointed in the mean direction to Earth. The Earth's pull tends to bring its satellite back into line following any movement (caused by variations in the Moon's orbital speed) away from the true Earth direction. Consequently, the actual rotation forced by gravitational attractions includes a complex pattern of small motions that can exceed 100 arc seconds in selenocentric (Moon-centred) rotation. These motions are called physical librations. Since they arise entirely from gravitational action, they can be calculated from mathematical theory, following principles laid down by Sir Isaac Newton and Leonhard Euler. Such theories have only arisen in the past century, owing to the difficulties of observation; they are now constructed both by analytic methods and by numerical integration, similarly as (and sometimes simultaneously with) the lunar orbit theories.

Free librations. The lunar surface shows evidence of many impacts by external bodies, both large and small. In any such impact, much of the energy of the impacting object is transmitted to the Moon. Some of that energy produces yet another rotational motion, a type of Eulerian oscillation that, in specific reference to the Moon, is called free libration. The mathematical theory of rotating bodies specifies that these free librations will occur at only three frequencies, the periods of which are predictable from the physical properties of the Moon. Since these librations depend on the size and direction of the impacting object, however, their amplitudes cannot be predicted and can be determined only from observation. Once stimulated by an impact, the free librations gradually disappear by internal friction, and so are expected to be very small. The first claim of detection of a free libration was made in 1967, but this now seems spurious; modern data appear to show a free libration in longitude, but it cannot be larger than a selenocentric rotation of 2 arc seconds, 0.01 arc second as seen from the Earth.

Interpretation of small irregularities in the motion. The lunar motion provided Newton with the key to universal gravitation, and it has remained for three centuries one of the most difficult and fruitful practical problems in celestial mechanics. In conjunction with new methods of observation of unprecedented accuracy, recent work has confirmed the classical value of tidal friction, given evidence of a dense lunar core, determined the location and motion of the equinox, invalidated the Brans-Dicke scalar-tensor relativity theory, suggested a weakening of gravity, and improved knowledge of the shape, mass, and gravity field of the Moon. In the 1970s, the accuracy of the theories of both the orbit and rotation of the Moon was improved by a factor of nearly 100.

Such discoveries were made possible by the analysis of small irregularities in the observed motion of the Moon. For each observation, the mathematical theory is used to compute the place at which the Moon (or, more correctly, the observed point on the Moon) should have been if all aspects of the theory were correct. When this computation is compared with the observation, there remains a small difference, called a residual. When residuals from many observations are compared, systematic patterns appear that are clues to errors or omissions in the mathematical theory. Each contributing phenomenon (*e.g.,* mass, gravity, or observer's location) has a characteristic "signature" in the residuals. Improvement of the mathematical theory, or the discovery of new phenomena, is then derived from an analysis of the residuals to recognize and explain various signatures, many of which are present in the same data. It is essentially a problem in mathematical decoding. For example, in the 19th century, gravitational theory could not explain all of the motion of the Moon then observed. At first, this failure was thought to be due to a deficiency in the orbit theory. It was later shown to be caused, however, by previously unknown fluctuations in the Earth's rate of rotation. These fluctuations affect the apparent motion of all celestial objects, but the magnitude of the effect is proportional to the object's orbital speed, and so it was first discovered in the lunar motion. From 1954 to 1976, observations of the Moon were used to refine the determination of time from the rotation of Earth; more precise refinements are now achieved by the use of atomic clocks.

Using new techniques of occultations, radio interferometry and laser ranging, much finer details of the lunar orbit and rotation are now observable than was formerly possible. Such data are, in fact, so precise that it is no longer feasible to consider the orbital and rotational motions of the Moon as unrelated subjects.

An occultation is the disappearance or reappearance of a star from the view of an Earth-based observer as the Moon passes in front of it. In a normal occultation, there is one disappearance and one reappearance in a given event. In a grazing occultation, however, the edge of the Moon's disk passes in front of the star so that the star disappears behind each of a series of lunar mountains and reappears in each of the valleys between them. In both cases, measurements of the lunar surface can be made; these measurements can be accurate within a few metres at the Moon, but their further refinement is limited by the Moon's topography. Occultations have been used primarily to study the shape of the Moon and the long-term evolution of its orbit, including the effects of both tidal friction and the possible weakening of gravity.

Very long baseline (radio) interferometry (VLBI) is used to determine the direction of a celestial source of radio waves. It measures the parallax effect of two simultaneous observations of radio waves by Earth-based stations separated by large distances. The Moon is not a natural source of radio waves, and so radio transmitters were landed on the Moon by various spacecraft to send such radio signals back to Earth. Another technique of lunar observation is that of lunar laser ranging (LLR), which provides measurements of the Moon's distance from Earth. Like VLBI, LLR required the placement of man-made devices—in this case, passive reflectors, or mirrors—on the lunar surface. In LLR, a short light-pulse is emitted through an Earth-based telescope to the Moon, where it is reflected from one of the mirrors back to Earth, where the light travel-time is measured. Although LLR observations have been made at several places, more than 95 percent of them have been made at McDonald Observatory, where an LLR program has been in operation since the landing of Apollo 11 on the Moon in 1969.

Both VLBI and LLR provide data on the position of the Moon that is accurate within a few centimetres. Using these techniques, important discoveries have been made since 1969. In the question of gravitational theory, for example, the lunar motion continues in its important role; new data have shown the so-called Nordtvedt effect, a violation of the principle of equivalence, to be essentially zero, thereby eliminating the Brans-Dicke-Jordan theories of relativity as serious alternatives to Einstein's general relativity theory. No convincing evidence has yet been found for a variation of the gravitational constant, a variation required by some theories of relativity. In other results, terrestrial tidal friction has been found to affect the lunar

Nordtvedt effect

longitude by an acceleration of −24 arc seconds per century per century, a value consistent with that discerned by Sir Harold Spencer-Jones in 1938, but only half that believed to be correct in 1970. The principal moment of inertia of the Moon as now measured indicates a significant increase of density toward the Moon's centre, which suggests the existence of a small metallic core. Finally, measurements of free libration are consistent with, but do not prove, the hypothesis that the Giordano Bruno crater was formed on the lunar surface by a meteorite impact witnessed from Earth in AD 1178.

THE MOON'S MASS AND GRAVITY FIELD

Until the era of space exploration, the mass and the gravity field of the Moon were only determinable by their effects on telescopic observations of the apparent motions of the Moon and stars. As late as 1965, the ratio, μ, of the Moon's mass to that of the Earth was calculable only to three figures, and knowledge of the lunar gravity field was so scant that a hollow Moon was still a mathematical, if not a physical, possibility. It was evident, however, that the Moon was of significantly lower density and of less spherical shape than the Earth. Knowledge of both the mass and the gravity field of the Moon was markedly improved by the spacecraft missions of the 1960s and '70s and by the use of radio interferometry and laser ranging techniques.

The mass of the Moon. The mass of a celestial object can be determined only from the gravitational effects of that object on others and can be expressed only in a ratio with respect to a "standard" celestial mass. In studies of the solar system, as in stellar astronomy, the standard mass used for such calculations is that of the Sun; a secondary standard mass, that of the Earth, is also used for the mass of the Moon. The total mass of the Earth–Moon system, in solar units, is best determined by comparing the geocentric orbit of the Moon with the heliocentric orbit of the Earth–Moon system, using Kepler's third law relating the orbital periods, distances, and masses of bodies in the solar system. Recent determinations from very long baseline interferometry (VLBI) and lunar laser ranging (LLR) calculations give a value of 328,900.5 for the mass ratio Sun: (Earth + Moon).

Results of VLBI and LLR

Accurate separation of the masses of the Earth and Moon requires an analysis of their gravitational influence on other nearby bodies, such as spacecraft leaving the Earth–Moon system for other planets. Roughly speaking, the Earth and Moon both describe elliptical paths about their common centre of gravity, or barycentre, and the ratio of the sizes of these ellipses is the inverse of the ratio of the Moon's mass to that of the Earth, or μ. (Since μ is about 1:81, the Earth's centre is about 4,700 kilometres from the barycentre, which is itself, therefore, always within the body of the Earth.) For distant objects, the gravitational influence of the Earth–Moon system is essentially that of a single planet of their combined mass. For a spacecraft beyond the Moon's orbit but not yet in the vicinity of Venus or Mars, however, the Earth and Moon each produce distinct effects that appear as monthly perturbations in the spacecraft's motion. Data of such perturbations have been obtained from all interplanetary spacecraft since Mariner 2 was launched in 1962; the resulting calculation of the value of the Moon/Earth mass ratio is $\mu = 1/81.3007 = 0.012300$. The Moon is thus only 1/26,410,940 the mass of the Sun.

By far the least accurate step in determining the mass of the Moon is the next, which requires the value of the mass of the Earth as measured in grams and relies on measurements of the Earth's surface gravity and radius. With a value for the Earth's mass of 5.977×10^{27} grams, the mass of the Moon is 7.352×10^{25} grams.

The gravity field of the Moon. A celestial object that is spherically symmetrical exerts a gravitational influence on other bodies that is equivalent to that of an infinitesimal point of the same mass. Any irregularities of shape, however, affect the gravitational field of such a body. Thus, to speak of the gravity field of a body is to discuss its departures from sphericity. A discussion of a gravity field can be conveniently separated into three categories: triaxiality, gravity harmonics, and local gravity anomalies.

Tri-axiality. Any body that is not a perfect sphere has an axis of greatest moment of inertia, *C*. This axis is usually that about which the body rotates, perpendicular to the body's equator. The axis of least moment of inertia, *A*, is usually in the equator. Perpendicular to both axes is the intermediate axis of inertia, *B*. These three axes are the parameters of any tri-axial body. In the case of the Earth, the Equator is virtually circular, so *A* and *B* are essentially equal, and the Equator bulges like a belt around the planet. The Moon, however, apparently never rotated fast enough to distribute its mass evenly in circumference and, consequently, it bulges in only one direction—along the Earth–Moon line. The differences of the principal moments of inertia of the Moon are not enormous, but they are important: $(B-A)/C = 0.00023$, $(C-A)/B = 0.000631$, $(C-B)/A = 0.00040$. These figures represent the ellipticities of the lunar equator, the polar section in the Earth–Moon direction, and the polar section normal to that direction, respectively.

The lunar tri-axiality was known and taken into account in the lunar orbit theories of the 19th century and the lunar rotation theories of the early 20th century. More detailed knowledge of the lunar gravity field was not to come until a spacecraft, Luna 10, was first sent into orbit around the Moon in 1966. The Moon is exceedingly lumpy, and the three-dimensional average given by the principal moments of inertia does not describe the gravity field accurately enough for present-day observations.

Gravity harmonics. In mathematical theories, it is the preferred practice to describe the gravity field of a body in spherical harmonic functions. In such a scheme, the non-spherical perturbing potential acting at an external point of distance, *r*, longitude, *l*, and latitude, *b* (all three being referred to the perturbing body's equator), is written as:

$$\Sigma\Sigma\ (1/r)^{\mathrm{n}}\, P_{\mathrm{nm}}(\sin b)\ (C_{\mathrm{nm}}\cos \mathrm{m}l + S_{\mathrm{nm}}\sin \mathrm{m}l)$$

where, in principle, the summations cover all values of $\mathrm{n} > 1$ and all non-negative values of $\mathrm{m} \leq \mathrm{n}$. In practice, the term $(1/r)^{\mathrm{n}}$ diminishes rapidly as n increases, so only a limited number of terms is required. The coefficients *C* and *S* are numerical quantities, called the gravity harmonic coefficients, that describe the shape of the gravity field; they can be determined only from observation. The principal moments of inertia are equivalent to the parameters C_{20}, C_{22}, and S_{22}. A description more detailed than tri-axiality requires a determination of the harmonic coefficients for $\mathrm{n} > 2$. Most of the values for $C_{3\mathrm{m}}$ and $S_{3\mathrm{m}}$ are obtained from LLR and VLBI observations by analysis of the residuals. More detailed structure of the harmonic field is only possible from analysis of the motion of artificial lunar satellites. A combination of LLR and lunar orbiter data has been used to determine the values of the gravity coefficients up to $\mathrm{n} = 5$.

Gravity harmonic coefficients

For most bodies, the values of *C* and *S* are expected to diminish significantly as n increases. For example, for the Earth (using units of 10^{-6}), $C_{20} = -1{,}083$, $C_{30} = -2.4$, and most of the remaining coefficients are negligible except for close artificial satellites. The irregularity of the Moon is illustrated by the corresponding values of its harmonic coefficients: $C_{20} = -202$, $C_{22} = +22$, $C_{30} = -12$, $C_{31} = +31$, $C_{32} = +5$, $S_{31} = +6$. Even some of the coefficients with $\mathrm{n} = 4$ and $\mathrm{n} = 5$ are significant at this level.

Local gravity anomalies. The Moon is the only body known for which the usual harmonic representation of the gravity field does not suffice for all purposes. Early analyses of tracking data from the first U.S. lunar satellite launched in 1966 produced unexplainable irregularities. The residuals showed curious oscillations that were reported by subsequent spacecraft, but not in any decipherable pattern, constituting a serious operational problem for two years. Attempts to solve the problem at first concentrated on improving and extending the set of harmonic coefficients. In 1968, it was shown that these effects were due to anomalous accelerations as the spacecraft passed over certain of the lunar maria, indicating unexpectedly large masses in those regions. It has since been shown that such concentrations of mass—popularly known as mascons—exist in all of the ringed maria. Though large enough to produce observable effects on the spacecraft, the mascons are too small to be represented by the harmonic coefficients.

In terms of the Moon's gravity field, Mare Orientale is unique. It looks rather like a giant target, complete with rings, as though a meteorite had splashed into a partly molten surface that solidified before the waves subsided. In fact, the local gravity field shows the same features, with alternating rings of excess gravity and gravity defect. This is one of the few cases on the Moon in which there is a clear-cut correspondence between the shape of the gravity field and the shape of the lunar surface.

(J.D.Mu.)

THE PHYSICAL NATURE OF THE MOON: SELENOLOGY

Observations from the Earth and from space vehicles. *Limits of resolution.* Lunar craters of the order of 100 metres in diameter may be seen from the Earth under the best conditions using moderately sized telescopes, and, because, in practice, resolution depends on the shape of an object and on its contrast with the surroundings, the same instruments can be used to view lunar lineaments (linear features), such as rilles, considerably smaller than 100 metres in width, provided they are many times longer than 100 metres. The best Earth-based photographs of the Moon show some (the sharpest) craters as small as 500 metres in diameter. Photographs from space vehicles have resolved details in some cases as small as a metre.

The first far-side pictures

The first camera was put into an orbit passing behind the Moon (October 1959) by the Soviet Union. Its useful photography was confined to some 70 percent of the Moon's averted hemisphere lit by a high Sun, and both the picture quality and the resolution were poor by the standards of Earth-based telescopic photography. Useful far-side photographs were obtained from another Soviet space vehicle, which completed the lunar far-side coverage in July 1965, and from a third, which provided higher resolution coverage of a limited portion of the far side. A number of ray craters (see below) were readily identifiable, but the most surprising observation was that most of the far side of the Moon contained no maria. Major crater chains, up to 1,000 kilometres long, were found.

Small-scale details. The photography from a series of vehicles designed to crash-land on the Moon and relay television pictures to Earth prior to crashing was not successful until July 1964, when a U.S. probe approached and crashed in a part of Mare Nubium; the area was renamed Mare Cognitum by the International Astronomical Union. For the first time, more and more small craters were seen as the vehicle effectively decreased the lunar-surface resolution to a limit of about 50 centimetres. Craters of various shapes, rock blocks, and ridges 100 to 1,000 times smaller than any that had been seen before were recorded during the last minutes of descent.

The Moon was found to be cratered more heavily than had been predicted, and the substantial excess of craters was attributed both to the mechanism of secondary cratering—that is, crater formation by objects ejected from a primary meteoric (or volcanic) cratering event—and to endogenic cratering—that is, crater formation by internal processes such as outgassing or collapse. A few dimple-shaped craters, probably the result of drainage of the Moon's veneer of fragmental rock into subsurface fractures, were discovered. Many more were later found in Mare Tranquillitatis and in the structure of Alphonsus, a crater 110 kilometres in diameter. None of these craters has the bowl- or saucer-shaped profile characteristic of a fresh, sharply sculptured (eumorphic) explosion crater or of an older, denuded (submorphic) explosion crater. On the other hand, the majority of the craters less than a kilometre in diameter could well be explosion craters produced either by high-speed impact or by violent outgassing (explosive volcanism).

Mare Cognitum is traversed by faint rays of lighter hued material. The lunar rays that emanate from many craters are generally thought to be composed of small craters and blocks of rock. A certain proportion of the craters set in rays may be secondary-impact craters. In one cluster of craters, chains (linear arrays) of both sharply sculptured and older craters trend in a given, narrow range of azimuth. It is not known whether the two types of crater were formed as respective secondaries when hard and soft material landed to produce craters at the same time and at the same distance from some primary crater; or if the eumorphic and submorphic craters were formed internally at widely different times, the latter being formed first and then undergoing denudation.

The lunar regolith

The principal erosive force on the Moon at the present time is probably that of meteoric impact. Primary meteoroids are increasingly the more numerous with diminution in size, so that micrometeoroids are probably the principal factor involved in the attrition of freshly formed surfaces on the Moon. The net direction of movement of material sprayed from a meteoric cratering event is downslope; and this downward migration of rock fragments tends to level the lunar terrain and to generate a ubiquitous veneer of particles that is variable in thickness from one locality to the next. This layer, known as the lunar regolith, is composed of fines, particles less than one centimetre and mostly much less than one millimetre in size, and of larger fragments of rock. The process of impact mixing of the lunar soil and the particles of the regolith is sometimes referred to as "impact gardening."

Some of the blocks of rock photographed on the Moon appear to be partially buried in the regolith, and a first estimate of the strength of the lunar-surface materials was made from considerations of the depth of penetration of these blocks and the distances they had been thrown. A bearing strength of about one kilogram per square centimetre found for this part of the Moon showed that it was strong enough to support a man clad in a spacesuit or any of the planned soft-landing craft. Following the first soft landing of a spacecraft, Luna 9, on the Moon, in February 1966, the Soviets determined the soil in Oceanus Procellarum to have a bearing strength of at least several kilograms per square centimetre.

It appears to have been confirmed that some of the small craters of the Moon are of internal origin, though it is still believed that the vast majority of these craters are of impact origin. Several chains of contiguous craters are generally thought to be secondary craters from Theophilus, a crater 85 kilometres in diameter, even though many of them run nearly parallel to tectonic lineaments such as linear ridges and troughs—a situation indicative, instead, of an endogenic origin of these craters. The blocks of rock, up to 100 metres in diameter, are commonly scattered in and around craters that were probably deep enough to cut through the friable regolith into cohesive bedrock. It has been argued that this bedrock would have been excavated and tossed, as blocks, from a crater during its formation.

The rilles

Many rilles (lunar trenches of lengths up to a few hundred kilometres and widths up to a few kilometres) have been photographed by space vehicles. Those recorded in Alphonsus are particularly interesting because of the intimate connection between the rilles in the floor of the crater and dark halo craters that lie in the rilles. Exceptionally dark deposits encircling these craters have tended to obliterate underlying relief and, locally, fill the rilles. The craters themselves are elongated along the axis of a rille and are without raised rims—characteristics that point to internal origins. The associated dark deposits were probably erupted as sprays of pyroclasts.

In the next phase of lunar exploration, closeup television photographs of the landing feet of a soft-landing vehicle in June 1966 (within a large ring structure, known as Flamsteed P, in Oceanus Procellarum) showed that the feet had penetrated a few centimetres into the ground and had disturbed the soil around them, giving it a darker appearance in the process. Stones and blocks of rock were found to vary considerably in appearance. Some appeared to be partly buried, their surfaces smoothed by erosion. Some appeared to have been put on the surface and to have remained in their original, angular configurations. Some had different shapes and numbers of pores (vesicularity) and different structure (layering and differential denudation) or, possibly, composition (tonal differences and the presence of inclusions).

Similar results were found from other probes soft-landed in Mare Tranquillitatis and in Sinus Medii in missions

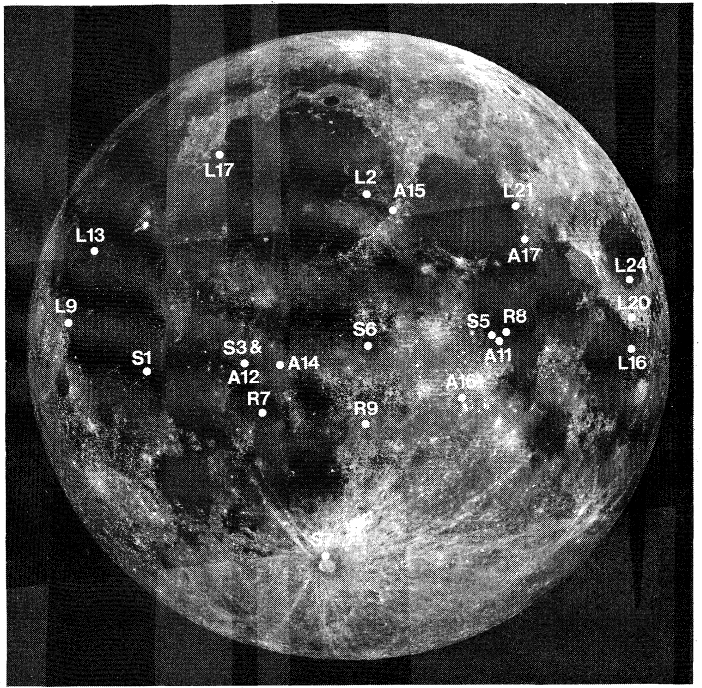


Figure 21: The side of the Moon visible from the Earth, photographed at Full Moon. The approximate points of landing are shown for selected U.S. and Soviet spacecraft on the Apollo (A), Ranger (R), Surveyor (S), and Luna (L) series. Three dark "seas," left to right, just above centre, are Mare Imbrium, Mare Serenitatis, and Mare Tranquillitatis. Tycho crater lies at bottom centre, just below landing site S7.

By courtesy of the Lunar and Planetary Laboratory, University of Arizona

meant to examine possible landing areas and analyze the soil. One mission was sent 42° south of the lunar equator to examine, for the first time, part of the Moon's highland domain. It landed (January 1968) successfully within what was, by comparison with the area of the highlands around Tycho, a fairly smooth area. The television cameras, however, showed the terrain to be strewn with many more blocks of rock per unit area than had been found in the maria. The most revealing discovery concerned the composition of the rocks; the percent weights of elements in rocks were estimated, and the landing area was found to be similar in composition to the mare samples, with the exception that the iron-group elements (mass numbers 47 to 65) were less abundant in the highlands.

Data on rock composition

The elemental composition of the rocks at three sites approached that of some terrestrial basalts (types of dark igneous rock). The tested areas of the Moon were compositionally unlike chondrites—the most common kind of meteorite—and analyses indicated that the Moon had been differentiated and heated. The theory that the Moon originated as a cold body and has never been heated suffered a serious blow, although the heating might have been the local result of major impacts. Later analyses of rocks returned to Earth provided further information on heating (see below).

Results from manned lunar landings. The first manned lunar landing was made in July 1969 in Mare Tranquillitatis. Some 22 kilograms of lunar rocks were collected and returned to Earth for study.

Later manned landings were in the Fra Mauro highlands of Oceanus Procellarum, between the Apennine Mountains and the Hadley Rille in Mare Imbrium, in the highlands near Descartes, and in the Taurus Mountains. Each landing added to the store of lunar rocks returned to Earth, and each crew left behind functioning observing instruments. A seismometer, a device commonly used to measure earthquakes, was included in each of the first five scientific packages of the manned landing missions.

Though the first seismometer failed after only a short operating time, the others worked simultaneously. In registering the vibrations generated by impacts and moonquakes, they have provided valuable information about the frequency with which meteoroids strike the Moon and also about the slight, relative movements of rocks along lunar faults or other slip surfaces. Other data revealed by instruments included the behaviour of lunar material at depth and the probable nature of the rocks at various depths and localities.

In addition to impacts by meteoroids, the Moon is subject to thousands of minor quakes each year. Most of them are triggered by the Earth's changing tidal pull, which, during each lunation, is greatest when the Moon is at its closest point to the Earth. Lunar seismic signals are characteristically different from any received at terrestrial seismic stations, the difference indicating that the Moon's rocks scatter and transfer energy much more efficiently than do Earth rocks. The reason is probably to be found in the dryness and the fractured or fragmental nature of the outer parts of the Moon. Lunar quakes originate from depths about halfway from the surface to the centre of the Moon; observations are consistent with the theory that there is some plasticity or melting of some of the rocks at these depths. Indeed, heat-flow measurements at two localities showed a thermal output of 0.03 watt per square metre—a value as high as one-half of the mean output of heat from the Earth.

Origin of lunar quakes

Missions have also landed a type of optical reflector on the Moon, capable of returning to Earth extremely accurately timed laser pulses for reception by large telescopes. (See above, *Interpretation of small irregularities in the motion.*) Other instruments left behind on the various manned landings provide continuing physical and geophysical data. Earlier measurements of the physical properties of the lunar environment using crash-landing and orbiting vehicles were supplemented or superseded by the results drawn from the later missions. (For sites of the impact or landing of significant lunar missions, see Figure 21 and Table 8.)

Table 8: Impact or Landing Sites of Successful Lunar Spacecraft

name and launch date	remarks
Luna 2 Sept. 12, 1959	first man-made craft to impact the Moon
Ranger 7 July 28, 1964	first craft to obtain high-resolution (one metre) television pictures
Ranger 8 Feb. 17, 1965	similar mission to Ranger 7
Ranger 9 March 21, 1965	similar mission to Ranger 7, but in highland region
Luna 9 Jan. 31, 1966	first craft to soft-land and relay panoramic and close-up pictures from lunar surface
Surveyor 1 May 30, 1966	first U.S. craft to soft-land and relay panoramic and close-up pictures from lunar surface
Luna 13 Dec. 21, 1966	similar to Luna 9, but also tested lunar soil hardness
Surveyor 3 April 17, 1967	similar to Surveyor 1, also first craft to dig trenches
Surveyor 5 Sept. 8, 1967	similar to Surveyor 1, also first craft to analyze lunar mare soil
Surveyor 6 Nov. 7, 1967	similar to Surveyor 5
Surveyor 7 Jan. 7, 1968	similar to Surveyor 5 and 3, first soft landing in highlands
Apollo 11 July 16, 1969	first manned landing (Mare Tranquillitatis)
Apollo 12 Nov. 14, 1969	first manned landing near an unmanned craft (Surveyor 3)
Luna 16 Sept. 12, 1970	first unmanned craft to return lunar soil to Earth
Luna 17 Nov. 10, 1970	first unmanned craft to deploy a roving vehicle (Lunakhod 1)
Apollo 14 Jan. 31, 1971	first manned landing in highland region (Fra Mauro)
Apollo 15 July 26, 1971	first manned landing to use roving vehicle (Hadley Rille)
Luna 20 Feb. 14, 1972	similar to Luna 16, but highlands sample instead of mare
Apollo 16 April 16, 1972	similar to Apollo 15, but in highland region (Descartes)
Apollo 17 Dec. 7, 1972	similar to Apollo 16, but sampled highland and adjacent valley
Luna 21 Jan. 8, 1973	similar to Luna 20, but landed in Le Monnier
Luna 24 Aug. 9, 1976	similar to Luna 20, but landed in Mare Crisium

Data from closeup photographs. Increasingly valuable photographs of the Moon were taken by the astronauts on succeeding missions, both from orbit and from the surface. Many of the rock samples were photographed in relation to their surroundings prior to removal from the Moon's surface, to assist in later interpretations (the broad setting from which a rock is drawn is important to an investigator in any comprehensive geological study).

The lunar photographs by no means record all possible bits of lunar surface, but vast amounts of unmeasured, unprocessed data are stored in existing photographs. From them, lunar craters have been measured for size, shape, and depth (from shadow measurements): crater counts have been used in estimates of the terrain roughness and in statistical studies bearing on the origin of the craters. The roughness of lunar rock blocks, estimated from photographs and returned samples, has been related to theories of denudation (wearing down) and aging. Lunar-surface processes, such as transport of the material and impact mixing, have been studied using photographs alone. Lava flows covering large areas of the maria have been detected and mapped using Earth-based infrared and ultraviolet photographs in combination with photographs taken from lunar orbit. The brightness contrasts displayed by different rock units have been measured from Full Moon photographs. Slopes have been assessed from shadows under different Sun angles. Topographic maps have been prepared. Texturally distinct geological units have been mapped and used to produce a lunar stratigraphic column of relative ages. Faults and other tectonic lineaments have been recognized on the photographs.

Porosity of topsoils

The photometric function. Scientists correctly reasoned, in the early 20th century, that the Moon's photometric function (changes in the intensity of scattered sunlight with different angles of illumination and reflection) proved that the lunar surface was far from smooth on a centimetric scale. The respective scattering properties of highland and mare material were deduced in 1956, when it was shown that the lunar surface in both the lunarite and the lunabase was extremely irregular and porous and that both types of terrain displayed essentially similar surface structure. Slight differences were observed in the way light was scattered by lunarite (highland) and lunabase (maria) material, and these differences indicated that, in general, there was a slightly greater roughness (more irregularities, or pores, per unit of surface) in the lunabase. In a more sophisticated analysis made in 1971, it was estimated that the porosity (fractional pore space) of the highland topsoil had an average value of 60 percent and that the mare topsoil averaged 80 percent. The measured values of the Moon's surface porosity at specific sites are not generally in conflict with these overall results.

The reason for the different structure in the uppermost centimetre or so of the lunarite and lunabase may be found in a study of the differential denudation or of the different histories of outgassing of the two types of terrain. Processes such as mass wastage—that is, downslope migration of material under gravity—may be expected to play a relatively important role in the highlands. Therefore, there may well be a tendency for craters and pores in the highlands to be smoothed over more quickly than corresponding features in the maria. On the other hand, the outgassing of the mare materials is believed to have occurred later than much of that in the highlands; and the initial porosity of the two types of material probably differed as well.

The packing of lunar fines

A special use of photometry is in the determination of the packing of lunar fines (small particles and dust). The acceleration caused by the Earth's gravity is six times that of the Moon at their respective surfaces, so that fines scooped from the Moon and transported to Earth will undergo an increase in their state of compaction. Agitation, severe during parts of the journey back to Earth, assists the process of compaction. The undisturbed fines and samples returned to Earth will scatter light in different ways, and it is possible, by rearranging the compacted fines into various states on Earth and comparing them with remote lunar observations, to deduce their approximate surface state before they left the Moon.

It is difficult to visually assess large-scale colour contrasts on the Moon from the Earth. A profitable investigation of lunar colour contrasts may be made from ultraviolet and infrared photographs. The technique is useful for surveying extensive areas of the Moon for differing rock units.

Certain parts of the Moon strongly absorb ultraviolet light. Such an absorption can be produced by traces of the element sulfur, for example. At the other end of the spectrum, lunar rocks display characteristic absorption bands in their infrared reflection spectra. One such prominent band, at a wavelength of about one micron, is due to the presence of iron in certain lunar materials.

Another instructive technique is that of plotting the photometrically determined colours of discrete patches of lunar terrain against their respective albedos and comparing the resulting colour-brightness diagram with others constructed for a wide range of terrestrial rocks prepared in widely different forms. It is found that blocks of terrestrial-type rock, or meteorites, cannot cover more than a small fraction of the lunar surface, although many rocks ground to powders and then irradiated with protons (to simulate the effect of particles from the Sun striking the Moon) have been found to approach the lunar domain of the colour-brightness diagram.

The lunar soil

Lunar polarization studies. A much more powerful optical method of studying the nature of the particles of lunar soil is that in which polarization curves are utilized. Polarization is a wave property of light, and different substances have different characteristic polarization curves. At a given lunar phase, the amount by which light from different parts of the Moon is polarized depends on the brightness of those parts: the brightest spots polarize no more than about 5 percent of the incident light, whereas the degree of polarization attains 20 percent or so for the darkest spots. These maxima were recorded at times close to those of the First and Last Quarters of the Moon. Most terrestrial rocks, whether solid or pulverized, polarize much more light than the Moon. The Full Moon polarizes no light, and practically no light is polarized when the Moon offers a very narrow, sunlit crescent. When the light intensity component in the Sun–Moon–Earth plane exceeds the corresponding component at right angles to this plane of diffusion, the maximum polarization in the visible (yellow light) is found to be − 1.2 percent for all parts of the Moon. The independence of the negative part of the polarization curve on the location on the Moon's near side indicates a fairly high degree of textural uniformity of this part of the lunar surface. Extensive laboratory experiments made with visual and photoelectric polarimeters, in attempts to narrow the choice of lunar surface models by careful comparison of laboratory and lunar polarimetric functions, have shown that the best match to the lunar soil at large is in dark, fine-grained basic to intermediate rocks.

Angle of inversion

In changing from positive to negative, the Moon's polarization passes through zero at a phase angle of about 23°, called the angle of inversion. The angle of inversion remains constant only when the behaviour of light of a particular wavelength is considered: the angle increases with wavelength (whereas the degree of polarization generally decreases as the light becomes redder). The maximum polarization, at a given wavelength, generally decreases as the particle size decreases. Reducing the grain size of a rock usually has the effect of increasing its albedo, and this, in turn, generally decreases its maximum degree of polarization.

It has been suggested that solar protons not only damage but also darken particles on the lunar surface. It is difficult to simulate all the appropriate aspects of the lunar environment in the laboratory, but it is clear that, if the solar wind were to darken fine particles derived from certain terrestrial rocks, their polarization functions could be made to approach or be rendered identical with those pertaining to the lunar fines.

Polarization studies of light-toned rocks brought from the Moon have shown that they either do not occur in large concentrations on the Moon's surface or they are generally covered with the fines responsible for the uniformity of the Moon's near-side polarization.

Infrared (heat) studies. Evidence of a widely dispersed layer of dustlike material on the Moon is given by measurements of infrared radiation from the Moon. The lag shown by temperature changes of the subsolar point during eclipse shows the very poor conductivity of the lunar surface in the eight- to 14-micrometre (one micrometre = 10^{-6} metre) waveband. Highlands and maria behave similarly, so the poorly conducting layer must cover most of the surface.

The poor surface conduction also produces differences in heat curves at different wavelengths. Measured at a microwave wavelength of 1.25 centimetres, the Moon's temperature attains a lower maximum than that determined by the eight- to 14-micrometre infrared measurements, and this lower maximum temperature occurs about 3.5 days later than the infrared maximum. The eight- to 14-micrometre radiation is from a surface layer only a millimetre or so thick, whereas the centimetre radiation comes from layers a few tens of centimetres beneath the surface. The Sun's heat generates the highest temperatures on the surface of the Moon, but deeper rocks are also warmed as the thermal wave penetrates and is attenuated in the poorly conducting regolith. Lunar radio emission derives from thermal energy finding its way to different levels in the Moon. Increasingly longer radio wavelengths have been used to observe the Moon at increasingly greater depths; but, although the average depth of the emitted waves increases with wavelength, the radiation at radio frequencies originates over a range of depths. In addition, the limit of angular resolution of a telescope increases with wavelength, so that, whereas the optical heat measurements may refer to a small patch of the Moon, the specific origin of the longer radio waves may be impossible to locate on the disk. At wavelengths above about 10 centimetres, no diurnal change in the intensity of received radiation has been observed. In these cases the upper layers of the Moon, which show strong diurnal temperature changes, are contributing a negligible fraction of the total radiation received.

Changes in temperature with depth

In principle, radio observations provide a means of estimating the variation of temperature with depth in the regolith and, thus, of investigating any internal heat source such as one that could be produced by widespread radioactive elements. The thermal conductivity at depth in the Moon, however, is not yet well-known, nor is it in the troublesome regolith, in particular, where the dependence of conductivity on temperature and depth is marked and thermal transport by radiation through incompletely specified vacuum pores contributes to the difficulties.

Theoretical models of the Moon's upper layers have been attempted, but it is not possible to derive the conductivity directly: density and specific heat must be determined first. Furthermore, the specific heat of lunar rock depends on its temperature: the specific heats of both a vesicular basalt sample and a sample of regolith from the Moon have been found to change smoothly from 0.06 calorie per gram per K at 90 K to 0.2 calorie per gram per K at 350 K. Such dependence on the prevailing temperature can change the thermal inertia of lunar surface materials by a factor of up to 2. A representative value of the thermal conductivity of the uppermost part of the regolith is about 2×10^{-5} watt per centimetre per K. Apollo measurements showed that, at depths of the order of one metre, this value was exceeded by a factor of 10.

The so-called hot spots

During optical infrared scans across the eclipsed Moon with a resolution of details of about 20 kilometres, numerous patches of terrain—mostly eumorphic craters—have been found to be significantly warmer than their surroundings. The most easily recognizable thermal anomalies (hot spots) were associated with bright-ray centres such as Tycho, Copernicus, and Aristarchus. These craters remained cooler than their surroundings during the lunar daytime, and the hot spots could be areas of higher than average thermal conductivity, permitting heat at depth to flow out during eclipse and be radiated to space more readily than from adjacent areas.

The night spectrum of Tycho, Copernicus, and Aristarchus, observed over the three- to 15-micrometre waveband, proved that the surface rocks of these craters emitted radiation from areas at widely different temperatures. Parts of the floors of two, Tycho (Figure 22) and Aristarchus, appear to be of ashy or regolith-type material, but other parts are probably volcanic domes apparently not covered by detritus or pyroclasts (substances fragmented by volcanic action). With geologically dissimilar rock units found in and around the craters, including lava flows with contorted surface ridging, such craters must be expected to generate cooling curves that are made up from several different cooling functions.

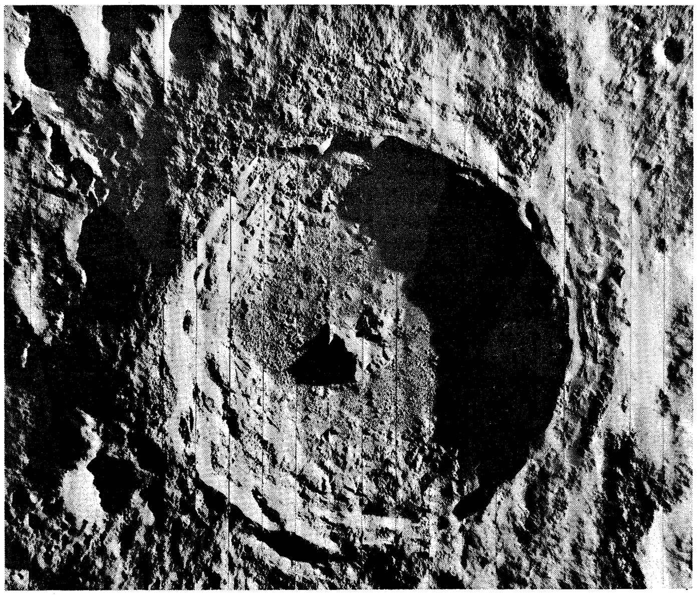


Figure 22: The lunar crater Tycho, in the lunar highlands. Picture obtained by the U.S. Lunar Orbiter 5.

By courtesy of National Aeronautics and Space Administration

The hottest craters are invariably the most sharply sculptured, freshest looking features. Large blocks of rock and characteristically rough terrain may be resolved around several of these craters, the hottest 20 of which have diameters from four to 16 kilometres in length.

There are relatively fewer hot spots per unit surface in the highlands than in the maria, and the number density of hot spots changes significantly from one mare to the next. The reasons are not well understood, though a tentative explanation may be offered in the case of Mare Tranquillitatis, which has an anomalously high number of hot spots per unit of its surface. First, there is a positive correlation between the number of eumorphic craters in the lunabase and the number of hot spots in the same area. Second, a study of the shape parameters of small, concentric craters suggests that at least the southern part of Mare Tranquillitatis has a thinner cover of regolith than most of the maria. It may be imagined that the sharpest eumorphic craters—which are almost certainly explosion craters—would be formed preferentially in areas of thin regolith, where the excavation of coherent bedrock could provide an abundance of ejected rock blocks.

Cold spots

A few cold spots have also been observed. Of the nine cold spots with the highest thermal gradients (those in excess of 5/7 K per second of arc), none is associated with a single, eumorphic crater.

Finally, there is the type of thermal anomaly in which a spot is observed to be hot during both eclipse and daytime conditions. Such a spot could be indicative of a source of internal heat.

Analysis of returned radar signals. For better understanding of the nature and origin of lunar thermal anomalies, more widely dispersed and improved estimates of the regolith thickness, layering, and conductivity are required. Analyses of radar signals returned from the Moon have proved of interest in this connection. The first part of a radar echo from the Moon returned to Earth derives from the central region of the Earth-pointing face. Later parts of the return signal come from rings of progressively increasing radii. Also, the lunar librational motions introduce frequency shifts in the return signals. As these effects are superimposed, it is possible to isolate the radar-reflection characteristics of two small patches on the Moon. Anomalously strong radar reflections were obtained in 1962 from Tycho and the corresponding area in the Moon's northern hemisphere. The latter had no topographic peculiarities, and it was concluded that Tycho was probably responsible for the anomaly: either the terrain in and around Tycho was much rougher than that in any other equal area of the Moon's near side, or the rocks of Tycho were much denser than those elsewhere. Since that time, photogeological studies of high-resolution pictures of the Tycho area have shown that it is indeed rough.

Radar has been used, further, to estimate the porosity of lunar-surface materials. On radar wavelengths of 10 to 100 centimetres, the effective dielectric constant (its ability to store electrical potential energy) of the lunar surface lies in the range of 1 to 2.7. To calculate its actual porosity, this can be compared with the assumed dielectric constant of its solid component. Returned fines have been used to estimate that the density of the solid rock is about three grams per cubic centimetre and that the corresponding dielectric constant at 450 megahertz is about 6. These figures, with a mean observed lunar dielectric constant of 1.6, show that the average porosity of the lunar surface layers may be more than 80 percent. The results compare favourably with those from Earth-based photometry, although the two methods refer to data from different areas and different effective depths.

Relative reflectivity

Radar studies of the relative reflectivity of small areas of the Moon's surface at 2.2 metres in wavelength involving the circumlunar satellite Explorer 35 show that, on average, the maria reflect approximately 30 percent more power than the highlands; and reflectivity structure was observed in the lunabase, a particularly high reflectivity being associated with a region near the crater Flamsteed. There appears to be a significant correlation among radar reflectivity, hot spots, and regolith thickness; both the radar reflectivity and the number density of hot spots increase as the thickness of regolith decreases. The effective (mean) dielectric constant of the lunar rocks (lunarite and lunabase together) down to depths of several tens of metres is about 3, slightly more than the mean value of the dielectric constant of layers nearer the surface measured on shorter wavelengths from the Earth. The porosity of the lunar rocks should decrease, on average, with depth as, in a fragmental, or granular, regolith, the weight of the overburden will cause increased packing as the depth increases. This would explain the observed increase of thermal conductivity with depth.

Gamma-ray spectroscopy

Physical experiments from orbiting spacecraft. Early physical experiments conducted by the Soviets from circumlunar orbits included a gamma-ray-spectroscopy experiment carried aboard the first artificial lunar satellite. Gamma rays are emitted by lunar radioactive elements, particularly potassium, uranium, and thorium, and the character of the emission can indicate whether the Moon has been chemically differentiated. Cosmic rays incident on the Moon induce a much higher intensity of gamma-ray emission than that deriving from natural radioactive decay in the lunar rocks. Nevertheless, Soviet scientists deduced that the surface rocks of the Moon were of the basaltic type and that a slightly greater intensity of gamma radiation derived from the maria.

Gamma-ray spectrometry from two Apollo modules in lunar orbit achieved further results. Not only were the maria found to be more radioactive than the highlands but the western mare units (*e.g.,* Oceanus Procellarum) were, in general, higher in the radioactive elements of uranium, thorium, and potassium than were the eastern maria. In general, the higher and the more light toned the lunar rocks, the less they exhibited radioactivity. The same modules also carried X-ray fluorescence spectrometers that were used to determine the ratios of aluminum

By courtesy of National Aeronautics and Space Administration

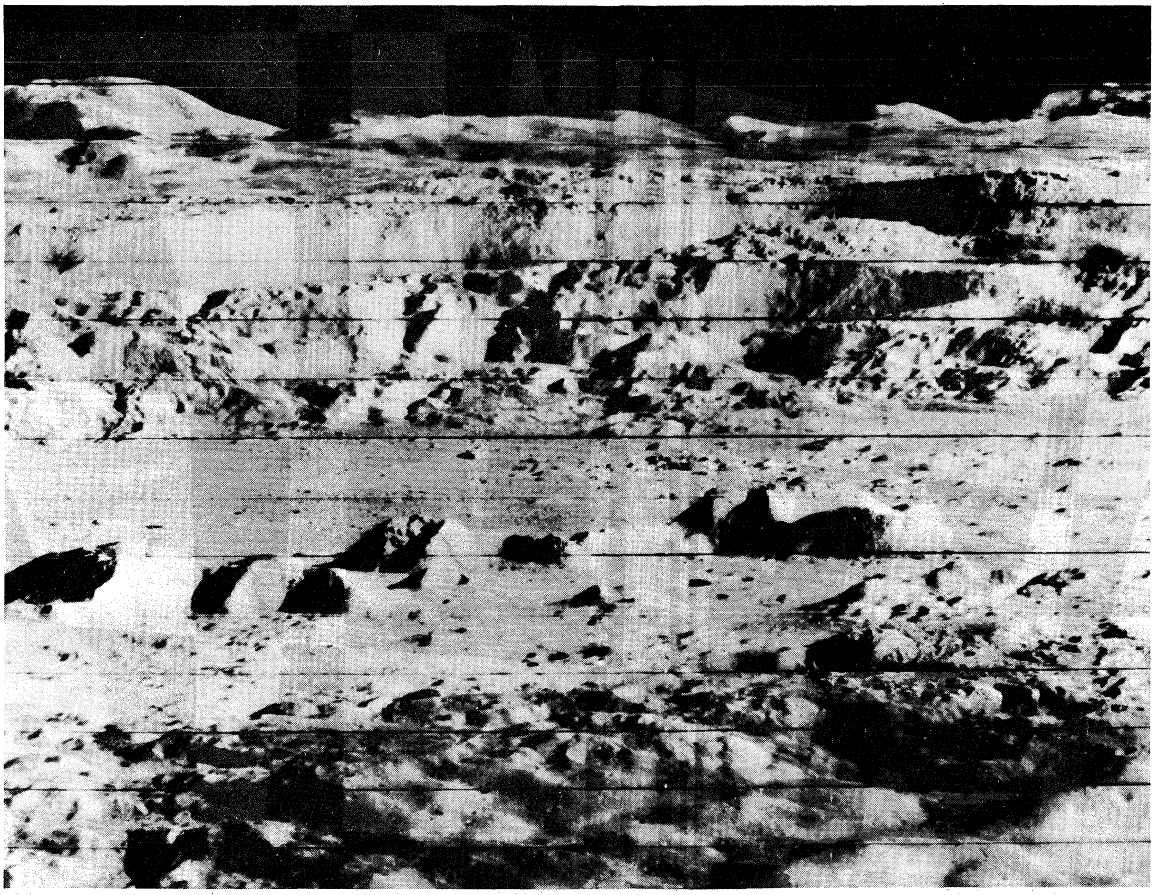


Figure 23: Crater Copernicus as photographed in November 1966 by Lunar Orbiter 2 from a height of 45.7 kilometres, looking due north from a position above the crater's southern rim. The mountains rising from the flat floor of the crater are about nearly 460 metres high; the 914-metre mountain Gay-Lussac, on the horizon, is part of the Carpathian range.

to silicon and magnesium to silicon. The lighter toned highland materials were found to be higher in aluminum and lower in magnesium than the dark-toned maria. All these results were consistent with the theory that the Moon had been chemically differentiated. Other U.S. experiments from orbit included gravity profiling—in which some of the mascons (see above), determined earlier, were evaluated in more detail and other, smaller mass concentrations were discovered—and heat flow measurements, in which hot spots and cold spots, similar to those discovered telescopically, were found.

TYPES OF LUNAR FEATURES

Lunar craters. The ubiquitous nearly circular features on the Moon have all come to be called craters; the traditional terms "ring plain" and "walled plain" describe, respectively, round and irregular lunar features with walls encompassing relatively smooth, level floors.

Forms of craters

Lunar craters, in the broad sense, may be classified, first, on the basis of their form. Many intermediate-sized craters, in particular, display mountain blocks or hills near their centres. The inner (and, to a smaller extent, the outer) walls of many craters are terraced, or benched.

Although in a lunar photograph (see Figure 21) the majority of the lips of crater rims (walls) appear to be approximately circular, it is common to find that, in detail, the actual shapes better approximate some polygon or an ellipse. A number of lunar craters have ellipticities greater than 30 percent. Some of the so-called craters have a spiral inner ringwall lower than the outer; there are craters with eccentrically placed inner craters and plateau craters that seem to have been filled with dark material.

Bowl-shaped craters with parts of the floor darker than the inner slopes are common among lunar craters less than 20 kilometres in diameter. They may be as much as three kilometres deep, with inner slopes of up to 50° and outer rims sloping at up to 20° near the lip. The shallowest "saucer" craters have the surfaces of their rims inclined to the horizontal at no more than a few degrees. Crater pits are those without raised rims. Funnel- or dimple-shaped crater pits with convex inner walls near their lips have been discovered in photographs.

Ray and dark-halo craters

A crater may be classified also by the neighbouring craters or other features that appear to be associated with it. Thus, craters such as Tycho (Figures 21 and 22), Aristarchus, Copernicus (Figure 23), and Kepler are known as ray craters, being major centres of subradial systems of bright rays that are now known to consist of numerous small, bright craters and blocks of light-toned rock. The rays of Tycho are up to 1,900 kilometres long. A number of kilometre-sized craters are known as dark-halo craters; these have associated dark-toned rings or halos around them that may be pyroclastic (volcanic) deposits. Other craters are associated with subradial linear formations that are distinct from rays and others again with adjacent positive elements of terrain known as deltoids. Dome-top craters are conspicuous because each occupies the summit of a broad, topographic swelling. Small craters often seen on the summits of hills or mountain peaks may form a class of their own (summit craters). It is common to find a high number of craters on ridges in the maria or centred in the crests of crater walls; such craters have been referred to as parasitic or ringwall craters. A crater that is one member of a linear progression, or chain, of craters may be described as a chain crater, particularly if other members of the progression are of a common and distinct morphological type. Other mutually similar craters form arcuate chains. Consanguineous craters with strong morphological similarities may be numbered among the members of a chain, but they are commonly found in close pairs also. The walls of interlocking craters may pass across one another without interruption. Finally, rille craters are those single or chained craters that are set in and form part of the lunar rilles, which are trenchlike features. A rille crater may be circular or markedly elliptical and may be the centre of a dark halo.

Height measurements

Height measurements form an essential element in the construction of a profile of a lunar feature. In principle, the height (h) of a lunar mountain peak above the lunar ground where the tip of the mountain's shadow, of some length (l), lies is determined by simple geometry provided the mountain is near the centre of the Moon's disk or measured on a photograph taken vertically. Taking the elevation of the Sun over the mountain as θ degrees, the height h of the mountain is calculated with the formula $h = l \tan \theta$; the angle θ can be obtained from regularly published tables. For mountains not viewed under these conditions, a more general formula for h is used.

The near side of the Moon has been surveyed fairly intensively, but, until recently, methods employed to measure the Moon's overall shape, or figure, have been complex and laborious. Unlike the Earth, with its oceans, the Moon has no natural datum from which to measure altitudes. The physical librations (see above) show that the Moon cannot be both spherical and homogeneous. An accurate evaluation of the Moon's figure has become important in discussions of the origin of its surface features as well as in discussions of its internal constitution.

Apollo laser altimetry, used in computing circumlunar orbits, showed that particular belts around the Moon, close to its equator, averaged a few kilometres lower than a reference circle on the Moon's near side and a few kilometres higher than that circle on the Moon's far side. In other words, there are several large, topographically low, maria on the near side of the Moon, while the far side is dominated by highlands. Individual mountains attain heights of up to four or five kilometres above their bases, and this is also the order of depth of the deepest lunar craters, commonly measured from the rim crest of the enclosing ramparts, which are in some cases more than 200 kilometres in diameter.

Smaller craters are progressively less deep: whereas the deepest and largest craters of the Moon have a depth-to-diameter ratio of only 0.08, lunar craters with a diameter of 10 kilometres may have a depth-to-diameter ratio of 0.15. The latter, however, are the eumorphic (sharply defined) craters; and all craters with a 10-kilometre diameter, taken together, show a wide spread in the depth–diameter ratio. The tendency in the case of eumorphic lunar craters is for the depth-to-diameter ratio to increase as the size of a crater diminishes, a generalization known as Ebert's rule.

Placement of craters

The deepest and largest craters of the lunar near side are concentrated in the central meridional belt of the Moon's southern hemisphere. The deepest craters tend not to be near the maria. The encircling walls of the oldest craters have slumped into the craters or have been cut across by faults or graben; the craters themselves have been overlapped and partially obliterated by smaller craters formed later. In addition, the topography of large craters is often softer than that of many of the sharply sculptured, more recent craters, presumably because of changes resulting from denudation.

Dense cratering is found in the lunar highlands, and many of the older craters are difficult to observe. About four times as many craters between 10 and 50 kilometres in diameter are found per unit area of the near-side highlands as in the same unit area of the average mare. Standard statistical methods applied to lunar-crater data have shown that the lunar craters are not distributed across the surface in the random fashion that would be expected if they resulted from meteoric impact.

The maria abut on the highland domains, and mare materials are found in craters with breached walls on the peripheral zones of maria. Dating of lunar rocks by the rubidium-strontium isotope method has established that the mare lavas are younger than the highlands, and it is clear that late phases of mare lava outpourings could blanket craters and thereby render their overall distribution non-random. The maria themselves differ in age, as stratigraphic maps and dated rocks show. There are, even in a single mare, distinct geological units (such as individual lava flows), each with its own characteristic population of craters.

An attempt by statistical methods to distinguish between impact (including secondary-impact) craters and endogenic craters in the lunar maria has been made. The ratio of internally formed craters probably varies from one unit to the next, and, of the craters with diameters between 50

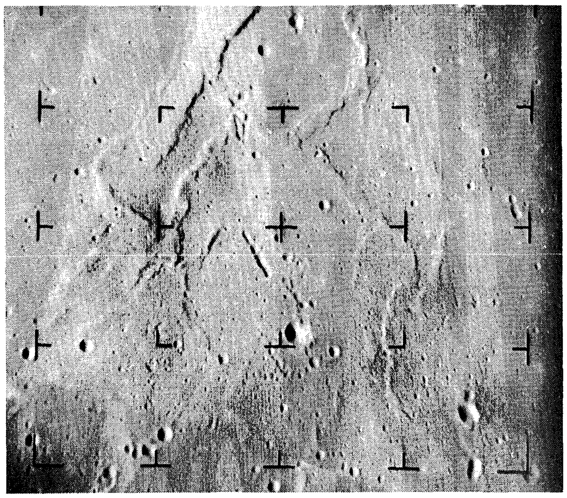


Figure 24: Part of Mare Tranquillitatis, showing chain craters and mare ridges. Area shown is about 19 kilometres on a side.

By courtesy of the Lunar and Planetary Laboratory, University of Arizona

metres and hundreds of metres, it is not uncommon for up to half to be arrayed with the trends of known tectonic lineaments (internally formed linear features). Such craters are probably of internal origin.

Lunar lineaments. *Mare ridges.* Lunar lineaments (linear features) are of many different types. Mare ridges (Figure 24) are numerous in the maria and may be much longer than the brighter highland-type or lunaritic ridges. Mare ridges are also found in dark-floored craters and, rarely, may be traced beyond a mare into highland-type material. In transverse profile they consist, typically, of broad, low rises sloping at only a degree or so for 10 or more kilometres and capped by slightly steeper, relatively narrow ridges. Typically, the major mare ridges rise to an altitude of about 200 metres above the surrounding country. They lack continuity and are broken, offset, or arranged *en echelon.* The higher, narrower parts of the ridges show a strong tendency to be in short sections, or segments. While the overall trend of these ridges in a subcircular mare is generally aligned with the curved border of radially more distant mountains, the directions of the mare-ridge lineaments often follow other preferential directions. Like the lunaritic ridges, they may also belong to families of ridges that trend over much of the lunar surface in set ranges of azimuth—a fact that labels many of them as tectonic in origin.

Tectonic origin

The lunar grid system. Other lineaments, such as linear graben (depressed regions between faults), crater chains, rilles, and faults, together with the ridge network, form the lunar grid system. Near the centre of the Moon's nearside face the principal trends of the grid system are at about 45° to the central meridian. Interference between these dominant directions and other directions of lineaments that are subradial to certain of the maria, notably Mare Imbrium, makes recognition of the Moon-wide topographic pattern difficult. Many of the mare ridges have probably been built up by extrusion of basic lavas from fractures. Others may have been the result of the intrusion of lavas. Because of their tendency to form an overall grid pattern, many of the fractures must have been produced by uniformly acting crustal stresses of tectonic origin.

Regional lineaments subradially oriented (*i.e.,* somewhat offset from a radial direction) with respect to Mare Imbrium and Mare Orientale are easy to locate; fewer ridge lineaments are associated with Mare Nectaris and Mare Humorum. Such regional fracture or fault systems extend into and are modified by the older terrain around the mare. The circular mare sites on the Moon are at present thought to have been blasted out by gigantic impacts, and the basins so created later filled with lavas from inside the Moon. Much faulting and possibly the extrusion of lavas in the regions surrounding each mare basin would be generated in such major events. Fractures in a previously unstressed homogeneous medium will be radial and tangential at some distance from an explosion site. The systematic deviations of the mare lineament systems from these directions in association with the more general grid system suggest influence from preexisting fractures or weaknesses in the lunar crust.

Similar patterns are found around certain large craters, and these may be detected most readily in mare material. Around the crater Aristillus, bundles of lineaments, rather than a set of precisely radial ridges and troughs, are found. Only the central, longest lineament of each bundle is radial to the crater, while the other members are progressively offset from the radial direction. The central lineament of a bundle tends to be aligned with one of the principal directions of the grid system. The ultimate explanation of these minor sets of subradial lineaments is expected to be analogous to that pertaining to the maria.

Where the grid system abuts on the larger craters, northeast- and northwest-trending ridges commonly intersect and sometimes pass inside the northern or southern ringwalls. Well-developed nests of crisscrossing ridges found in such localities have been called lattice patterns. A quite different feature, called a deltoid, also commonly occurs outside the northern or southern walls of certain craters. Deltoids are raised elements of the lunar terrain in the form of a delta, or of a gothic arch, the base of which abuts on the rim of a crater.

Of the subcircular maria, some tend to be polygonal. As in the case of the polygonal craters, the shape might be the result of crustal weaknesses existing before the mare was formed. The peripheries of other areas of lunabase (dark, mare-type material), including some maria, are irregular. Maria also contain features known as rilles (German *Rille:* "trench," "groove," "channel"; Latin *rima,* a term used principally on official maps), which are, however, found also in the lunabase generally, in certain of the larger craters, and in the lunarite (brighter, highland-type material).

Normal rilles

Rilles. The normal rille is a linear or broadly arcuate trench without raised rims (except when it cuts through higher ground) and with gently insloping walls. The largest rilles are five kilometres wide, a few hundreds of kilometres long, and several hundreds of metres deep. The floors of normal rilles are relatively level (unless they are so narrow that fallen rocks and dust from the opposing walls meet in the bottom) or are humped or strewn longitudinally with blocks. They slice through the mountainous walls of some craters and traverse the lunarite for great distances. They tend to occur in parallel families, in arcuate paths generally just within the mountain ranges or hills that border maria or the larger craters, and to run in preferential directions suggestive of lines of probable crustal weakness and tectonic control. Some are arranged *en echelon,* a characteristic of strike-slip faults on Earth.

Single or contiguous craters without raised rims define parts of the walls and floors of some normal rilles; such rilles have sometimes been termed crater rilles. The craters appear to be of the collapse type, with flattish floors. Lunaritic ridges traverse some normal rilles (such as the Ariadaeus and Hesiodus rilles) and appear to have sagged or subsided into the rilles. Such ridges seem to be older than the rilles that they cross without a break. These observations suggest that normal rilles are not partially filled fissures, but are graben, or collapse features, and are probably controlled by underlying fractures or faults. In many instances, normal rilles run into or connect with other normal rilles at sharp angles and with faults, chains of craters, or mare ridges, again suggesting that the rilles in question are controlled by tectonic fractures.

Sinuous rilles

Examination of space photographs stimulated much discussion of another type of lunar rille—the sinuous rille. The direction of some parts of sinuous rilles, like normal rilles, appears to be controlled by tectonic lines of weakness. Other parts have features resembling meanders in a terrestrial river. The superficial resemblance to dry rivercourses—particularly to those cut by rivers draining inland into desert areas—led to a revival of interest in the possibility of water on the Moon. But later theoretical work using detailed measurements of the geometrical parameters of sinuous rilles lent weight to another hypothesis—that these rilles were shaped by the flow of lavas. Many of the sinuous rilles have one end marked by an elongated crater that would be a probable source of lava, and many degenerate into a chain of craters that are commonly elliptical and have their longest axes set along the probable courses of underlying lava tubes. In a few cases, smaller sinuous rilles meander along the floors of larger ones, such as Schröter's Valley. Several sinuous rilles have distributaries or reentrant loops, both of which are common in terrestrial lava channels.

Like normal rilles, sinuous rilles tend to concentrate around the borders of the circular maria or at the junctions of lunabase and lunarite. Exceptionally strong concentrations of sinuous rilles are found in the densely domed region near the crater Marius and also around the crater Aristarchus. Thus, sinuous rilles show a preference to regions that have been assessed from other evidence as showing volcanic characteristics.

Lunar faults

Opposite sides of a few stretches in some normal and sinuous rilles are observed to be at different levels. In principle, a rille may form by the collapse of a weakly cohesive regolith into an open fracture. Faulting is initiated when there is relative movement between the two blocks that define the fracture. The displacement of a fault may remain constant, or there may be continuing or intermittent movement. The direction of displacement may be steeply inclined to the horizontal (normal, or dip-slip, fault); steeply inclined to the vertical (thrust fault); or lateral (strike-slip, wrench, or transcurrent fault). Faults on the Earth and Moon may show a combination of these idealized directions of displacement. On the Moon, however, no thrust fault has yet been recognized. The term "wrinkle ridge," a synonym for "mare ridge," is suggestive of compression or thrusting, but there is little or no real evidence that any of the mare ridges (Figure 24) are compressional features.

Some rilles are directly associated with probable dip-slip faults. The characteristic *en echelon* pattern of strike-slip faults is reflected in other rilles, in segments of wrinkle ridges, and in lunaritic ridge systems. More than one major strike-slip fault appears to have offset blocks of the lunar Caucasus Mountains; and there are a host of probable minor strike-slip faults on the Moon, one of the most notable being the fault that has displaced the walls of the crater Capella and of another two craters further to the south. Strike-slip faults are perhaps of more importance on the Moon than the better known dip-slip faults, such as the inappropriately named Straight Wall. Major faults, dissecting large areas of lunarite, have been recognized. These "selenofaults" measure hundreds, rather than tens, of kilometres in length.

Bright rays. The bright rays that diverge from craters such as Tycho (Figures 21 and 22) are thought by some to be fracture controlled, but the majority of lunar geologists believe the rays to be the result of ejecta of secondary particles from a primary impact site. One of the strongest arguments for the impact origin of rays is that they look like splash marks. The particles that formed the end of the longest ray of Tycho must have been ejected from Tycho at a velocity of at least 1.4 kilometres per second if the ejection hypothesis is correct. The velocity of escape of a particle from the lunar surface is 2.4 kilometres per second, so potential ray-forming particles would not escape from the Moon. The evidence from closeup photographs is that more than the average number of craters are in unit areas crossed by rays; and evidence gathered by astronauts points to there being a complement of bright-toned rock blocks in a ray-strewn zone.

Contour maps

Lunar mountains. Systematic estimates of the altitudes of lunar mountains were first made in the 18th century. The most comprehensive program of lunar height measurement ever attempted was started in 1959 by the United States Air Force's Aeronautical Chart and Information Center. The results of this concerted program, in which many thousands of lunar photographs were taken at different Earth-based observatories, are incorporated in lunar near-side charts prepared by the Aeronautical Chart and Information Center. Charts of the whole surface of the Moon, compiled using space probe data when appropriate, include the spot heights of mountains. A number of detailed contour maps are based on accurate measurements of high-quality, controlled photographs taken from lunar orbit, and laser positional measurements.

The classical program of heliometer observations of the angular separation of points on the Moon, made in the 19th and early 20th centuries, is of fundamental importance in the present understanding of the lunar figure and physical libration. The time-consuming reductions involved the computation of the distances between Mösting A (a crater near the centre of the Moon's disk) and many control points scattered over the Moon's near side and, finally, solution for constants characterizing the librations.

Transient lunar phenomena. Ever since serious telescopic observation of the Moon began, astronomers have reported mistlike phenomena or temporary obscuration of detail in specific patches of the Moon. In 1966 a significant positive correlation was found between the times of more than 100 observations of transient lunar phenomena made between 1749 and 1963 and the distance of the Moon from the Earth. The majority of phenomena occurred close to the time of the Moon's perigee or apogee passages. The indication was that the Earth's tidal forces on the Moon's solid crust had contributed to the cause of the transient lunar phenomena, although the phenomena themselves may be manifestations of seismic (moonquake) or seismically triggered volcanic activity or merely caused by dust either blown out from the regolith and dispersed by cold gases escaping from cracks in the Moon or set in motion by seismic activity. Similar periodicities in some types of seismic signal from observing stations planted on the Moon show that lunar seismicity occurs most commonly within a day or so of perigee and of apogee, and this suggests that the majority of the transient lunar phenomena are connected with some kind of seismic activity.

Other kinds of lunar change have been suggested. Perhaps the most frequently discussed case of a possible permanent or long-term change of topography is that of the crater Linné, which may be located on many photographs as a whitish patch in Mare Serenitatis and appears on the best photographs as a crater about 1.5 kilometres in size,

By courtesy of National Aeronautics and Space Administration

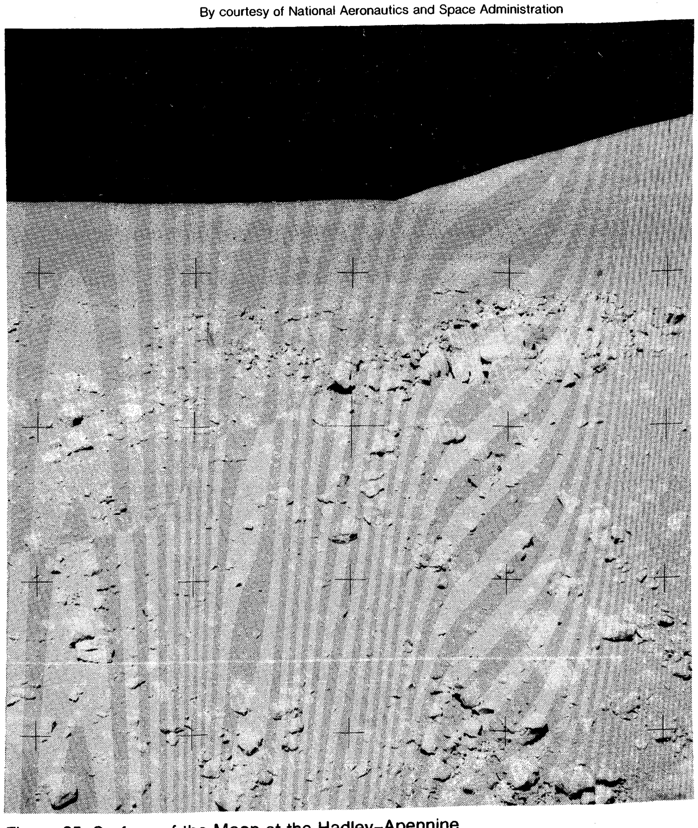


Figure 25: Surface of the Moon at the Hadley-Apennine landing site, photographed during the third Apollo 15 EVA (extravehicular activity), August 2, 1971.

centred in a bright halo. It cannot be concluded from the evidence available that the area has suffered any morphological change.

Observed eruptions

In 1958, a Soviet astronomer, N.A. Kozyrev, claimed to have detected a reddish cloud of volcanic ash, followed by an efflux of gas lasting for at least half an hour, from the vicinity of the central mountain block of the crater Alphonsus. He recorded a lunar emission band having its maximum intensity at 4,737 angstroms and gradually fading toward the violet end of the spectrum. This and other bands, he proposed, had their origin in carbon molecules. Some volcanoes on Earth emit carbonic gases. In 1959 Kozyrev again obtained an unusual spectrogram of Alphonsus and maintained that the central block of the crater had given birth to an eruption of lava at a temperature exceeding 1,270° C. Although these observations and interpretations have not been universally accepted, the reports did stimulate interest among other professional astronomers in a Moon long considered to be dead and changed only by impacts.

Notable reddish glows on the crests of ridges in the Aristarchus region were observed in 1963 at the Lowell Observatory. A month later, similar observations of the Aristarchus region were reported at two observatories. At this time, coloured patches persisting for more than an hour seem to have been authenticated.

THEORIES OF ORIGIN OF THE SURFACE FEATURES

In the early 17th century the lunar craters were thought to be endogenic; that is, of internal origin. The first part of the 19th century saw truly under way the argument that most craters could have been caused by the impacts of external bodies or by volcanism. Theories that could not be classified as either volcanic or meteoric (impact) have also been proposed; but lunar craters are now generally thought to include impact craters of all sizes, large impact craters modified by volcanism, and endogenic craters, mainly less than one kilometre in diameter.

Gas-bubble theory

Volcanic theory. One of the earliest hypotheses of the formation of the lunar craters was the so-called gas-bubble theory of Robert Hooke, who proposed in 1655 that they were formed by the collapse of huge bubbles of gas; the discharge of gas is a vital part of the volcanic process. On Earth, such bubbles that occur in fluidized mud, as at La Solfatara, Italy, commonly measure centimetres or decimetres in diameter, increasing in size with the viscosity of the containing medium and the gas pressure. In Iceland there are hollow domes of lava measuring metres in diameter; on the Moon the low gravity and lack of atmosphere could lead to significantly larger domes. In a later theory that was more attractive as far as the larger lunar craters were concerned, the starting point was thought to be a conventional eruptive centre such as might have given rise to a cone volcano of the Vesuvius type. According to this theory, volcanic materials would have been ejected from the central orifice in a symmetrical fountain and would have landed to construct, layer on layer, a roughly circular ringwall. Sometimes, dying activity would build up a central eminence; in other cases, the floor of a crater would be flooded by dark lavas.

Comparison of lunar features with terrestrial ones

Volcanoes on Earth generally produce domes, or cratered cones, rather than rimmed depressions. There are many tuff or lava and tuff rings of kilometre size in terrestrial volcanic fields, however; for example, in Iceland, Hawaii, Arizona, and Mexico, where central collapse or engulfment of materials tends to leave landforms resembling many of the lunar craters.

Both the bubble hypothesis and the volcanic-fountain theory were argued without any appreciation of the physical or geological principles involved. More competent geological arguments for lunar volcanism and caldera collapse were proferred in 1944 by Josiah Edward Spurr, who developed a theory that the larger lunar craters were volcanic collapse structures (calderas); he developed a classification of craters and described in detail the different modes of origin of different classes of crater. More importantly, he was the first to recognize the uniformity and extent of the overall structural pattern in the Moon's crust and to name it "the lunar grid system." Spurr's analyses were closely tied to his meticulous photogeological analyses of the lunar craters. Calderas may have developed during an intensive phase of lunar outgassing, while many of the low-walled ring formations on the Moon marked the positions of ring fractures from which lavas had been erupted in the precaldera phase of lunar crater formation.

Impact theory. The impact hypothesis was formulated comparatively late. Few people had seen a large meteoroid fall on Earth, though many had witnessed terrestrial volcanic eruptions, and there was practically no evidence for live, meteoric cratering events on the face of the Moon. Even the geologist who had been one of the first to argue convincingly (in 1893) for impact craters had disputed the impact origin of the world's best authenticated meteoritic crater, in Arizona. Until recent years, few craters on Earth had been described as of probable impact origin; many structures on Earth now regarded as impact craters had been termed cryptovolcanic craters by geologists. The predominantly circular shape of lunar craters was thought to be impossible to produce by impact, except from relatively rare, vertical impacts. It is now believed that a nearly circular crater would be blasted out of the lunar rocks by a large meteorite arriving at a high speed even if the angle of approach to the Moon's surface were far from the vertical.

This, essentially, is the basis of the modern impact theory. A vast array of statistical data on both terrestrial and lunar craters was collected, compared, and analyzed by Ralph Baldwin, who developed a thesis on the assumption that the great majority of the lunar craters were of impact origin, and who argued for the cataclysmic formation of Mare Imbrium following a major collision between a large object and the Moon that resulted in a gigantic explosion. Later, it was reasoned that the nearly circular maria, all the large craters, and the majority of the small ones were impact scars and that later lavas, flooding impact basins such as that of Mare Imbrium, flowed further afield than the confines of the basins and created the irregular maria. Later impact theories differed considerably from Baldwin's in some details.

Impact velocities

The velocities of interplanetary meteoroids in the vicinity of the Moon are commonly of the order of tens of kilometres per second. Because, in such a meteoric-impact process, the energy is transformed and released very rapidly, little of the initial kinetic energy is converted into heat, and the rocks in the floor of a crater are not likely to be melted to any appreciable extent. A longer period of time for the impact is needed to give the Moon rocks time to absorb more of the heat that would otherwise be dissipated in other ways.

A ring 1,000 kilometres in diameter (such as Mare Imbrium; Figure 21) could be formed by impact of a high-velocity meteoroid of less than one-tenth of that diameter or with a larger, lower velocity object. Asteroids, for instance, would collide with the Moon at relative velocities around six kilometres per second. Objects circling the Earth in a ring at about the Moon's distance could collide with the Moon at velocities as low as the velocity of escape from the Moon's surface, 2.4 kilometres per second.

Evidence against large-scale melting

H.C. Urey developed an argument based on the concepts that the Moon was nonspherical and that it had a nonequilibrium shape (even allowing for the Earth's tidal effects) and a peculiar mass distribution. The most important irregularity in the Moon's distribution of mass was found in the excess mass in or around the axis directed toward the Earth. Support of the excess mass in the Moon's Earthward-pointing axis implied considerable mechanical strength. If the Moon had undergone any appreciable melting phase in the past, it would have had insufficient strength in its deep interior to support the supposed bulge. This was regarded by some scientists as an indication that the Moon could never have been melted to any significant extent. The mare lavas would have been derived from the impacts themselves, and, hence, the impacting bodies must have been large, low-velocity objects. The observed asymmetric pattern of radial grooves and ridges beyond the periphery of Mare Imbrium could be generated by splashlike ejecta following an oblique, low-velocity collision, whereas a high-velocity collision at a similar angle of impact would produce an ejecta pattern essentially symmetrical about the explosion centre. It is to be pointed out, however, that one prediction of this theory was that the Moon had not undergone widespread igneous differentiation, whereas such differentiation is now known to have occurred.

Modern views on the origin of the craters have been much influenced by Baldwin's statistical work. He showed that the ratio of the depth to the diameter of terrestrial shell, bomb, and meteoric craters and of the craters of the Moon could all be fitted on smooth curves showing logarithmic plots of the ratio against the diameters of the craters. Terrestrial calderas (collapsed volcanic structures) also fit the depth-to-diameter curves of Baldwin. Further studies extended Baldwin's data, relating the ratios of rim widths of lunar craters to their diameters to those for volcanic craters and calderas on Earth; the six calderas studied had much wider rims than lunar craters of the same diameter (1–10 kilometres), though this may be because of underestimations of the lunar rims.

The shape of plots of the logarithm of cumulative number of craters larger than a certain diameter against the logarithm of that diameter has been said to be evidence for impact origin of the craters; probable amounts of meteoric debris in the solar system would collide with the Moon to produce craters giving a similar statistical result; that is, the numbers of lunar craters of various diameters seem to be relative to the numbers of meteoric bodies of various sizes, in space, available to produce the craters. Others, however, found that a high proportion of craters less than one kilometre in diameter were not impact craters and that their proportion varied with position.

Existence of both impact and volcanic features

There is now no real doubt that both impact and volcanic features exist on the Moon. Many details and some major problems remain to be solved. The origin of the circular maria, as has been suggested, may be that basins were created by impacts and were later flooded by lavas from within the Moon. The same argument may apply to many of the larger craters; following an explosive cratering event, the terrain beneath the excavated material will be fractured and will tend to adjust upward, while the extra mass deposited in the form of ringwalls will tend to adjust downward. Through the friction, gravitational energy will be dissipated under the crater as heat. This heat might be sufficient to melt the underlying rocks, which would then have access to the surface through fractures. The volcanic floor of Tycho (Figure 22)—and possibly its central peaks (which might also have been formed by impact)—may have been produced in this way. For Mare Orientale, there is much to be said for an explosive origin, since, for one thing, many characteristics of the surrounding terrain may be so explained. Other maria may be predominantly lava-filled calderas of igneous origin. The maria differ in age and were not all formed in one short-lived catastrophe.

(G.Fi.)

THE ORIGIN AND EVOLUTION OF THE MOON

The evolution of the Moon's orbit. The Moon is gradually changing its orbit because of a continual increase in its angular momentum. This increase, in turn, is due largely to the Moon's own effect in raising tides on the Earth. The action of both lunar and solar tides causes an energy loss through internal friction, transforming mechanical energy into heat; most of the dissipation of energy is thought to occur in shallow seas, principally the Bering Strait. The immediate effect of tidal friction is a slowing of the Earth's rate of rotation; the length of a terrestrial day, as determined by observation, is increasing by about 2×10^{-3} seconds of time per century.

As its spin slows, the Earth loses angular momentum. The total angular momentum of the Earth–Moon system, however, must remain nearly constant (there is a small increase from the Sun's tidal action), and the lunar orbit must acquire that which is lost by Earth's rotation. This is accomplished by the gravitational action of the solid Earth tide, exerting a force in the direction of the Moon's orbital motion. According to Kepler's laws, this must cause the Moon to move ever farther from the Earth, while the Moon's orbital angular speed decreases. Because it amounts only to a few centimetres per year, the Moon's recession had not been observed directly by the early 1980s, although it is expected that laser ranging observations will soon detect it. "Tidal acceleration" was first measured by Sir Harold Spencer-Jones in 1938; his value of −22 arc seconds per century^2 differs by only 10 to 15 percent from values obtained with the most modern techniques; much higher values have been published on the basis of erroneous analyses, however. Precise determination of the tidal effect is complicated by the possibility, required by some theories of relativity, that the gravitational "constant" is changing with time; if true, this would cause an acceleration that is distinguishable from the tidal acceleration only with difficulty.

When discussing changes in the Moon's orbit, it is important to bear in mind that the hour is a defined quantity and therefore constant, and that the year can be considered as having been essentially constant since the formation of the Earth, or shortly thereafter. The day, however, did not always have 24 hours, nor was the synodic month always 1/12.4 of a year. Astronomical observation, as all human activity, has taken place in a miniscule fraction of the history of the Earth, and there is no direct evidence of how the decelerations of the Earth and Moon proceeded in the distant past. Some scientists have interpreted the growth rings of tidal life forms (such as mollusks, corals, and stromatolites) as evidence that both the day and the month have lengthened since the Paleozoic Era. They estimate that, 400,000,000 years ago, the day was only 22 hours long and there were a bit more than 13 months in a year; these estimates give an average rate of slowing that is near to presently measured values. According to these data, the Moon appears to have been at a distance of 30 Earth radii, only half its present distance, about 2×10^9 years ago, with 34 months per year and 10 hours per day.

Simple mathematical calculations of the lunar orbit into the distant past, assuming the Earth's frictional response to be constant, lead to the conclusion that the Moon would have been at a distance of only 2.8 Earth radii 2,000,000,000 years ago, a distance that is closer than that of communications satellites; the day would have been only 5 hours and the month only 6.5 hours in length. Such calculations are doubtful because they are contradicted by geological evidence, they assume a constancy in the Earth's rheological properties that is physically unrealistic, and they include an unreliable time-scale. In 1982 D.J. Webb employed an idealized hemispherical ocean for a mathematical model in which the Earth's frictional response increased rapidly as its rotation decreased. In any case, it is reasonable to assume that the Moon was much closer to the Earth in the geologically distant past, and this fact has given rise to several theories of the origin of the Moon.

Dynamical theories of lunar origin. The physical mechanism for tidal evolution was recognized around 1880 by Sir George Darwin. There followed three fundamentally different hypotheses of how the Moon was formed. Plausible qualitative arguments have been advanced for each one, but each has major physical and/or geochemical flaws, and the problem remains unsolved.

Terrestrial fission. Sir George Darwin proposed the simplest possible extrapolation of tidal friction, that the Earth was formed as a rapidly spinning fluid spheroid that became rotationally unstable because of solar tides and then broke in two, the smaller piece receding from the larger ever since because of tides between the two. Harold Jeffreys showed in 1930 that such instability was not likely, and the theory was little considered for three decades. Because the density of the Moon is about that of the Earth's outer layer, or mantle, it was proposed in the 1960s that the Moon was ejected from the solid Earth, leaving the vast depression known as the Pacific Ocean basin. This theory was proved untenable by recent dating that shows the lunar surface to be nearly as old as the Earth's and the ocean floor to be much younger. A number of more recent variations on the fission theory have appeared, but the ejection mechanism is always unclear, and none of these theories can account for the present obliquity of the lunar orbit.

Untenability of the fission theory

Circumterrestrial coagulation. A theory of the Moon's origin from a "circumterrestrial swarm"—a respectable extension of the condensation theory of the origin of the solar system—was proposed in 1950. According to this theory, bits of dust, meteorites, and larger bodies are to have come together to form the primitive Earth and a swarm of debris; the remaining matter was then to have precipitated to form the Moon. Differing densities were explained by differential precipitation, the heavier metallic content of the "swarm" going into the Earth, leaving mostly silicates for the Moon. Variations on this idea include a "many Moon" theory, in which the circumterrestrial debris coagulated into several Moon-sized objects, some of which are accreted onto Earth and only one surviving the hazards of coexistence. Dynamical problems with such theories include the accumulation of the Moon's present angular momentum, the obliquity of its orbit, and the accumulation of two planets instead of one. Present gravitational theory suggests that the formation of one large object will sweep up all loose debris, blocking the possible formation of other large bodies in the vicinity.

"Many Moon" theory

Orbital capture. Sir George Darwin calculated the lunar orbit backward in time to a point at which it was extremely close to the Earth; he then took an intellectual leap to the fission hypothesis. In 1955, Horst Gerstenkorn continued Darwin's calculations. According to his result, the Moon did not come from the Earth, but from the distant reaches of the solar system, and it was captured into its present orbit during a chance close approach toward the Earth. The assumptions made in Gerstenkorn's calculations were physically implausible; they have since been replaced by improved models in which the major unknown factor is the dynamic and viscous response of the Earth to such an encounter. Extreme disruption and melting is implied, so the encounter would have happened very early in the history of the solar system, before the formation of the lunar maria some 4,000,000,000 years ago. According to the assumptions made in the calculations, the event can be located at any time in the distant past, and the Moon can be made to come from almost any direction, with either a direct or retrograde initial orbit. If a retrograde orbit is assumed, the action of the tides would cause the inclination of the orbit to flip over the Earth's poles, bringing the Moon into its present direct orbit. The most serious objection to this theory is the requirement that such a massive object as the Moon come from elsewhere in the solar system and pass within a few radii of the Earth. It has been said that the only dynamical objection to capture is its improbability. (J.D.Mu.)

Evidence from the composition and physical properties of the Moon. *Chemical analysis of lunar samples.* The first orbital measurements of the broad composition of the rocks covering the Moon's surface served to confirm the speculations of photogeologists, who had argued for the presence of basic rocks, such as the mare basalts.

By courtesy of National Aeronautics and Space Administration

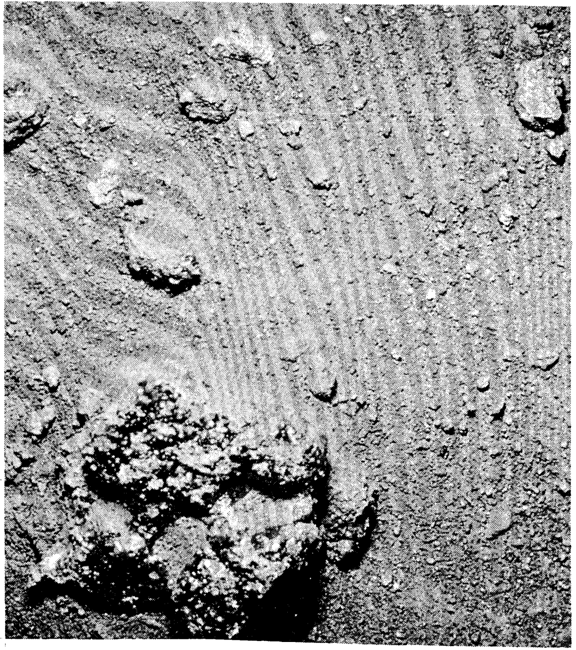


Figure 26: Closeup of a lunar rock taken during the Apollo 12 EVA on the lunar surface. The image scale is half actual size.

Detailed and more accurate measurements of the compositions of lunar samples were made *in situ,* in Mare Tranquillitatis, in Sinus Medii, and on the edge of a flow unit in the highlands near Tycho beginning in 1967. The findings indicated compositions close to those of some basalts on Earth, but there were differences between the highland analyses, on the one hand, and the earlier analyses in two patches of the lunabase. In the highland site, undisturbed lunar soil, disturbed soil, and a small rock block were analyzed separately. The rock block had a chemical composition that closely paralleled that of the undisturbed soil except for the elements magnesium and iron, which were slightly less abundant in the rock, and aluminum, which was slightly more abundant. Taken as a whole, the results indicate that the Tycho flow unit is compositionally similar to anorthosite, a granular igneous rock on the Earth.

These results led many authors to suggest that the lunar highlands in general were higher in aluminum and calcium and lower in the iron-group elements (those with mass numbers between about 47 and 65) than the maria. More iron in the maria could contribute to their generally darker tones and could also contribute to the higher rock density of the mare units.

Chemical differentiation

By 1968, scientists were beginning to appreciate the fact that the Moon had more in common with the Earth than many of them had previously been prepared to accept. It appeared that the Moon had suffered some degree of chemical differentiation and that its face presented not only physiographic domains but also true geological domains (the highlands and the lunabase) that differed in density. Furthermore, if these areas were to be established as separate geological domains, it would be clear that the Moon had been differentiated and heated. The evidence for the mare–highland rock contrast, however, was slender. First, it was based on only one analysis in the highlands. Second, the rocks analyzed were not necessarily typical highland material. Magnetic experiments indicated that the lunar samples were of basalt with little (about 1 percent by volume) or no addition of meteoric nickel–iron. This level of iron is less than would be expected if meteoroids were the only agent responsible for breaking up and gardening the lunar rocks.

The first samples of Moon rock to be returned to the Earth for exhaustive study were collected in Mare Tranquillitatis in July 1969. Since that time, rocks from Oceanus Procellarum, Mare Fecunditatis, the Fra Mauro highlands (Figure 26), Hadley Rille, the highlands near Mare Crisium, the southern highlands (near Descartes), and the Taurus Mountains have been returned to the Earth and have been analyzed.

All the rocks are volcanic basalts, except for small meteoric, anorthositic, and glassy components in the surface debris. The coherent, basaltic rocks subdivide into those of vesicular, crystalline type and breccias (consisting of compacted regolith materials), although the breccias and regolith contain fractions of meteoric nickel–iron, anorthosite, and glass, and, among themselves, glass particles show a range of composition that includes that of basalt.

Evidence that meteoroids have been an important factor in shaping the Moon is found in the layering of returned core samples and in the shocked minerals present in the returned rocks. The compositional similarity between the anorthosites and the rocks of the Tycho highlands pointed to the anorthosites and some of the shocked materials found in the mare regions having been ejected from the highlands; following later missions, the highlands were found to consist of anorthositic gabbros, including anorthosites, norites, and troctolites—the so-called ANT suite of differentiated rocks—as well as numerous surface breccias. The glasses that have undergone selective volatilization and that range widely in overall composition have probably derived from the melting of materials at a variety of impact sites. Certain glass globules of more uniform composition may be volcanic in origin.

Density at different depths

The astronomically determined mean density of the Moon as a whole is 3.34 grams per cubic centimetre, and the mean density of the basalts returned from the Moon's surface has a similar value. Yet the Moon cannot be composed of these mare basalts all the way through, since, at the pressures existing at moderate depths in the Moon, such materials would suffer a phase change and transform to eclogite, which has a density of 3.7 grams per cubic centimetre. One hypothesis is that the lunar highlands are composed of anorthosites of density between 2.8 and 2.9 grams per cubic centimetre, and certainly the bulk of the Moon's outer parts must be of rock with lower density than basalt. Yet, by comparison with the average chondrite (the most common meteorite), the returned lunar basalts are, in general, enriched in silicates and oxides and depleted in iron-group elements, and this is again consistent with the theory that the Moon has undergone general magmatic differentiation that might have led to the formation of a nickel–iron core. Dynamical considerations limit the mass of iron in any lunar core to 6 percent of the Moon's mass, so that the core, if it exists, must be very small compared to that of the Earth.

Differences in the abundances of the elements are observed at different lunar sites. Some differences could be related to crystal fractionation in near-surface melts. If the final products of fractionation had neared the lunar surface, the residual liquid rocks might have crystallized as granites; and a few such fragments have been reported. In the rocks from Mare Tranquillitatis, titanium (occurring principally as ilmenite, $FeTiO_3$) was found to be much more abundant than in either chondrites or most terrestrial rocks. Evidently, the lavas in Mare Tranquillitatis contain more titanium because they were extruded at a stage of fractionation in which certain elements, including titanium, were enriched in the melt.

The principal minerals of which the outer layers of the Moon are composed appear to be pyroxene, plagioclase, and olivine, all of which are common on the Earth. One of the most striking differences between the Earth and the Moon is the Moon's enrichment in refractory elements and its depletion in volatile materials. In particular, water—common elsewhere in the solar system—is essentially absent from the Moon in either free or combined form. The analysis of returned samples led to the discovery of a few lunar minerals that had not been found on the Earth: the first was named Armalcolite for the astronauts Neil Armstrong, Edwin Aldrin, and Michael Collins.

Rocks collected from the Fra Mauro site were the first found to be high (with reference to rocks from other parts of the Moon) in basalts enriched in potassium, rare-earth elements, and phosphorus. The element europium was found to be less enriched than most other rare earths in the mare lavas. Later, it was confirmed that europium was more prevalent than many other rare earths in the highland rocks. Together with the evidence for the divalent state of most of the lunar europium, this argues for past lunar magmatic processes occurring on a large scale in a chemically reducing environment. Indeed, it is the comparative abundance of elements, rather than of minerals, that distinguishes the Moon rocks from those of the Earth (or from chondrites) and identifies the Moon as a special object with its own geochemical evolution.

Origin of mare lavas

Formation of the maria and highlands. The geological evidence for magmatic differentiation, partial melting, and crystal fractionation seems to rule against the theory that the mare lavas were generated by impact; rather, it seems that they came from deep within the Moon. Furthermore, the lavas were extruded at widely different times. Standard rock-dating methods showed that, broadly speaking, the crystalline rocks from the maria solidified at different times lying in the interval 2×10^9 years to 4×10^9 years ago. At the present time, certain radioisotopes of the elements potassium, uranium, and thorium are, together, the most important source of radioactive heating in the Moon. Calculations relating to the heat flow through, and net heating of, the Moon have been conducted for plausible compositional models of the Moon, and it is likely that the mare lavas derived from magmas heated by this kind of radioactivity, since the thermal models show that the Moon would probably have reached its state of maximum heat content at about 2×10^9 years ago. The extent of the melting would depend on the initial temperature of the Moon, but the period of maximum heat would appear to approach, or coincide with, the outpourings of the mare lavas. There are many examples of major lava flows in the maria, the most notable in Mare Imbrium.

By contrast, most of the samples from the lunar highlands appear to have solidified 4.6×10^9 years ago. The Moon's highlands are almost certainly much older than any of the Earth's rocks. Although these ages confirmed the view of lunar stratigraphers that the lighter toned highlands were overlain by the mare lavas, the fact that the highland suite of rocks was now observed to be differentiated posed the interesting problem of the source of the heat that led to their melting and very early differentiation. If the Moon were formed by accumulation, the gravitational energy of the accumulation would provide an early source of low-level heating. A more promising source, however, may be found in the decay of short-lived radionuclides, such as aluminum-26. In principle, an early interplanetary magnetic field might be capable of heating the Moon by an inductive process. Finally, the presently observed secular acceleration of the Moon and the general tectonic (grid) system of fractures observed on the Moon's surface appear to require that the Moon was once much closer to the Earth during its early evolution; and the dissipation of tidal energy in the Moon would provide yet another way of heating it. Whatever the solution to this problem of the formation of the lunar highlands, there seems to be a record of lunar magmatic processes in operation long before any processes that can be deduced directly by terrestrial geological studies. Possibly this early heating phase, and its associated volcanism, was of some relevance in the loss of volatiles from the Moon; but an alternative theory is that the volatiles were lost before the Moon accumulated as a single object. The crater counts in distinct geological units suggest that parts of Tycho and Aristarchus are considerably younger than the surrounding materials, so there may be a dispersion of ages in highland material as well as in mare material.

Remanent magnetization

Magnetism. Remanent magnetization has been found in lunar crystalline-rock samples. Further studies of the specimens have raised questions about the origin of that magnetization. Rocks may have been magnetized as a result of electrical discharges in the primitive solar system or by the early field of the Earth when the Earth–Moon distance was small. There is some evidence, however, that the early magnetizing field was associated with the dynamo effect of a hot lunar core. Evidence drawn from surface instruments has been combined with that derived from other sources to deduce the bulk electrical conductivity and temperature profile of the lunar rocks at different depths. Although the electrical conductivity of rocks has a low value near the surface of the Moon, it increases to a depth of the order of 200 kilometres. A mantle with the conductivity of basalts and a core with the conductivity of olivine-like materials would be consistent with these observations. The precise temperature of the core is not yet decided; the core's temperature must be related, for example, to the initial temperature of accumulation of the Moon and to the extent and rate of redistribution of radioactive elements in the upper layers of the Moon. The elemental abundances of the stony meteorites and of the rocks of the Earth seem to be much the same, but they differ from the overall abundances of elements on the Moon. (G.Fi.)

Mars

Mars, symbol ♂ in astronomy, is the fourth planet in order of distance from the Sun and seventh in order of diminishing size and mass. It completes its orbit about the Sun once in 687 Earth days and spins on its axis with a period only 37 minutes longer than a terrestrial day; it has become common to refer to a Martian solar day as a sol. Orbital data for Mars are listed in Table 9; physical parameters in Table 11.

Mars is named for the Roman god of war. This astrological association of the planet with destructive forces can be traced back as far as 3,000 years ago to Babylonian astronomer-astrologers, who named the planet Nergal for their god of death and pestilence. The early Greeks named the planet Ares, for their god of battle; the planet's two satellites, Phobos (Fear) and Deimos (Terror), were named for two of the sons of Ares and Aphrodite.

HISTORY OF OBSERVATION

Work of Johannes Kepler

Mars was an enigma to ancient astronomers, who were bewildered by its apparently capricious motion, sometimes direct, sometimes retrograde, across the sky. Johannes Kepler solved this problem in 1609. Using Tycho Brahe's superior naked-eye observations of the planet, he was able to deduce empirically its laws of motion and so pave the way for the modern gravitational theory of the solar system. Kepler found that the orbit of Mars was an ellipse along which the planet moved with non-uniform, but predictable, motion. He was able to demonstrate why the planet's motion had confounded earlier astronomers who had based their theories on the older Ptolemaic idea of hierarchies of circular orbits. In the words of Kepler: ". . . Mars alone enables us to penetrate the secrets of astronomy which otherwise would remain hidden from us."

Early visual observations. The earliest telescopic observations of Mars in which the disk of the planet was seen were those of Galileo Galilei, who, in 1610, noted that the planet did not appear to be perfectly round. In 1636, Francesco Fontana produced the earliest drawings that show markings on the disk of the planet. Some of these drawings show Mars in the gibbous phase, *i.e.,* with more than half the visible disk illuminated. Unfortunately, the markings drawn by Fontana bear no resemblance to the planetary details now known, and Christiaan Huygens is credited with the first accurate drawings of surface markings. In 1659, Huygens made a drawing of the planet showing a major dark marking on the planet now known as Syrtis Major. Polar caps on the planet were first noted by Giovanni Domenico Cassini around 1666.

The accomplishments of visual observers were many. The tenuous Martian atmosphere was first noted by Sir William Herschel, who also measured the tilt of the planet's rotation axis and first discussed the seasons of Mars. The rotation of the planet was discovered by Huygens in 1659 and measured by Cassini in 1666. Cassini found the rotation period to be 24 hours and 40 minutes, which is in error by only three minutes. The moons of Mars were also discovered visually, by Asaph Hall in 1877.

Visual observations accumulated a rich compendium of meteorological and seasonal phenomena that occur on Mars, such as numerous cloud phenomena (yellow, blue, white, and gray clouds), the waxing and waning of the polar caps, seasonal changes in the colour and extent of the dark areas, a "wave of darkening" in the markings that sweeps across the planet in phase with the melting of the polar caps, and a green haze and bright transient spots. But the explanation of most of the phenomena had to await the scientific exploration of Mars by spacecraft.

Space probes. In the interval 1960–80, space missions to Mars were the major objective of the U.S. and Soviet programs for the exploration of the solar system. In an extended scientific program consisting of six missions, U.S. spacecraft successfully flew by the planet (Mariners 4, 6, and 7), orbited the planet (Mariner 9 and Viking 1 and 2), and placed lander modules on the planet (Viking 1 and 2). Three Soviet probes (Mars 2, 3, and 5) also investigated the planet, two of them reaching its surface. Before orbiting the planet, on November 27, 1971, Mars 2 placed into the Martian atmosphere a capsule bearing the Soviet coat of arms that may or may not have reached the surface. Mars 3 was the first spacecraft to soft-land an instrumented capsule on the planet on December 2, 1971; landing during a planet-wide dust storm, the device returned data for approximately 20 seconds. The Mars 5 mission entered orbit about the planet in February 1974 and returned important data that are believed by Soviet scientists to demonstrate a very weak magnetic field on Mars. The Mariner 9 and Viking missions played a major role in the development of planetary science during 1970–80. Much of the discussion in this section is based on the results of these missions.

Mariner 9 was placed in orbit about Mars in November 1971 and operated until October 1972. It returned a wide variety of spectroscopic, radio propagation, and imaging data. Some 7,329 pictures covering 70 percent of the surface demonstrate a history of widespread volcanism, ancient fluvial erosion, and extensive tectonics. The largest known volcano in the solar system, which is possibly still active, has a summit that rises 27 kilometres above the mean surface of Mars.

The Viking missions

The central theme of the Viking missions was the search for extraterrestrial life, Mars being the most likely place. No signs of biological activity were found, but the various instruments on the four spacecraft (two orbiters and two landers) returned an enormous wealth of detailed knowledge concerning Martian geology, meteorology, and aeronomy that encompasses a wide range of Martian environments and seasons. Vikings 1 and 2 were placed into

By courtesy of the U.S. Geological Survey

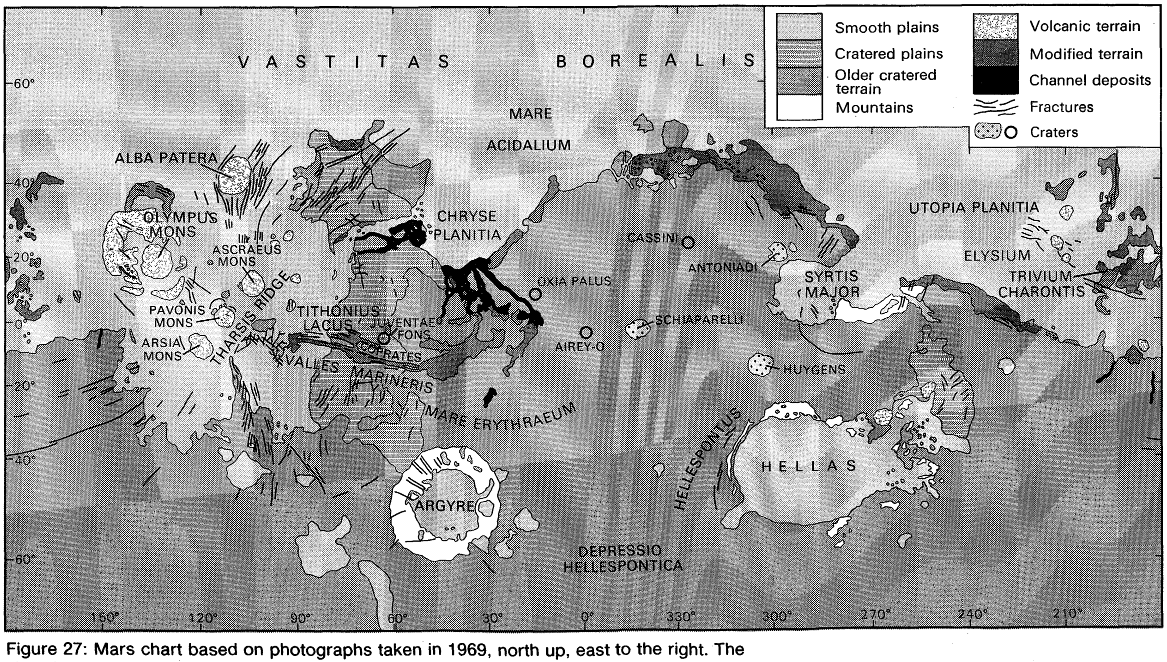


Figure 27: Mars chart based on photographs taken in 1969, north up, east to the right. The chart is a Mercator projection and extends from latitude 70° N to 70° S. It is centred on longitude 0°.

orbit during June and August 1976, respectively. Lander modules descended to the surface from the orbiters after suitable sites had been determined. Viking 1 landed in the region of Chryse Planitia (22° N, 48° W) on July 20, 1976, and Viking 2 landed in Utopia Planitia (48° N, 226° W) on September 3, 1976. All of the spacecraft operated far beyond expectations, with the last pictures from the Viking 1 orbiter being transmitted on August 7, 1980, more than four years after its arrival at Mars. The Viking 2 orbiter ceased to operate in July 1978. The Viking 1 lander was still operational in 1981; it was programmed to operate on a severely limited basis until early 1990.

Martian maps. The first known map of Mars was produced in 1830 by Wilhelm Beer and Johann H. von Mädler. The earliest system of longitude and latitude, not too different from that in use today, was defined by these pioneers. Better maps with the first systematic attempts to name the markings on the planet were produced by Richard A. Proctor (1867), Nathaniel E. Green (1877), and Camille Flammarion (1877). The nomenclature for the markings was based on the names of prominent scientists and observers of the planet such as William Herschel, William R. Dawes, Pierre-Simon Laplace, Wilhelm Beer, Dominique Arago, and Johann Schroter.

First modern map of Mars

Giovanni Schiaparelli produced what might be called the first modern astronomical map of Mars in 1877; and it contained the basis of the system of nomenclature still in use. The names on his map are in Latin and are formulated predominantly in terms of the ancient geography of the Mediterranean area. This map also showed, for the first time, indications of a system of canals, or channels, on the bright areas. Schiaparelli is usually credited with their first description, although the idea of canals on Mars may have originated 10 years earlier in the work of Angelo Secchi.

Since Schiaparelli's time, the map of Mars has been constantly improved, and hundreds of new details have been added to it. Moreover, photography has played an increasingly important role in this process. In 1958 the International Astronomical Union adopted a formal list of Martian place-names related to a map of Mars drawn by G. de Mottoni and based on a large number of photographs of the planet taken by Henri Camichel at the Pic du Midi Observatory (see Figure 27). The remarkable quality of Mottoni's map was borne out by comparison with pictures later returned from approaching spacecraft.

Until Mariner 9, maps of planets were of large-scale features distinguished by their albedo, or their ability to reflect light. The detailed observations made by the Mariner and Viking orbiters led to an explosion of the number and types of maps, the new maps presenting topography, geologic provinces, temperatures, craters, and so on. Figure 27 is a modern geologic map of Mars based on Mariner 9 data. The responsibility for producing high-quality maps from spacecraft data has been undertaken by the U.S. Geological Survey, and the responsibility for naming the plethora of features found on the surface of the planet has been taken by the International Astronomical Union. The names used in early maps are kept for features that can be identified telescopically, and Latin terms (Planitia, Tholus, Mons, etc.) are used to extend this nomenclature to the different types of geologic features that have been discovered. Craters whose diameters are between five and 100 kilometres in size are named for small towns and villages on Earth. The basic geodesic control net for a system of longitude and latitude on Mars is based on data from Mariner 9 and has the prime meridian defined by a small crater named Airey-O. This tiny crater is the Martian equivalent of Greenwich.

BASIC ASTRONOMICAL DATA

The orbit of Mars. Mars moves around the Sun at a mean distance approximately 1.52 times that of the Earth from the Sun. The relatively large eccentricity (0.0934) of its orbital ellipse brings Mars, at its closest approach, to within 206,600,000 kilometres of the Sun and, at its furthest distance, to 249,200,000 kilometres. Mars completes an orbit in roughly the time in which the Earth completes two, and, therefore, Mars spends most of its year far from the Earth in directions that are near the Sun. At its closest approach, Mars is less than 56,000,000 kilometres from the Earth, but it recedes to almost 400,000,000 kilometres; thus, it is not generally available for detailed telescopic observation.

Table 9: Orbital Elements of Mars*

Mean distance from Sun	227,941,963 km
Eccentricity	0.093387
Inclination	1.86641°
Mean longitude of node	49.4032°
Mean longitude of perihelion	335.6909°
Sidereal period	686.9804 mean solar days
Mean synodic period	779.94 mean solar days
Mean orbital velocity	24.1 km/sec

*For January 1, 1980.
Source: Mars Scientific Model, JPL Document No. 606-1 (March 1, 1972).

Most Earth-based observations of Mars are best made when the planet is at opposition; *i.e.,* when the planet is in the opposite direction in the sky to the Sun. In this position, the planet culminates near local midnight and is also close to the Earth. Successive oppositions occur at an interval of approximately two years and seven weeks; because this interval, called the synodic period, is not commensurate with the orbital period of Mars, an opposition may occur at any point in the Martian orbit. Oppositions near perihelion are called favourable, or perihelic, because the planet is then closest to both the Earth and the Sun and, therefore, is brighter and appears larger than usual. The largest angular size that the planet can attain is 25 seconds of arc. Oppositions that occur near aphelion are the least favourable, for the planet is then farther from the Earth (101,000,000 kilometres) and at its greatest distance from the Sun. During such oppositions, called aphelic, the planet is less bright and smaller than at the average opposition, and it may attain an angular diameter of only 14 seconds of arc.

Inclination of the orbital plane

Another factor that critically influences observations of Mars at opposition is that the orbital plane of Mars is inclined at an angle of 1.85° to the Earth's orbital plane. At Martian perihelion, Mars is well south of the Earth's orbital plane. Consequently, at favourable oppositions Mars is badly placed in the sky for observations in the Northern Hemisphere and is best observed from sites south of the Equator. The situation is reversed for aphelic oppositions.

The north pole of rotation of the planet is at present pointed in the direction of a sixth magnitude star best known by its *Bonner Durchmusterung* catalog number, BD 52°2880, which is located near the bright star Deneb in the constellation of Cygnus. BD 52°2880 is therefore the Martian "pole star" about which the apparent daily motion of the heavens takes place. As is the case of the Earth, the direction of the pole in the sky changes slowly, owing to gravitational torques exerted by the Sun on the planet's equatorial bulge. This precession causes the Martian pole to move along a small circle in the sky with a period of 97,000 Martian years.

The axis of rotation is not perpendicular to the orbital plane of the planet, but is inclined at an angle of 24.935°, and, as for the Earth, the tilt gives rise to the phenomenon of seasons on Mars. The Martian year consists of 668.6 Martian solar days, called sol. For 371 sol, or more than one-half of the Martian year, the Sun appears overhead at some northern latitude and it is spring and summer in the northern hemisphere. The orientation and eccentricity of the orbit, however, leads to an important characteristic of the seasons on Mars: they are quite uneven in length. The summer solstice, which separates spring from summer, occurs considerably later than the midtime between the times of passage through the spring and fall equinoxes. The southern summer is short and hot, while the northern summer is long and cool (see Table 10). These are important factors in understanding the seasonal behaviour of the polar caps (see *The polar caps,* below).

Mars is a small planet. Its equatorial radius is about half that of the Earth, and its mass only one-tenth the terrestrial value. It follows that Mars possesses four times less surface area than the Earth, and its gravity is reduced by a factor of nearly 3. To escape the gravitational pull of Mars, a velocity of only 5.022 kilometres per second

Table 10: Length of Season on Mars

Northern Hemisphere	Southern Hemisphere	Martian days (sol)
Spring	Fall	194
Summer	Winter	178
Fall	Spring	143
Winter	Summer	154

Source: Mars Scientific Model, JPL Document 606-1 (March 1, 1972).

is required. Thus, a rocket need only supply to an object one-fifth of the energy that would be needed on Earth to free it from the gravitational attraction of the planet.

Shape of Mars

Mars is observed to be roughly spheroidal in shape. Early and extensive Earth-based observations indicated that the polar radius was about 40 kilometres less than the equatorial radius. These observations are now known to have systematically overestimated the equatorial radius as a result of a bias imposed on measurements by the Martian atmosphere. Radio occultation measurements and photographic observations from Mariner 9 showed that the actual amount of polar flattening is, in fact, much less, about 21 kilometres, and that the centre of figure is displaced from the centre of mass by about 2.5 kilometres toward the south.

General appearance. To the Earth-based telescopic observer, the Martian surface outside the polar caps is characterized by red-ochre-coloured bright areas on which dark markings are superimposed. In the past, the bright areas were usually referred to as deserts and the majority of large dark areas as maria (sing. mare; *i.e.,* oceans or seas). It is now absolutely certain that these dark areas are not, and in all probability never were, covered by expanses of water. Nor do they appear to be physically similar to the dark maria on the Moon. It therefore seems more reasonable to simply refer to dark and bright areas on the planet than to use the older terms.

The dark areas

Surface features. The most prominent dark areas of Mars occupy a band around the planet between 10° and 40° S. Their distribution is irregular, and the northern hemisphere has only two such major features, called Mare Acidalium and Syrtis Major. Dark markings cover about one-third of the total area of the planet's surface, and—according to the interpretation given to the photographs obtained from the Mariner 9 and Viking orbiters over a period of several Martian years—they originated with the interaction of atmospheric wind systems and surface topography. The dark regions of Mars, when viewed at high resolution, are covered with dark streaks and splotches that are associated with craters, ridges, hills, and other obstructions to the flow of local winds. The streaks are sometimes variable and are caused by the removal, from a darker surface material, of a layer of fine dust by atmospheric turbulence in the lee of obstructions. Some of the dark splotches seen in crater floors have been found to be vast fields of sand dunes.

One of the most remarkable facts about the dark areas discovered by telescopic observation is that they undergo both seasonal and long-term variations in size and colour. In the words of E.C. Slipher: "The markings are in a state of perpetual and bewildering fluctuation, the slower and more gradual ones partly through change in the illumination of the Martian air and partly through inherent change in the markings themselves." Charles F. Capen, an expert visual observer, has reported changes in the Trivium Charontis region. In the spring, the area is medium-dark gray with brown tints, olive drab with a darkening wave, later taking on intense black. In summer, the area stays black, but in fall the colour is dark gray and brown, turning to medium-dark gray with brown tints in the winter.

The reported colours of the dark features have been subject to considerable discussion, particularly because colour recognition is largely subjective. It has been questioned if the colours reported are accurate or are merely a reflection of physiological processes in the eye and mind of the observer. A straightforward answer is not entirely possible, but photoelectric spectrophotometry of the dark areas has shown that they are actually red, although less so than the bright areas of the planet.

The bright areas

The bright areas, which represent about 70 percent of the planet's surface, are also well endowed with subtle shadings and intricate features—the so-called Martian canals and oases. The explanation of these phenomena is now clear. The surface of Mars is largely covered by fine dust, which can be effectively transported by surface winds. The global pattern of this dust on the surface largely controls the pattern of albedo on the surface, and the basic pattern itself is set by reoccurring seasonal wind systems and their interaction with the local terrain.

The blue haze. Another phenomenon reported by many Martian observers from Earth-based photographs has been called the blue haze, or violet layer. A comparison of photographs taken through blue and red filters shows that the contrast of surface features is less in blue light. Slipher found, if the wavelength of the light was less than 4,600 angstroms, dark features on the surface were generally totally obscured. Occasional exceptions, called blue clearings, occur when the obscuration lifts for a short time and the visibility of the dark markings is reestablished. Many theories, none entirely satisfactory, have been devised to explain this phenomenon. Most presume that the blue haze is related to the presence of some scattering or absorbing agent in the Martian atmosphere, although others have considered the effect to be illusory. To investigate this phenomenon, some of the television pictures taken by the Mariner 6 and 7 spacecraft were exposed through a blue filter similar to those used in Earth-based telescopic studies of the blue haze. No evidence for any loss of contrast was found for real topographical features on the surface, but albedo markings did show a reduced contrast in the blue light. The blue haze is therefore a real phenomenon, and the explanation of its cause and variability is probably linked to the interplay of reflectance properties of intermixed bright and dark areas on the surface and of the large amounts of fine dust that are now known to be kept in suspension in the atmosphere.

The wave of darkening. Major seasonal changes occur on the planet and are closely associated with changes in the polar caps. The caps wax and wane, and corresponding periodic changes occur in the appearance of surface features. One of the more enigmatic sequences that take place is called the wave of darkening, first discussed in detail by Lowell in 1894. Near the edge of either polar cap, a general darkening of the surface markings appears in early spring as the cap begins to recede. The darkening then moves away from the receding polar cap and sweeps toward and crosses the equator in an indistinct band of heightened contrast, finally dissipating in the opposite hemisphere. The waves, one in each hemisphere, travel at an apparent speed of about 35 kilometres per day. Studies of the phenomenon were carried out by John H. Focas in 1962, using photometric techniques instead of subjective visual reports. While these waves have been well documented in observations from the Earth, all attempts by 1981 to study them from spacecraft had failed. No surface changes had been detected that could be associated with this phenomenon, and, if it is not illusory, it seems most likely to be some kind of atmospheric effect.

Dust storms. Frequently, and with some regularity, the surface features of Mars are obliterated completely as a result of atmospheric dust storms of planetary dimensions. While these events can apparently happen at any time, they occur more frequently when the planet is near perihelion. Observations from the Viking lander spacecraft, which experienced several such storms, have shown that these storms have a major effect on the structure of the Martian atmosphere. Perhaps the best known storm was that which occurred during the approach of the Mariner 9 spacecraft to the planet. The storm had begun several weeks before the arrival of the spacecraft, and it prevented photography and other measurements of the surface for the first few weeks of the mission.

STRUCTURE AND EVOLUTION OF THE MARTIAN INTERIOR

Knowledge of the interior of Mars and how it has evolved is of the broadest conceptual kind. Observations that bound these concepts include the relatively precise measurements

Density of Mars

by the Mariner and Viking missions of the planet's mean radius, its mass, and its moment of inertia. These known factors, together with hypotheses on the overall chemical composition of Mars, restrict the range of interior models to a surprising degree and show that the internal structure of the planet must be substantially different from that of the Earth or the Moon. The mean density of Mars, when corrected for internal pressure, is 3.75 g/cm^3 and is much lower than the corresponding value of 4.04 g/cm^3 of the Earth. Evidently the overall chemical composition of Mars is different from that of the Earth. Most likely Mars has a lower iron–nickel content or is enriched in lighter, more volatile elements.

Table 11: Physical Parameters of Mars

Equatorial radius	3,396.9 km
Polar radius	3,376.1 km
Geometric flattening	0.00612
Dynamical flattening	0.00525
Surface area	1.44 × 10^8 km^2
Volume	1.63 × 10^{11} km^3
Mass	6,418 × 10^{26} gm
Mean surface gravity	372 cm/sec^2
Mean density	3.933 g/cm^3
Mean escape velocity	5.022 km/sec
Martian (sidereal) day	24 hr 37 min 22.663 sec
Mean Martian solar day	24 hr 39 min 36 sec
Direction of North Pole (1980)	
Right ascension	21 hr 9.2 min
Declination	+52° 41.6′
Pole star	BD 52° 2880
Inclination of equator to orbit	24.936°
Visual albedo	0.159
Visual magnitude at mean opposition	−2.01
Magnetic moment	<2 × 10^{21} gauss cm^2
Central pressure	350–400 kbar

Sources: Various, *see* bibliography.

The geometric and gravitational figure, or shape, of Mars places constraints on the internal distribution of density, its more spheroidal shape indicating a less centrally condensed object. Evidence of gravity anomalies, a range in topographic altitudes that is twice that found on the Earth, photogeological evidence of extensional stresses, and the fact that the mean geometric figure of the planet does not coincide with that of the gravitational field raises some doubt as to the veracity of the theory that is used to interpret observations of the planet's shape. It is thought, however, that any deviations that are present will have only a minor effect on the implications for internal structure.

Many interior models have been constructed based on the above facts. While they differ in many important details and assumptions, the general picture that emerges is of a planet that is differentiated into three well-defined zones—a dense core, a mantle, and a thin lithospheric veneer, or crust. The radius of the core is about 1,700 kilometres, with a density in the range of 5 to 8 g/cm^3 (depending on assumed composition), and the mantle extends to perhaps within 200 kilometres of the surface. The best estimates of the density of the mantle are in the range of 3.33 to 3.58 g/cm^3, substantially higher than mantle densities estimated for the Earth.

Experiments were conducted by the Viking landers and Mars spacecraft to examine the chemical composition of the crustal layers on Mars and, by Viking, to probe the interior seismically. Results of these experiments obtained by 1981, while of considerable interest regarding detailed phenomena, were not useful as a probe of the interior. The X-ray fluorescent spectrometer on Viking examined surface materials at both lander sites and found them to be remarkably similar, even though they were separated by a distance of 6,500 kilometres. The chemical nature of the Martian soil is unlike any rock type found on the Earth or known on the Moon and seems to be similar to a heavily weathered, hydrated clay. The Mars spacecraft found direct evidence for the occurrence of natural radionuclei of potassium, uranium, and thorium in the soil in similar proportions to those found in terrestrial soils; the energy released by these long-lived radioactive elements has probably determined the thermal evolution of Mars over most of geological time. Only one of the Viking seismometers was able to operate on the Martian surface, and so it was not possible to probe the structure of the deep interior by listening to the reverberations of Mars quakes from two stations, as has been done on the Moon. The Martian core is probably solid, since the planet appears to have no strong intrinsic magnetic field of global dimensions that is of the dynamo type.

The small amount of factual data regarding the evolu-

By courtesy of the Jet Propulsion Laboratory/National Aeronautics and Space Administration

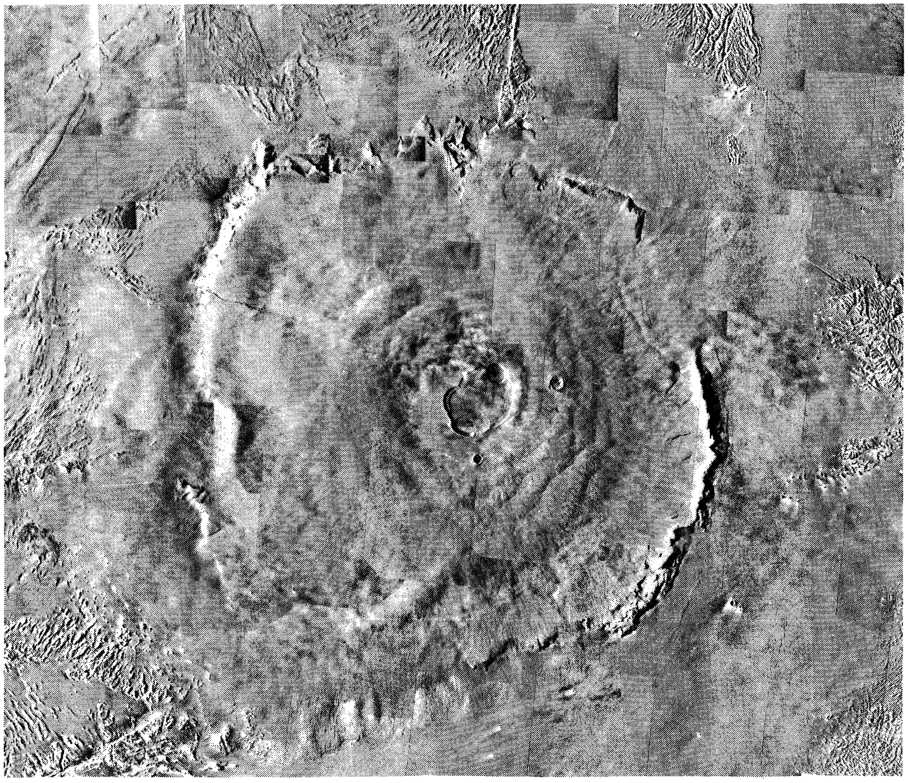


Figure 28: Viking photographs form a mosaic image of Olympus Mons looking directly into the top of the caldera. The circular escarpment, which borders the mountain, is at places as high as seven kilometres.

Evolution of the interior

tion of the Martian interior is primarily comprised of the geologic record of large-scale extensional stresses in the crust, the presence of volcanism (lava plains, volcanic domes), and the densities of impact craters (see below). It is thought that there were four identifiable stages in the evolution of the planet. The first was its accretion over a relatively short time scale. This was quickly followed by the second stage, the separation of the core and the formation of a veneer of crust on the otherwise hot planet. These first two stages are thought to have been completed in the first 1,000,000,000 years of the planet's existence. In the third stage, which occupied the next 2,000,000,000 years, strong heating throughout the body of the planet by long-lived radioactive elements in the interior governed the development of the basic physical and chemical structure of the mantle. This was accompanied by a considerable degassing of the interior and formation of early atmosphere, volcanism, and a general physical expansion of the planet. It was during this stage that the large-scale topography, volcanoes, and enormous chasms and channels found by Mariner 9 were formed. This activity probably stopped some 1,500,000,000 years ago, as the planet reached an equilibrium referred to as its mature state. The fourth stage of evolution is one of cooling and contraction as the interior energy sources become exhausted, and it is not known whether Mars is yet in this terminal stage.

CHARACTER OF THE SURFACE

The character of the Martian surface on linear scales below 100 kilometres was a matter of pure conjecture until the pictures of Mariners 4, 6, and 7 were returned to Earth in 1965 and 1969. But Mars did not reveal the true character of its surface until the Mariner 9 orbiter mission, which photographically surveyed more than 70 percent of the total surface of the planet at spatial resolutions between one and three kilometres. The Viking missions in 1976 then provided much of the detailed information that explains the forces and processes that have shaped the surface.

The early fly-by missions of the 1960s seemed to indicate a planet much like the Moon, a planet whose interior had been essentially inactive throughout its history and whose inhospitable surface was saturated by impact craters. Only glimpses of the presence of a "chaotic" terrain, characterized by irregular areas of jumbled short ridges and depressions, seemed to indicate exceptional surface modification processes. The reaction from many scientists at the time was that Mars was simply uninteresting.

Olympus Mons

Mariner 9 showed that this impression was quite inaccurate and simply the result of bad luck. In fact, a surface of widely varying character was revealed. Any possible relationship between the classical "canals" of Lowell and real surface features was immediately dispelled as an optical illusion resulting from the poor spatial resolution achievable by telescopes on Earth. Some of the classical "oases," such as Juventae Fons and Oxia Palus, were identified as large, dark-floored craters. Nix Olympica, discovered by Schiaparelli in 1879, was found to be a truly enormous and complex caldera at the summit of the largest volcanic mountain presently known in the solar system. This volcano, now called Olympus Mons, has 10 times the volume of the Earth's largest volcano, Mauna Kea, and, at an elevation of 27 kilometres, is three times the height of Mt. Everest. Olympus Mons covers a circular area some 550 kilometres in diameter and is skirted by an escarpment that in places rises as much as seven kilometres above the adjacent terrain.

The Mariner and Viking pictures show the surface to be composed of an almost innumerable variety of land forms, the origin of many of which was not yet understood in the mid-1980s. Many of the thousands of images and other scientific data taken by Viking still had not been studied in detail. Nevertheless, there are clear indications of planet-wide order in the location and variety of the primary terrain types.

Two regions of relief. Except for the polar regions, the character of the surface terrain divides the planet into two hemispheres. In the south, the surface is mostly one or two kilometres above the mean level and is heavily cratered. At low resolution, photographs of the surface are reminiscent of those of the lunar uplands or the planet Mercury, and the density of impact craters indicates a very old and primitive terrain. In the north, which is generally below the mean level, there are many craters, but the surface is chiefly characterized by widespread, relatively smooth plains. These plains were probably formed by widespread flooding of the original heavily cratered surface by molten lava from the interior in the third stage of the planet's development. This type of hemispheric dichotomy is not unique in the solar system and is present on the Moon and the Earth. The physical cause of the dichotomy may, however, be different in all three cases.

The boundary of the two regions is not parallel to the equator, but roughly follows a great circle inclined to it by about 50°. The boundary is broad and irregular and slopes downward toward the north. Along this interface, the terrain is both varied and at places unique. Here are found "fretted" terrain, "chaotic" terrain, large "channels," "knobby" terrain, and, in the region of Tharsis, a large area of relatively young volcanic and "grooved" terrain that is surmounted by four of the largest volcanic mountains in the solar system.

Types of craters

The southern craters. The southern cratered terrain appears to be extremely old. The number of very large craters implies an uncertain age of about 3,000,000,000–4,000,000,000 years, depending on the precise history of impact cratering events. There are several distinctive types of craters: large and flat-bottomed craters; smaller and fresh-looking bowl-shaped craters, as on the Moon; pedestal craters; and rampart craters. The last two types are unique to Mars, with rampart craters distinguished by smooth layers of ejecta that appear to have splattered from the crater, possibly because of the presence of a deep permafrost layer just below the surface and of groundwater below the permafrost. Pedestal craters are otherwise normal-looking craters that sit on elevated circular platforms. It is thought that these platforms may be debris from the original impact that was left intact as the rest of the surface was stripped away in severe erosional episodes. The largest recognizable impact feature on Mars is the Hellas basin, which is roughly 1,600 kilometres in diameter. Its floor, once erroneously thought to have a special featureless character, is the lowest topographic region on the planet, lying some four kilometres below the mean surface level.

The high spatial resolution of the Viking photography has yielded one further tantalizing feature characteristic of this ancient southern terrain: the pervasive presence of fine channels and gullies that strongly resemble networks of desert runoff channels. It is speculated that, because of the wide occurrence of these features and their similarity to runoff systems in terrestrial deserts, they were formed as a result of widespread rainfall in a primitive epoch in the history of the planet, when its atmosphere was much more massive than it is today.

Possible fluvial channels

The interface terrain. In fretted, chaotic, and knobby regions, the Martian surface has apparently undergone vigorous modification processes leading to the collapse of large areas, sometimes followed by erosional processes (*e.g.,* landslides, wind, outflow of subsurface water) and volcanic flooding with fluid lavas. Collapse in the case of chaotic terrain is thought, in one hypothesis, to be the result of the melting of a subsurface permafrost layer by volcanic activity. If the water was released rapidly and was in great enough quantity, it may have flowed away from the site, leaving chaotic terrain at the head of a fluvial channel. Such channels, some very large, are seen in the region near Chryse Planitia (the Viking 1 spacecraft landed at the end of such a channel). The channels often show sinuous scour lines that diverge around obstacles that have themselves been eroded into streamlined shapes. The hypothesis of fluvial (water) origin is, however, not entirely certain, since liquid water is not stable under the present conditions of the Martian surface, and the probability of a former climate more favourable for such processes is uncertain. Several alternative, but less satisfactory, hypotheses for the formation of the channels include the flow of highly fluid lava or of fluids other than water, or wind erosion.

The Tharsis region. The youngest surface on Mars is found in the region of Tharsis. The surface is a vast,

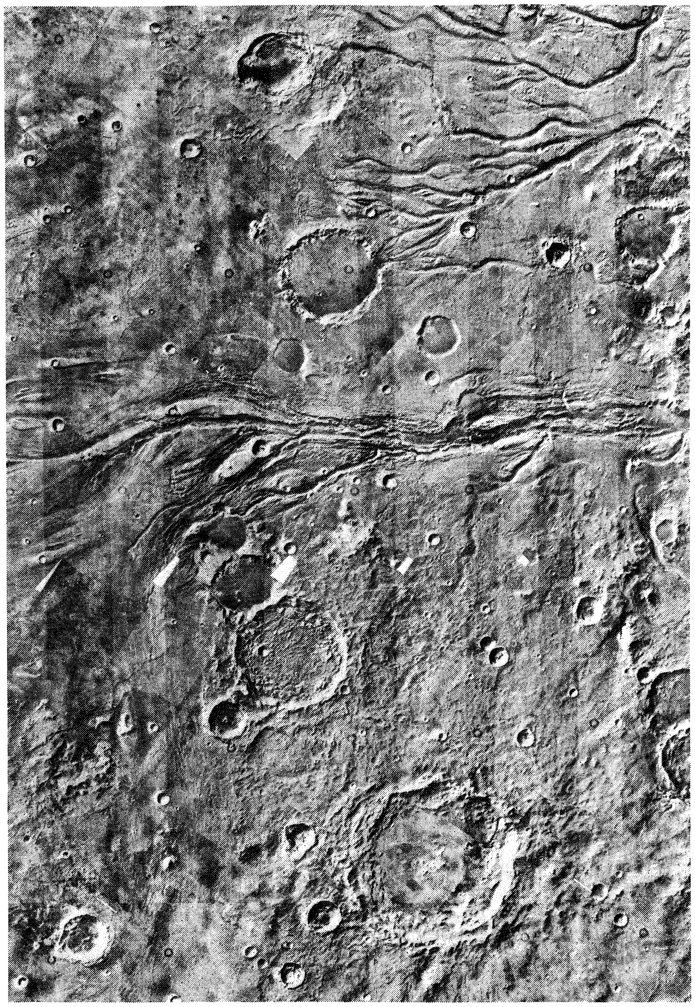


Figure 29: Viking photograph of the cratered terrain in Lunae Planum. The dry channels are thought to have been caused by a catastrophic flood during the early history of Mars.

By courtesy of the Jet Propulsion Laboratory/National Aeronautics and Space Administration

domed, volcanic plain some 2,000 kilometres in extent. This Tharsis "bulge" is raised in excess of 10 kilometres above the mean surface level. Superimposed on it is a ridge formed of three massive volcanoes of the shield type: Arsia Mons, Pavonis Mons, and Ascraeus Mons. On the flanks of the bulge are Olympus Mons and the remnants of a much older volcanic mountain, Alba Patera. It is not possible to estimate accurately the age of these volcanoes, but, with the exception of Alba Patera, they could be less than 1,000,000,000 years old. The possibility of their being active today cannot be completely discounted. Martian volcanic features, in smaller and in different forms, are almost ubiquitous in the northern plains. Many other volcanic mountains have now been located on the planet, with concentrations in Elysium and in both polar regions.

Tharsis area's volcanoes

The flanks of the Tharsis bulge are also the site of a "grooved" terrain that is formed of largely parallel fractures radial to the centre of the bulge. Some of these fractures are very large and complex; the largest, Valles Marineris, is a canyon system that runs roughly east–west near longitude 70° west and some 10° south of the equator. Its length of 4,500 kilometres stretches almost a quarter of the way around the planet, and, with a maximum width of roughly 600 kilometres and depth of seven kilometres, it is comparable in dimensions to the East African Rift Valley. The walls of its canyons are very steep and show ample evidence of large landslides. The walls have receded in many places to form a complex pattern of tributary valleys in the surrounding plains. Flat-topped mesas found in the bottom of the canyon have been interpreted by some scientists as evidence for past flooding and sedimentation in the floor of the valley.

Topography. Early visual observers of the planet concluded that high mountain ranges similar to the Himalayas or the Alps were not present on Mars. Early estimates suggested that, if mountains existed with altitudes higher than 8.4 kilometres, they would have been detected as irregularities at the terminator. The calculations evidently did not apply to the gently sloping volcanic shields that have now been found on the planet and that rise to such enormous altitudes. The evidence against folded mountain building from the Mariner and Viking photographs also seems clear, for there are no signs of the compressional folding that arises as a result of plate tectonics. There is, however, plenty of evidence for topographic relief on the surface of the planet on the continental scale. Elevation differences of as much as 30 kilometres are known to exist over scales of thousands of kilometres. These differences in elevation, while large, nevertheless imply gentle slopes.

History of topographic studies of Mars

The history of the search for large-scale topographic relief is fascinating and has an ironic twist. In a sense, the search for large elevation differences grew out of the controversy over the nature of the dark areas and the possible presence of life. The proponents of the "vegetation" theory of the dark areas considered them to be low lying and hence warmer and moist, and therefore more conducive to life. The temperatures of the dark areas are indeed a few degrees warmer than those of the surrounding bright areas, but the opponents of the above hypothesis preferred to attribute the temperatures of the dark areas and their curious ability to change their appearance on a seasonal basis to inorganic causes. The fact that they are warmer is simply a result of their being darker, which is just another way of saying that they absorb more energy from the Sun. Their apparent ability to change their appearance was presumed to be the result of their being elevated regions and of the transport and deposition of small dust particles by the prevailing winds.

In 1965 it was noticed that white clouds that tended to form and remain stationary over certain bright areas were aligned with the boundaries of adjacent dark areas. These clouds were thought of as lee clouds formed in the prevailing wind as it crossed the dark areas and that the latter must therefore be high ridges. Further circumstantial evidence came from radar observations. Surface regions with high radar reflectivity were found to be correlated with dark areas, although they were not actually coincident with the dark areas, but were always somewhat displaced in longitude. The conclusion reached from these observations was that the dark areas must be elevated regions with low slopes that extend over extremely large distances. Elevation differences of about 10 kilometres were thus implied. In 1967, precise radar-ranging methods were used to scan the surface along a narrow swath centred on latitude 22° N. The timing of the radar echoes led to the first direct evidence that large elevation differences in fact did exist, but it showed, ironically, there was little correlation between the dark areas and elevation. Another investigation using entirely different techniques further confirmed and extended these measurements. The amount of carbon dioxide gas in the atmosphere was measured at hundreds of points of the disk from latitudes 25° S to 40° N and provided the data to make the first crude topographical maps of the planet.

There are now detailed maps of the topography and gravity field of Mars that cover the entire globe. These maps were made by combining information from various experiments, including Earth-based radar, timings of occultations of orbiting spacecraft behind the planet's limb, and ultraviolet and infrared photometry. The main topographic characteristics of Mars are as follows. In the south, except in the Hellas basin, the surface is elevated above the mean by about two kilometres. The reverse is the case in the northern hemisphere. In equatorial latitudes, the circumference of the planet shows a sinusoidal distortion, with high altitude bulges near 110° W longitude (Tharsis) and 210° W (Elysium). The Tharsis bulge rises to an altitude of 10 kilometres, while in Elysium the altitude reaches six kilometres. The frequency of occurrence of a particular surface elevation shows only a single peak near the mean level. This is unlike the Earth, which shows a bimodal distribution as a result of its surface being di-

Main topographic features

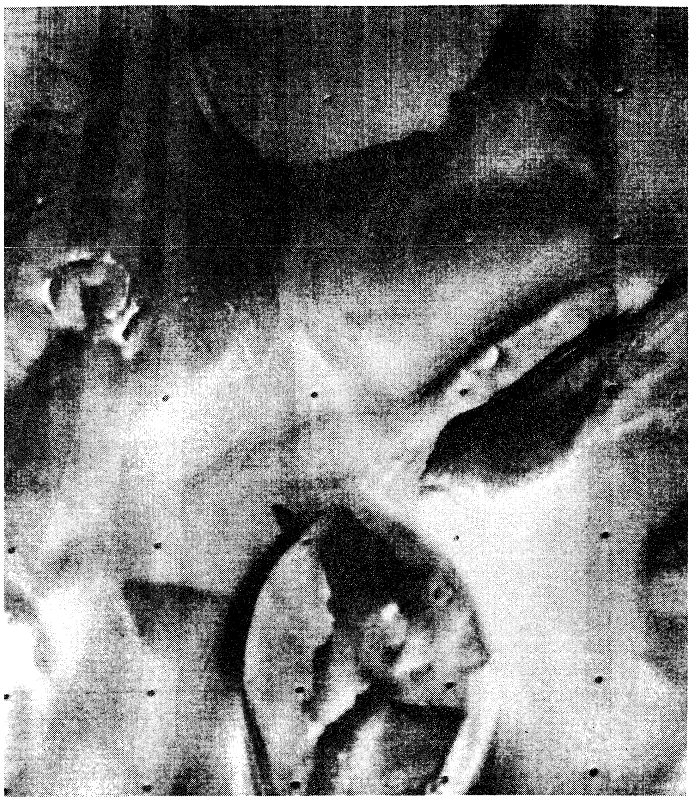


Figure 30: Mariner 9 photograph of the complex pattern of delicate swirls and irregular dark blotches that cloak the terrain of the south polar region of Mars.

By courtesy of the National Aeronautics and Space Administration

vided into ocean floors and continental regions. On Mars, gravity anomalies are strongly correlated with topographic features, indicating that the Martian crust is probably not isostatically compensated. The topography of Mars may still be relaxing from its primitive form, may possibly be supported by internal mantle convection, or is simply supported by uncompensated stresses in the crust.

THE POLAR CAPS

Seasonal changes. The seasonal behaviour of the polar caps is one of the more elaborately documented phenomena on Mars. In early spring the southern polar cap has its greatest dimensions and stretches from the pole to a latitude of about 50° S. As spring proceeds, the cap sublimates and begins to recede at a rate of about 1° of latitude every five days. During this time the edge of the cap takes on a ragged appearance. Eventually, gaps appear that break the cap into well-defined fragments. This process is repeated each year, and the location of the fragments is essentially always the same; the Mariner and Viking cameras, however, have detected small variations in detail from year to year. As the polar snow recedes, a dark blue band, first noted by Wilhelm Beer and Johann von Mädler in 1830, is seen in telescopic observations at the edge of the cap. One-third of the Martian year later, the cap has receded to its smallest extent; it is then not usually visible from Earth, but spacecraft observations show that a small remnant remains.

The polar hood

A brief interlude of atmospheric clarity then follows, which terminates in early fall with the rapid formation of obscuring clouds referred to as the polar hood. Occasionally, the hood is sufficiently transparent in the red region of the spectrum to allow photographs to be taken of the new southern polar cap then forming beneath it. The hood is far more extensive than the cap itself and has been reported to extend to within 35° of the equator. The advance of the polar hood is associated with an atmospheric frontal system, the passage of which is very abrupt. The composition of the particles that make up the polar hood is either water ice, solid carbon dioxide, or, possibly, a mixture of the two.

The behaviour of the northern polar cap is similar, although not exactly the same in detail. The differences are mainly due to a combination of the effects of the lengths of the seasons and the variation in distance from the Sun. The atmospheric frontal system associated with the advance of the polar hood was observed in the northern hemisphere by the Viking 2 lander. Since the front was observed near midday, it was found to be accompanied by a substantial drop in illumination. The northern summer is longer than the southern. The planet is at a greater distance from the Sun during the northern summer, however, and receives less heat during that period, allowing the remnant of the northern cap to be more substantial than its southern counterpart.

Theories of cap composition

Composition. The composition of the remnant Martian polar caps in summer is not precisely known, despite the considerable data on their thermal and radiative properties acquired by the Viking craft. One early hypothesis—that the caps were simply water ice—can be traced to William Herschel, who imagined the polar caps of Mars to be just like those on Earth. In 1898, George Johnstone Stoney questioned this theory and suggested that the caps might consist of frozen carbon dioxide (CO_2); physical evidence to support this idea was not available until 1947, when Gerard P. Kuiper discovered CO_2 to be present, although in unknown proportions, in the planet's atmosphere. Stoney's hypothesis was particularly attractive because no water vapour had then been detected in the atmosphere, but Kuiper, in 1948, was nevertheless persuaded that the caps had reflection spectra that were characteristic of water ice, rather than of CO_2. Similar spectra, published by V.I. Moroz in 1966, indicated the same result, but R.B. Leighton and B.C. Murray published, also in 1966, an important numerical study of the thermal environment on Mars that threw considerable doubt on the water ice hypothesis. Their calculations, although based on simplifications of the actual conditions on Mars, were later supported by measurements taken from lander and orbiting spacecraft. The calculations predicted that atmospheric carbon dioxide gas would freeze out at the poles and showed that polar caps so formed would behave similarly to actual observations.

The question of the composition of the cap remnants remained unsolved, however, until the Viking 2 mission returned information regarding the north polar remnant. The ice of the northern remnant is at temperatures of 205 K, well above the frost-point of CO_2, and is unquestionably mainly water ice. It represents the largest reservoir of available water on the planet. Observations at the south pole are more ambiguous, and it is highly probable that the southern cap remnant is CO_2 ice, rather than water.

Data from the Mariner 6 and 7 spacecraft in 1969 resolved the problem of composition as far as the fully devel-

By courtesy of the Jet Propulsion Laboratory/National Aeronautics and Space Administration

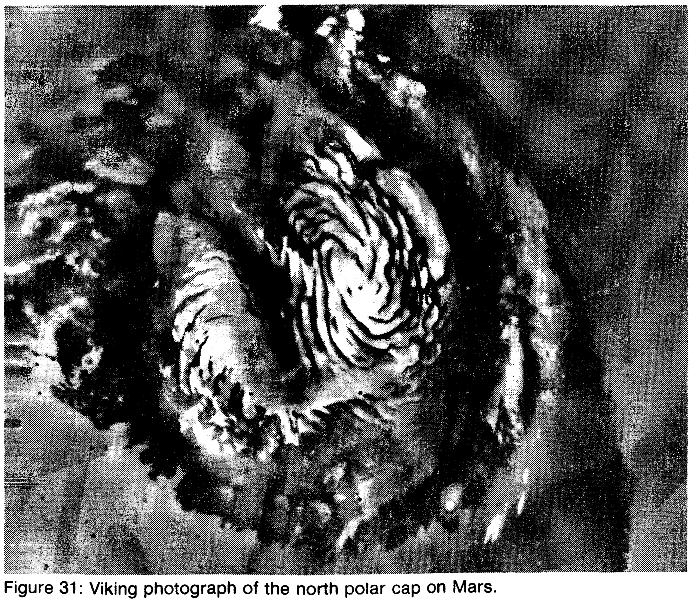


Figure 31: Viking photograph of the north polar cap on Mars. The bright areas are composed of water ice. The dark lines cutting into the cap are valleys, the sides of which are the site of a laminated terrain unique to Mars.

oped caps are concerned. On board the spacecraft were an infrared radiometer that could measure the temperature of the surface and an infrared spectrometer that could look for characteristic features in the spectrum of the reflected light from the polar cap without the interference of absorption in the Earth's atmosphere that plagues Earth-based astronomers. As the field of view of the radiometer swept over the polar cap, the temperature dropped to the value predicted for frozen carbon dioxide, thereby demonstrating that a polar cap composed of CO_2 would be stable. If the measured temperature had been only 10° higher, the CO_2-ice hypothesis would have been untenable. The infrared spectrometer then provided conclusive evidence that the spectrum contained absorption features at 3 and 3.3 micrometres, levels that are characteristic only of solid CO_2.

Relief features. The terrain in the polar regions is among the most distinctive on the planet. Except for winter, when the surface is covered with what are probably several metres of CO_2 snow, wide areas of the surface show characteristic layers, or "laminations," of alternately dark and bright materials. The layers are exposed on the flanks of escarpments and valleys that border or cut into the polar caps. (The valleys have cut into the caps in such a way as to give them a distinctive swirl-like appearance.) The layers may indicate a history of regular climatic variations, as the planet's orbital parameters have varied in response to gravitational interaction with the other planets (primarily Jupiter and Earth), which controls the relative amounts of dust and water ice deposited during successive winter seasons.

Sand dunes

In the north polar region is the largest sea of sand dunes, or "erg," in the solar system. This erg forms a band of windblown sand entirely around the north polar remnant cap. The dunes in this region are spectacular in their regularity over hundreds of kilometres.

THE ATMOSPHERE

Basic atmospheric data. The Martian lower atmosphere consists mainly of carbon dioxide that has been released slowly from the crust and interior over the age of the planet. There is strong evidence that the lower atmosphere was much more extensive in the remote past, and there are signs that rainfall may have occurred in some regions. While only small amounts of water are found in the lower atmosphere today, it supports just about as much water vapour as it is physically able. In this sense, the lower atmosphere is humid, or saturated; it is not in any way to be considered moist, however. Ice clouds have been observed at high altitudes, and in several localities morning ground ice-fogs have been seen. Nitrogen and various rare gases, oxygen, and carbon monoxide are also present in small quantities.

By courtesy of the National Aeronautics and Space Administration

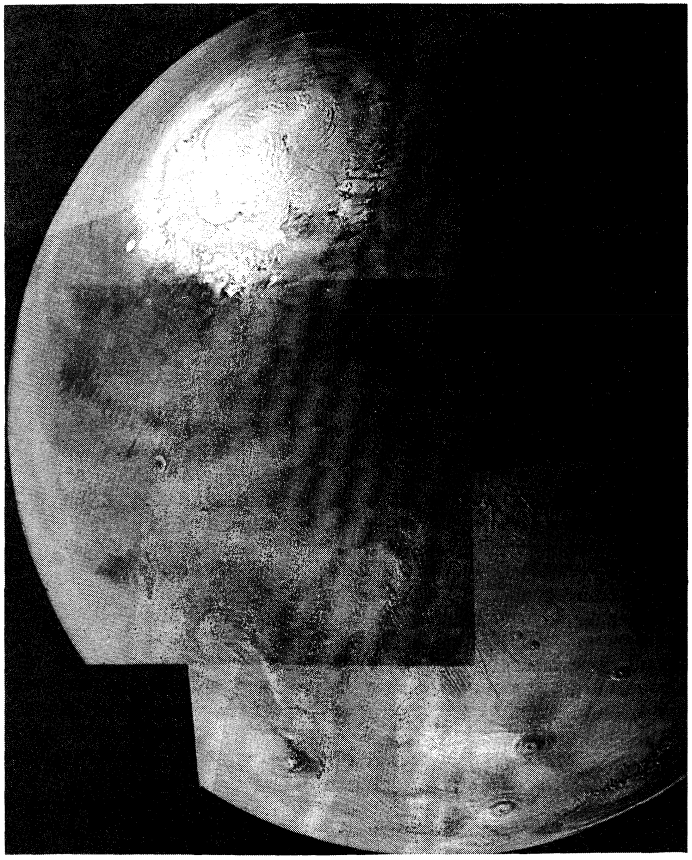


Figure 32: Mariner 9 composite photo of the northern hemisphere of Mars taken during the late Martian spring. The shrinking north polar cap can be seen at top, and Olympus Mons can be seen at lower left.

Table 12: Composition of the Martian Atmosphere
(fraction by weight)

Carbon Dioxide (CO_2)	96.5%
Molecular nitrogen (N_2)	1.8%
Argon (Ar)	1.5%
Molecular oxygen (O_2)	0.1%
Carbon monoxide (CO)*	0.05%
Water vapour (H_2O)*	0.02%
Neon (Ne)	0.0001%
Krypton (Kr)	0.00003%
Xenon (Xe)	0.00002%

*Uncertain, variable, or unevenly distributed in the atmosphere.

The characteristic temperature in the lower atmosphere is about 200 K. This is generally colder than the average surface temperature, which in daytime is about 250 K. The difference is due to the presence of temperature gradients that have an average value of about 1.5 K per vertical kilometre. In the middle and upper atmosphere, large daily waves of temperature and density have been found, indicating a confused and variable structure. There is also an ionosphere that, unlike the Earth's ionosphere, consists primarily of oxygen ions at an altitude near 130 kilometres.

A thin atmosphere

At the planet's surface, the atmosphere is very thin, the average surface pressure being less than one-hundredth that on the Earth. Because of large variations in the altitude of the surface, however, some locations may experience a surface atmospheric pressure up to five times greater than others. The total mass of atmosphere, unlike that of the Earth, undergoes major seasonal variations as the major atmospheric component, CO_2, "snows out" at the winter pole. The annual variation of surface pressure is about 26 percent of the mean value.

Despite its thinness, the Martian atmosphere supports a variety of clouds. One variety, the so-called yellow clouds, are intense dust storms that throw dust into the atmosphere up to heights of 10 kilometres or more. Under special conditions, localized storms have been found to trigger storms of global dimensions that can totally obscure the surface from view; such global storms were experienced by Mariner 9 and the Viking lander craft. Even in their absence, the atmosphere is never clear of dust, the presence of which plays an important role in determining the structure of the atmosphere.

Composition and surface pressure. Precise measurement of the Martian atmosphere was made possible by the Mariner 9 and Viking 1 and 2 missions (see Table 12). Prior telescopic observations had already determined that CO_2 was the atmosphere's primary constituent, that traces of carbon monoxide (CO) and water were definitely present, and that the mean surface pressure was near 5.5 millibars. Kuiper discovered the presence of CO_2 in 1947, at a time when the meager evidence available suggested that, as on Earth, nitrogen was the principal atmospheric component. It was not until 1963 that it was possible to conclude that CO_2 was, in fact, the major atmospheric component.

As a result of more than 500 radio occultation experiments from orbiting spacecraft and direct sampling and chemical analysis of atmospheric gases through mass spectroscopy and gas chromatography, precise knowledge of the atmosphere, both locally and globally, and of its variation in time is now available.

Molecular composition

Below 125 kilometres in altitude, where the atmosphere is well mixed by turbulence, 96.5 percent of the atmosphere by weight is CO_2. This is a comparatively large amount of CO_2, being some nine times the amount now in the Earth's much more massive atmosphere. When consideration is made, however, of the tremendous amount of CO_2 that has been cycled through the Earth's atmosphere and is now chemically locked in terrestrial sedimentary rocks, the Martian CO_2 content actually represents slightly less than 1/1000 of that on the Earth. The balance of the Martian atmospheric content is molecular nitrogen (N_2), argon (Ar), water vapour (H_2O), and rare gases. In addition to these molecules, there are also trace amounts of gases that have been produced from the primary gases by photochemical reactions, generally high up in the atmosphere; this component includes molecular oxygen (O_2), carbon monoxide (CO), nitric oxide (NO), and small amounts of ozone (O_3). The atmosphere supports an ionospheric layer at the top of the atmosphere, where the densities are low enough for the components to separate diffusively according to their mass. The Martian ionosphere is less developed than the F-layer on Earth, but is formed in a similar fashion. Its peak density of 10^5 ions/cm^3 occurs near 130 kilometres altitude and is formed primarily by O_2^+ molecules that have been produced in local chemical reactions. These photochemical reactions are often quite energetic; Michael B. McElroy has shown that the electronic recombination of O_2^+, N_2^+, CO^+, and CO_2^+ ions high in the atmosphere (in the exosphere) can produce energetic neutral nitrogen and carbon atoms that can escape ballistically from the Martian gravitational field. His calculations show that Mars has lost significant amounts of oxygen and nitrogen in this way over geologic time. Finally, an important constituent of the outer limits of the atmosphere is atomic hydrogen (H). This is a relatively easy atom to detect and monitor, for it strongly scatters solar radiation, which is present in abundance at 1,216 angstroms. The rate of decrease in the brightness of the scattered radiation with increasing distance from the planet allows for the determination of the temperature at the top of the atmosphere. This temperature, in turn, is important for understanding the long-term evolution of the atmosphere.

The total mass of gas in the Martian atmosphere is not fixed, and measurements at the Viking lander sites show that there is a regular seasonal variation as CO_2 "snows out" on the winter polar cap and is released from the receding summer polar cap. Seymour Hess and his associates found that this variation has a peak-to-peak amplitude of about 26 percent of the mean atmospheric pressure (7–8 millibars) at the lander sites. Hess has calculated that some 7.9×10^{12} metric tons of CO_2 are involved in this yearly exchange, equivalent to a thickness of at least 23 centimetres of dry ice over the vast area of the polar caps. This translates into several metres of CO_2 snowfall. The Martian atmosphere seems unique in this respect, and several atmospheric phenomena—such as the frequency of dust storms and wind patterns and velocities—are ultimately controlled by it.

Presence of water

Water (H_2O) is a minor constituent of the atmosphere, primarily because of low atmospheric and surface temperatures. At most, H_2O is present at a level of only a few molecules per 10,000, but it seems nevertheless to play an important role in atmospheric chemistry and in meteorology. The atmosphere is effectively saturated at all times with water vapour, yet there is no liquid water. The temperature and pressure of Mars are so low that water molecules can exist only as ice (solid) or as vapour. Although the presence of water vapour on Mars is exceedingly difficult to determine by telescopic observations because of the large amounts of water in the Earth's atmosphere, the discovery of water on Mars was made in this way in 1963 by Lewis D. Kaplan, Guido Munch, and Hyron Spinrad.

Most of the detailed knowledge of the water on Mars comes from the Viking orbiter, which was able to follow the seasonal patterns of water in the atmosphere over a full Martian year. Unexpectedly, atmospheric water vapour—except in a few special localities that show surface ice or local morning fogs—undergoes little daily exchange with the surface, despite the very cold surface temperatures recorded during the Martian night. Water vapour is found to be mixed uniformly up to altitudes of 10 or 15 kilometres and shows strong latitudinal gradients that depend on the season. According to Crofton B. Farmer, the total atmospheric content of water vapour can be most easily understood by imagining a cube of ice some 1.3 kilometres on each side, a size little more than that of a medium-sized terrestrial iceberg. It is inferred that this atmospheric water vapour is in contact with a much larger reservoir in the Martian soil; subsurface layers of ice are thought to be ubiquitous on Mars at latitudes poleward of 40°.

Surface pressure is highly variable because of the meteorological processes mentioned above and the enormous range of topographic elevations on the surface. The pressure at the lowest elevations (for example, at Hellas; −4 kilometres) is as high as 8.4 millibars, while at the crest of Olympus Mons (+27 kilometres) it is only 0.5 millibar. At locations near the mean elevation of the topography, the average pressure is about 5.9 millibars.

Isotopic composition

In addition to knowledge of molecular constituents of the atmosphere, the Viking mass spectroscopy and gas chromatography experiments yielded measurements of the isotopic composition of the major species and rare gases. Values of the ratios of isotopes, carbon-12 to carbon-13 and oxygen-16 to oxygen-18, are similar to those on Earth. The ratio of the isotopes argon-40 to argon-36—a ratio thought to be important in judging the amount of volatile gases that have escaped from the interior of the planet over geologic time—was found to be roughly 10 times the value for the Earth, however. In addition, the ratio of the isotopes nitrogen-14 to nitrogen-15 was found to be 1.62 times larger than the terrestrial value. These latter two measurements, together with many local indications of liquid water on the surface in the remote past as well as similar measurements of the Venutian atmosphere, have had a significant effect upon ideas concerning the evolution of the Martian atmosphere. The enhanced nitrogen ratio is thought to have originated in the selective escape of the lighter nitrogen isotope (nitrogen-14) as a result of photochemical reactions at the top of the atmosphere noted earlier. This idea, if correct, implies that much larger amounts of nitrogen were present in the atmosphere in the past. The similarity of carbon and oxygen isotopic ratios to those of the Earth implies that large volatile reservoirs for these elements exist on Mars. The enhanced argon isotopic ratio—coupled with the fact that the observed relative elemental abundance of the rare gases (argon, xenon, krypton, neon) is similar to terrestrial values—can be interpreted as a sign that Mars was formed with an overall deficiency of volatile elements relative to the Earth. Also, the low total mass of atmosphere may indicate that there has been a much slower release of gases from the interior of Mars than that experienced by the Earth.

The ratio of carbon dioxide and molecular nitrogen content relative to that of the rare gases is 10 times smaller on Mars than on the Earth. This is interpreted by some scientists as a sign that large amounts of CO_2 and N_2 have been lost to Mars during its history. This hypothesis agrees with the ideas derived from the discussion of atmospheric escape, and it is conjectured on the basis of these ideas that Mars may once have had an atmosphere the major constituents of which were carbon dioxide, molecular nitrogen, and water and the dimensions of which were similar to the atmosphere that exists on the Earth today.

Atmospheric structure. The vertical structure of the Martian atmosphere—*i.e.,* the relation of temperature and pressure to altitude—is determined partly by a complicated balance of radiative, convective, and advective energy transport and partly by the way energy (from the Sun) is at first introduced into the atmosphere and then lost by radiation to space. Before 1965, essentially all knowledge of the Martian atmospheric structure was based on extrapolations of indirect data with the help of theoretical models. As a result of a wide range of spacecraft observations, detailed knowledge of the real structure of the atmosphere and how it reacts to changing conditions is now available. Many of the more general concepts, such as the predictions of large diurnal fluctuations in temperature and of large thermal inversions at the surface during the Martian night, were correct. Detailed comparisons have also shown, however, that many of the early ideas were naive.

There are two factors that control the vertical structure of the lower atmosphere—its composition of almost pure carbon dioxide and its content of large quantities of suspended dust. Carbon dioxide, which has a strong infrared spectrum, radiates energy with great efficiency at Martian temperatures and causes the atmosphere to respond rapidly to changes in the amount of solar radiation received. The suspended dust absorbs great quantities of heat directly from the Sun's light and provides a distributed energy source throughout the lower atmosphere.

Temperature variations

Surface temperatures depend on latitude and fluctuate over a wide range from day to night. Under normal weather conditions at the Viking 1 lander site, the temperature regularly varied from a low near 189 K, just before sunrise, to a high of 240 K in the early afternoon. This is a swing equal to 92° of temperature Fahrenheit, much larger than occurs, for example, in desert regions on Earth. This diurnal variation is greatest near the ground and occurs because of the ability of the surface to quickly radiate its heat to space during the night. During dust storms, this ability is impaired and the daily temperature swing is reduced. Above altitudes of a few kilometres, this diurnal swing is damped out, but other oscillations appear throughout the atmosphere as a result of the direct input of energy from the Sun. These oscillations, which are wavelike in nature and are akin to tides, have been measured as pressure and temperature variations and lead to a very complex wavelike, rather than smooth, vertical structure. The mean temperature of the lower atmosphere falls off with altitude at a rate (or lapse) of about 1.4 K per kilometre to a height of about 40 kilometres and is roughly constant at 140 K above that level. This low lapse rate, caused by the large amount of suspended dust, was unexpected; the lower atmosphere, or troposphere, was anticipated to be convectively unstable and to have an adiabatic lapse rate near 5 K per kilometre, and the tropopause, the upper limit of the troposphere, was expected to occur at much lower altitudes (15 kilometres).

Above an altitude of 100 kilometres, the structure of the atmosphere is determined by the onset of diffusive separation of the various atmospheric molecules. The tendency of heavier molecules to lie below the lighter overcomes the tendency of turbulence to mix the atmospheric gases. Absorption of ultraviolet light from the Sun then disassociates and ionizes the atmosphere's constituents and leads to complex sequences of chemical reactions. The net result is the embedding of an ionized layer near 130 kilometres and an increase in temperature, particularly of the ionic component, at higher altitudes. The top of the atmosphere, or exosphere, is characterized by a temperature of about 300 K averaged over a Martian year.

Meteorology and atmospheric dynamics. The global pattern of the atmospheric circulation of Mars shows many superficial similarities to that of the Earth, although the causes of motion arise from different circumstances. The most important of these are the ability of the Martian air to adjust rapidly to local conditions of solar heat input, the lack of oceans that have a large thermal inertia, the great range in altitude on the surface, the strong internal heating of the atmosphere because of suspended dust, and, finally, the seasonal deposition and release of a large fraction of the Martian atmosphere by the polar caps.

Martian winds

There are no direct measurements of wind velocities in the free atmosphere more than one or two metres above the surface, and knowledge of winds is based on inferences from a global series of atmospheric temperature measurements made from the Mariner 9 orbiter. Such observations can be directly related to the zonal wind field if it is assumed that the pressure forces that drive wind motions are balanced by coriolis forces that arise in a moving atmosphere as a result of planetary rotation. Such a geostrophic balance applies on Earth and it is expected that it should apply equally on Mars. The predicted wind fields show a strong dependence on the Martian season, because large horizontal temperature gradients build up across the edge of the newly forming polar cap in autumn and in winter. The result is the formation at high latitudes of strong jet streams with eastward velocities in excess of 100 metres per second. Elsewhere in the atmosphere, the velocities are much lower, on the order of 10 metres per second. Circulation is less dramatic at the equinox, with light winds predominating everywhere. Surface winds in the lowest one or two kilometres of the atmosphere are strongly controlled by frictional forces and turbulence.

Direct measurements of wind velocities at these low altitudes are available at the Viking lander sites, as are indications of the mean direction of local winds from streaks in windblown dust, patterns in dune fields, and patterns in many varieties of clouds. The low-altitude winds are generally regular in their behaviour and light, the mean wind velocity at the Viking lander sites early in the mission being 0.63 metre per second. On Mars, unlike the Earth, there is also a relatively strong meridional circulation that transports the atmosphere to and from the winter and summer poles. The general circulation pattern is occasionally unstable and exhibits large-scale wave motions and instabilities. The effects of a regular series of cyclones and anticyclones can be clearly seen in the pressure and wind records at the Viking lander sites.

Smaller scale wave motions and tides, driven both by the Sun and by surface topography, are ubiquitous in the atmosphere, as attested to by its irregular vertical structure discussed earlier. This is probably why the atmosphere maintains a uniform composition to extremely low pressure levels. The pressure level of the turbopause, or level of mixing, on Mars is some 50 times lower than that of the Earth. This turbulence is presumably also an important factor in maintaining the large quantity of dust supported in the atmosphere. The state of the atmosphere is highly susceptible to the presence of this dust, and one of the most dramatic and interesting of Martian meteorological phenomena is the global dust storm. Mariner 9 and Viking cameras recorded in excess of 20 local dust storms during their primary missions and in two cases experienced their development into global dimensions.

Dust storms

The dust storms tend to begin at preferred locations in the southern hemisphere during the southern spring and summer. The activity is at first local and vigorous (for reasons that are not yet understood), and large amounts of dust are thrown up high into the atmosphere. If the amount of dust reaches a critical quantity, the storm rapidly intensifies and large amounts of dust are carried aloft by the winds to all parts of the planet. This leads in a few days to the total obscuration of the surface. Measurements at the Viking lander sites indicate that atmospheric transmission is reduced by a factor of 20 or more. The intensification process is evidently short-lived for, almost immediately, opacity of the dust begins to fall off, and in a few weeks the atmosphere returns to normal.

THE SATELLITES

The two satellites of Mars, Deimos and Phobos, were discovered in 1877 by Asaph Hall of the U.S. Naval Observatory. Almost 100 years later, a new phase in the exploration of these objects began with photographs of the shadow of Phobos on the surface of Mars by Mariner 7 and with extensive direct photographic observations from the Viking orbiting spacecraft. Viking 1 flew to within 100 kilometres of Phobos, and Viking 2 was navigated closer than 30 kilometres from Deimos.

Orbits of the satellite

The orbit of Phobos is exceptionally close to the planet. At a mean distance of 2.8 planetary radii from the centre

Table 13: The Satellites of Mars

property	Deimos	Phobos
Orbital radius	23,459 km	9,378 km
Orbital period	1.26244 days	0.31891 days
Mean orbital velocity	1.4 km/sec	2.1 km/sec
Dimensions (kilometres)	15 × 12 × 11	27 × 22 × 19
Area	400 km^2	1,000 km^2
Volume	1,000 km^3	5,000 km^3
Mass	2×10^{18} g	9.6×10^{18} g
Mean density	1.9 g/cm^3	2.0 g/cm^3
Mean escape velocity	4.7 m/sec	7.8 m/sec
Albedo	0.07	0.06

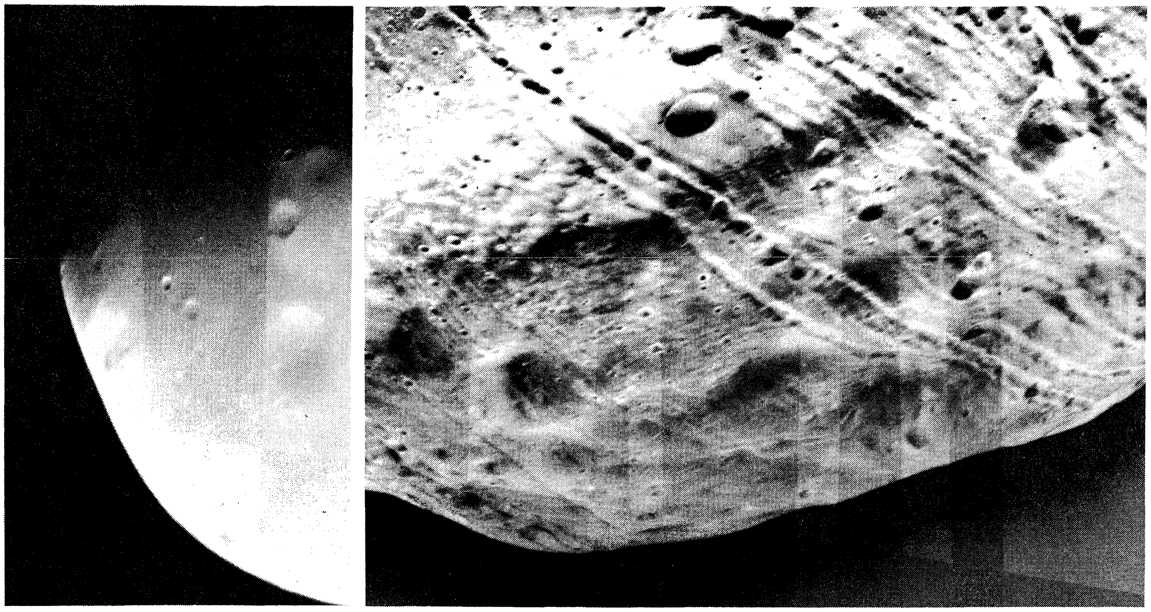


Figure 33: Viking photographs of portions of (left) Deimos and (right) Phobos. The smooth texture of the surface of Deimos stands in contrast to the grooves and pitted craters of the surface of Phobos.

By courtesy of the Jet Propulsion Laboratory/National Aeronautics and Space Administration

of the planet, it is within the Roche limit of 3 planetary radii. (The Roche limit is the minimum distance at which a satellite, if it has no internal strength, can approach its primary body without being torn apart by gravitational forces.) These gravitational, or tidal, forces cause the orbital motion of Phobos to be slowed and may ultimately cause the satellite to fall onto the surface of Mars, possibly in less than 100,000,000 years. The orbit of Deimos suffers an opposite fate, for it moves in a more distant orbit, and tidal forces cause a recession of the satellite from the planet.

The orbital period of Phobos about Mars is seven hours and 39 minutes. This short period means that it travels around Mars twice in a sol. An observer at a suitable point on the planet would see Phobos rise and set twice in a sol.

Not all observers on the surface of Mars could see the moons, however, because of their proximity to the planet and their near equatorial orbits. At latitudes greater than approximately 70° north or south, for example, Phobos could not be seen. The dimensions of the satellites, which cannot be resolved by Earth-based telescopes, have been revealed by photographs taken from the Viking orbiter spacecraft. Each satellite is roughly the shape of a triaxial ellipsoid, with Phobos the larger of the two. In both cases, the long axis of the satellite constantly points towards Mars, and, as with the Earth's moon, both have rotational periods equal to their orbital periods.

Surface features of the satellites

Pictures of the moons' surfaces show striking differences. Phobos' rugged surface is totally covered with craters caused by impacting objects. The largest of these, the Stickney crater, is distinguished by a diameter comparable in size to that of the satellite itself. The surface of Deimos is smooth by comparison, although there are many craters. The chief characteristic of Deimos' surface is that its craters are found to be almost buried in large quantities of fine debris. In addition to a large number of impact craters, Phobos' surface is also characterized by a widespread system of surface fractures, or grooves, many of which are geometrically related to the Stickney crater. It is thought that early in their history these satellites were both subjected to a similar intense bombardment by meteor-like bodies. Their surfaces were fractured and cratered, and large amounts of broken debris were produced. In the case of the more massive Phobos, this debris either remained in place or, if it flew off the satellite, was subsequently lost to Mars. In the case of the more distant and smaller Deimos, any debris that flew off the satellite remained in orbit until it was recaptured by the satellite as a kind of rain that blanketed the surface, inundating the craters and fractures that probably exist there.

The albedo, or reflectivity, of the surfaces of both satellites is very low and has similarities to that of the most primitive types of chondritic meteorites. One theory of the origin of the Martian satellites is that they are asteroids that were captured during the time of Mars's accretion.

The discovery of the satellites was of great importance because the period of their motion and other orbital characteristics provided an estimate of the mass of the parent planet and its internal structure. This information has now been superseded by more accurate and diverse data from orbiting and lander spacecraft. The study of these objects now centres on the contributions they might be able to make toward understanding the general origin of satellites and asteroidal bodies and the kinds of physical and chemical processes that occurred early in the development of the solar system.

THE QUESTION OF LIFE ON MARS

The question of the presence of life on Mars has been an essential element of general discussions of the planet since G.V. Schiaparelli first included an interconnecting system of "canals," or channels, in his maps of the planet in 1877. These canals were perceived in visual telescopic observations as systems of rectilinear markings on the Martian surface. These markings are now known to be an illusion that is probably caused by chance alignments of larger craters and other surface features. Canals, as visualized by Schiaparelli and Lowell, do not exist on Mars.

Observations of seasonal changes in the colour of certain markings on the planet, the ability of major dark areas to recover their appearance after being obliterated by global dust storms, and the springtime "wave of darkening" have also been used to support the speculation that a widespread biota capable of rapid reproduction could be present on the Martian surface. These changes are, in fact, both real and dramatic, but their cause—the movement of vast quantities of fine dust by atmospheric winds—is now known to be physical rather than biological.

The two Viking lander craft searched for the by-products of the metabolism of living organisms, but the results remain a subject for debate because the experiments were limited to samples from only two relatively inhospitable sites and because they addressed only a limited range of many possibilities. What does seem certain is that the surface of Mars does not support a widespread system of living organisms such as that existing on Earth. No trace of the organic detritus that would be associated with a biota such as is ubiquitous on Earth is known to exist on the surface of the planet.

Viking biological experiments

The biological experiments that were conducted by the Viking landers addressed the questions of the nature of organic material, if any, on the surface; the possible presence of objects on the surface whose appearance or motion would suggest living or fossilized organisms; and the possible presence in Martian soil of agents that, under prescribed conditions, could indicate metabolic processes. The answer to the first of these questions was definite and unambiguous; a direct and extremely sensitive chemical analysis on several samples at both lander sites showed no trace of any complex organic materials at all. The cameras on the lander spacecraft addressed the second question in great detail, but no evidence of biological agents or activity of any kind was found. Three separate experiments addressed the last question. One, the pyrolytic release experiment, was designed to look for signs of photosynthesis or chemosynthesis induced by samples of the Martian soil. While this experiment produced some indications of a positive result, the experimenters believe that these can be best explained by nonbiological processes. A second experiment, called the gas exchange experiment, was designed to measure any gaseous products from a soil sample as it was exposed to a humid atmosphere or was treated with a solution of organic nutrients. This experiment also produced a positive result in that the soil samples liberated substantial quantities of oxygen in response to the nutrient. Since this reaction was also found to occur even after the sample had been baked at 145° C for three hours, the experimenters believe that the cause is again nonbiological in origin. Finally, the labelled release experiment was designed to look for the release of radioactive gas when the soil sample was exposed to a radioactive organic nutrient solution. A positive result was again obtained, and in this case the "baked" control sample did become deactivated, as would be expected if the reaction was caused by a biological agent. The consensus of the Viking experimenters is that, while the labelled release experiment did give a positive result, it is still possible to explain it in nonbiological terms, and, taken in the context of the results of the other four experiments, the Viking mission found no persuasive evidence for life on the surface of the planet. (M.J.S.B.)

Jupiter

Jupiter, symbol ♃ in astronomy, is the most massive of the planets and the fifth in distance from the Sun. When ancient astronomers named the planet Jupiter for the ruler of the gods in the Greco-Roman pantheon, they had no idea of the planet's true dimensions, but the name is appropriate, for Jupiter is larger than all of the other planets combined. It has a narrow system of rings and 16 known satellites, one larger than the planet Mercury and three larger than our Moon. Jupiter also has an internal heat source; *i.e.,* it emits more energy than it receives from the Sun. This giant also has the strongest magnetic field of any planet, with a magnetosphere so large that it would exceed the apparent diameter of our Moon if it could be seen from Earth. Jupiter's system is the source of intense bursts of radio noise, at some frequencies occasionally radiating more energy than the Sun.

Knowledge about the Jovian system grew dramatically during the 1970s. The new information came in part from Earth-based observations, but especially from two sets of spacecraft—Pioneers 10 and 11 in 1974–75 and Voyagers 1 and 2 in 1979 (see Figure 34). The Pioneer spacecraft served as scouts for the Voyagers, showing that the radiation environment of Jupiter was tolerable and mapping out the main characteristics of the planet and its environment. The more sophisticated instrumentation on the later spacecraft then filled in the details, providing so much new information that it was still being analyzed in the early 1980s.

BASIC ASTRONOMICAL DATA

Table 14 shows physical and orbital data for Jupiter. Of special interest are the low mean density of 1.33 grams per cubic centimetre (g/cm³)—in contrast with Earth's 5.5 grams per cubic centimetre—coupled with the large dimensions and mass, and the short rotational period. The low density and large mass suggest that Jupiter's composition and structure are quite unlike those of Earth and other inner planets, a supposition supported by detailed investigations of its atmosphere and interior.

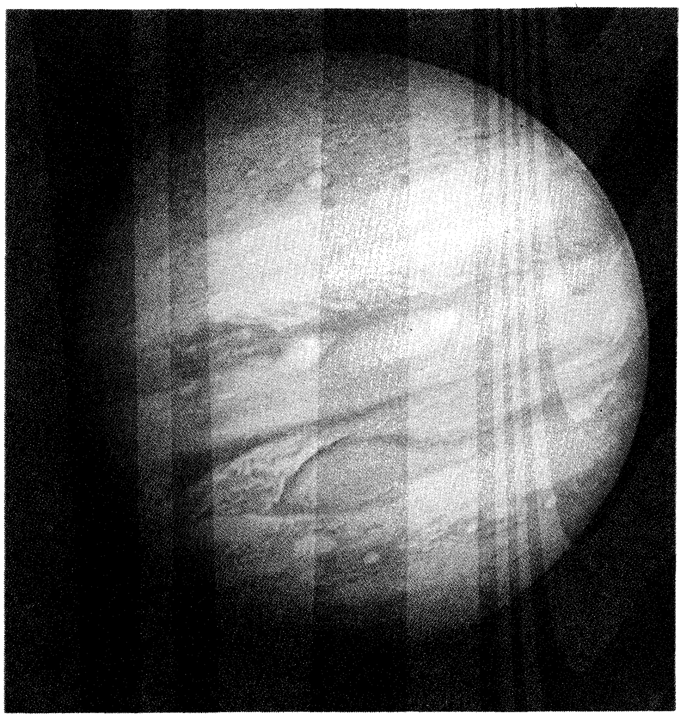


Figure 34: Photograph of Jupiter taken by Voyager 1 on February 1, 1979, at a range of 32,700,000 kilometres.
By courtesy of the Jet Propulsion Laboratory/National Aeronautics and Space Administration

Three rotational periods have been established. The two periods labelled Systems I and II (Table 14) are mean values and refer to the speed of rotation at the equator and at higher latitudes, respectively, as exhibited by features observed in the planet's visible cloud layers. Jupiter has no solid surface; the transition from the atmosphere to the core occurs gradually at great depths. This variation in rotation period at different latitudes does not imply, therefore, that the solid body of the planet rotates with either of these mean velocities. In contrast, the apparent constancy of the period deduced from observations at radio wavelengths, System III, during 20 years of observation,

Table 14: Basic Data

Orbital characteristics		
Semimajor axis	5.20 a.u. (483.4 × 10^6 miles)	
Sidereal period	11.86 years	
Eccentricity	0.049	
Inclination	1°19′	
Physical elements		
Mean apparent diameter	47 seconds of arc	
Equatorial radius	71,400 km	= 11.2 (Earth = 1)*
Polar radius	66,770 km	= 10.4 (Earth = 1)
Ellipticity $\left(\frac{R_e - R_p}{R_e}\right)$	0.065	= 18 (Earth = 1)
Mass	18.99 × 10^{29}	= 317.8 × Earth's mass = $\frac{1}{1047.39}$ × Sun's mass
Mean density	1.33 g/cm³	= 0.24 × Earth's density
Surface gravity	2,288 cm/s²	= 2.65 × Earth's gravity
Escape velocity	59.5 km/sec	= 5.45 × Earth's escape velocity
Rotation periods		
System I (±10° from equator)	9h50m30s	
System II (higher latitudes)	9h55m40s	
System III (magnetic field)	9h55m29s	
Inclination of equator to orbit	3°04′	
Miscellaneous data		
Dimensions of Great Red Spot	26,000 km × 14,000 km	
Mean apparent visual magnitude	−2.55	
Magnetic field strength at equator	4.3 gauss	= 13.8 × Earth's surface field

*Radius of Earth = 6,378 km or 3,963 mi

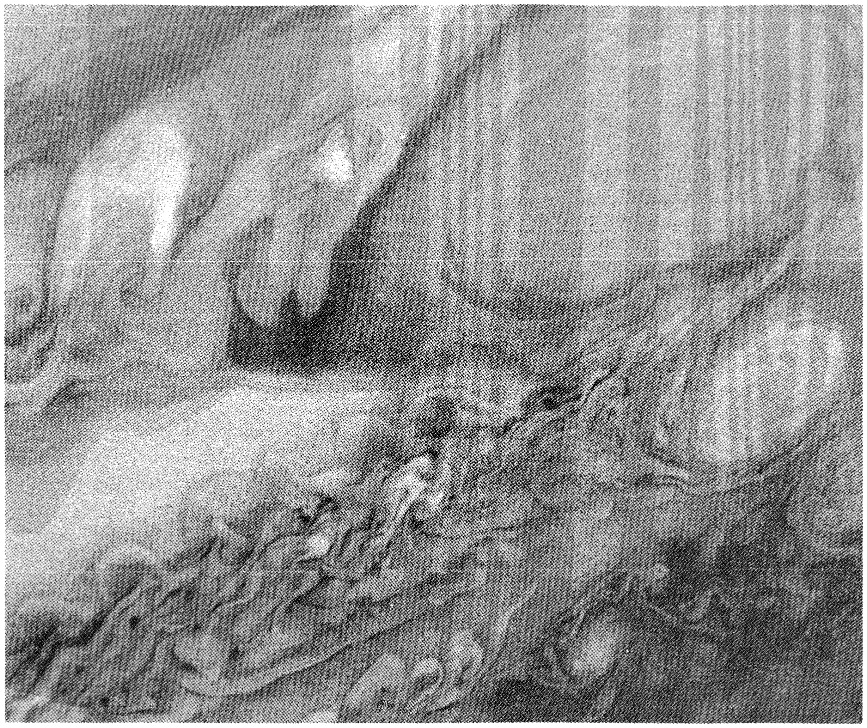


Figure 35: The Great Red Spot (top right) and the surrounding region photographed by Voyager 1 on March 1, 1979. At centre right is one of the white ovals visible from Earth.

By courtesy of the Jet Propulsion Laboratory/National Aeronautics and Space Administration

suggests that it is associated with the planet's magnetic field, formed deep in Jupiter's liquid or semisolid interior.

THE OUTER LAYERS

The clouds and the Great Red Spot. Even a modest telescope can show much detail on Jupiter. The region of the planet's atmosphere that is seen from Earth contains several different types of clouds that are separated both vertically and horizontally. Changes in these cloud systems can occur in a few hours, but an underlying pattern of latitudinal currents has maintained its stability for decades. (It has become traditional to describe the appearance of the planet in terms of a standard nomenclature for alternating dark belts and bright zones. The currents, however, seem to have greater persistence than this pattern.)

Various cloud forms

The close-up views of Jupiter from the Voyager spacecraft revealed a variety of cloud forms, with a predominance of elliptical features reminiscent of cyclonic and anticyclonic storm systems on Earth. All of these systems are in motion, appearing and disappearing on time scales dependent upon their sizes and locations. Also observed to vary are the pastel shades of various colours present in the cloud layers—from the tawny yellow that seems to characterize the main layer, through browns and grays, to the well-known salmon-coloured Great Red Spot, Jupiter's largest, most prominent, and longest-lived feature (see Figure 35). The vertical and horizontal segregation of the cloud systems is evidently accompanied by chemical differences as well.

Jovian meteorology can be compared with the global circulation of the Earth's atmosphere. On Earth huge spiral cloud systems often stretch over many degrees of latitude and are associated with motion around high- and low-pressure regions. These cloud systems are much less zonally confined than the cloud systems on Jupiter and move in latitude as well as longitude. Local weather on Earth is often closely tied to the local environment, which in turn is determined by the varied nature of the planet's surface.

Absence of a solid surface

Jupiter has no solid surface, hence no relief features, and the planet's large-scale circulation is dominated by latitudinal currents. The lack of physical boundaries on Jupiter's surface makes the persistence of these currents and their associated cloud patterns all the more remarkable. The Great Red Spot, for example, moves in longitude with respect to all three of the rotation systems, yet it does not move in latitude. The three white ovals found at a latitude just south of the Great Red Spot exhibit similar behaviour; white ovals of this size are found nowhere else on the planet. The dark brown clouds, evidently holes in the tawny cloud layer, are found almost exclusively at latitudes near +18°. The blue-gray or purple areas, from which the strongest thermal emission is detected, only occur in the equatorial region of the planet.

The nature of the Great Red Spot. The true nature of Jupiter's unique Great Red Spot was still unknown by the early 1980s, despite extensive observations from the Voyager spacecraft. On a planet whose cloud patterns have lifetimes often counted in days, the Great Red Spot has survived as long as detailed observations of Jupiter have been made—at least 100, and perhaps 300, years. There is some evidence that the spot may be slowly shrinking, but a longer series of observations is needed to confirm this suggestion. Its present dimensions are about 26,000 by 14,000 kilometres, making it large enough to accommodate, side by side, two planets the size of Earth. These huge dimensions are probably responsible for the feature's longevity and possibly for its distinct colour.

The rotation period of the Great Red Spot with respect to the rotation of Jupiter itself shows a variability that has not been successfully correlated with other Jovian phenomena. Earth-based observations in 1966 and 1967 revealed the counterclockwise circulation of the material within the spot itself to have a period of 12 days. This period was confirmed by the Voyager observations, which recorded a large number of interactions between the Great Red Spot and much smaller disturbances moving in the current at the same latitude (Figure 35). The Voyager pictures showed the interior of the spot to be remarkably tranquil, with no clear evidence for the expected upwelling (or divergence) of material from lower depths.

The Great Red Spot, therefore, appears to be a huge anticyclone, a vortex, or eddy, whose lateral dimensions are greater than the Earth's diameter. This lateral size is presumably accompanied by a huge vertical extent that allows the feature to extend well below and well above the main cloud layers. The extension above the main clouds can be observed directly, and it also is manifested by lower temperatures and by less gas absorption above the Great Red Spot than at neighbouring regions on the planet.

The tawny clouds

Cloud composition. The clouds that are recorded in pictures of Jupiter are formed at different altitudes in the planet's atmosphere. Except for the top of the Great Red Spot, the white clouds are the highest, with cloud-top temperatures of about 140 K. The white clouds apparently consist of frozen ammonia crystals and are thus analogous to the water-ice cirrus clouds in the Earth's atmosphere. At lower levels occur the tawny clouds that are widely distributed over the planet. They appear to form at a temperature of about 220 K, which suggests that they probably consist of condensed ammonium hydrosulfide (NH_4SH) and that their colour may be caused by other ammonia-sulfur compounds such as ammonium monosulfide, $(NH_4)_2S$, or ammonium polysulfides, $(NH_4)_2S_x$ (x may equal 2, 3, . . .).

The reasons for invoking sulfur compounds as likely colouring agents (chromophores) lie in the relatively high cosmic abundance of sulfur and the absence of any evidence of hydrogen sulfide (H_2S) in Jupiter's atmosphere. Jupiter is primarily composed of hydrogen and helium. Under equilibrium conditions—allowing all of the elements to react with one another at an average Jovian pressure and temperature—a predominance of methane (CH_4), ammonia (NH_3), water (H_2O), and hydrogen sulfide over other compounds involving these abundant elements would be expected. This outcome is true for carbon, nitrogen, and oxygen, but no form of sulfur has been directly detected on Jupiter. This can be understood if the presence of sulfur compounds—formed in part because of the ease with which hydrogen sulfide can be broken down by solar ultraviolet light—in the lower clouds is postulated. In this hypothesis the sulfur made available by photo-dissociation is free to combine with the more abundant ammonia to form tawny chromophores.

Sulfur compounds have also been proposed to explain the dark brown coloration of the clouds detected at still lower levels, where the measured temperature is 260 K. These clouds are seen through what are apparently holes in the otherwise ubiquitous tawny clouds. They appear bright in pictures of Jupiter that are made from the thermal radiation detected at a wavelength of five micrometres.

Localized cloud chemistry

The colour of the Great Red Spot has been attributed to the presence of complex organic molecules, red phosphorus, or yet another sulfur compound. All of these ideas find support in laboratory experiments, but there are counter arguments in each case. Dark regions occur near the heads of white plume clouds near the planet's equator, where temperatures as high as 300 K have occasionally been measured. Despite their blue-gray appearance, these dark features have a reddish tint. They may be clear gas exhibiting a blue colour (from Rayleigh scattering) overlain with a thin haze of reddish material. That these regions occur only at the equator, the elliptical dark brown clouds only near +18°, and the most prominent red colour on the planet only in the Great Red Spot implies a localization of cloud chemistry that is particularly puzzling in such a dynamically active atmosphere.

At still lower depths in the atmosphere, astronomers expect to find water-ice clouds and water droplet clouds, both consisting of dilute solutions of ammonium hydroxide (NH_4OH). These cloud layers are expected to be reached by a probe to be sent into the Jovian atmosphere by the Galileo Project planned for the late 1980s.

The atmosphere. *The proportions of constituents.* Until a probe has entered Jupiter's atmosphere, studies of the planet's spectrum must be relied upon to provide information about the composition, temperature, and pressure of the atmosphere. In this technique light or thermal radiation from the planet is spread out in wavelengths (colours, in visible light, as in a rainbow) by the dispersing element in a spectrograph. The resulting spectrum indicates that there are discrete intervals at which energy has been absorbed by the constituents of the planet's atmosphere. By measuring the exact wavelengths at which this absorption takes place and comparing the results with spectra of gases obtained in the laboratory, the gases in Jupiter's atmosphere can be identified.

The presence of methane and ammonia in Jupiter's atmosphere was deduced more than 50 years ago, while hydrogen was detected for the first time in 1960. (Although 500 times more abundant than methane, hydrogen has much weaker absorption lines because it is a molecule of two identical atoms that interacts only very weakly with electromagnetic waves.) Subsequent studies led to a growing list of new constituents, including the discovery of hydrogen cyanide (HCN) in 1981. Table 15 includes a list of Jupiter's atmospheric constituents and their abundances as determined by early 1982.

Table 15: Atmospheric Abundances

gas	mixing ratio*	element ratio	Jupiter/Sun
H_2	1		
He	0.12	He/H	1
CH_4	2×10^{-3}	C/H	2
NH_3	2×10^{-4}	N/H	1
C_2H_6	4×10^{-5}		
H_2O	10^{-6}	O/H	10^{-2}
C_2H_2	8×10^{-7}		
PH_3	4×10^{-7}	P/H	1
CH_3D	3.8×10^{-7}	D/H	†
CO	3×10^{-9}		
HCN	2×10^{-9}		
GeH_4	6×10^{-10}	Ge/H	10^{-1}
C_3H_8	(detected)		

*The mixing ratio is the number of molecules of a given atmospheric constituent in a unit volume divided by the number of hydrogen molecules in that same volume.

†Deuterium is not present on the Sun because of nuclear burning. The value of D/H on Jupiter is 3.5×10^{-5} ($\pm 1.5 \times 10^{-5}$), which is approximately 2.3 times the present interstellar value.

If the condition of chemical equilibrium held rigorously in Jupiter's atmosphere, molecules such as carbon monoxide (CO), hydrogen cyanide, acetylene (C_2H_2), and ethane (C_2H_6) would not occur in the abundances shown in Table 15. Sources of energy other than the molecular kinetic energy corresponding to local temperatures are evidently available. Solar ultraviolet radiation is responsible for the breakdown of methane and subsequent reactions of its fragments into acetylene and ethane. In the convective region of the atmosphere, lightning discharges contribute to these processes and may be responsible for the production of hydrogen cyanide. Still deeper, at temperatures around 1200 K, carbon monoxide is made by a reaction between methane and water vapour. Vertical mixing must be sufficiently strong to bring this gas up to a region where it can be detected from outside the atmosphere.

Differences between Jupiter and the Sun

Table 15 includes a comparison of elemental abundances in Jupiter's atmosphere with the composition of the Sun. If the planet had formed by simple condensation from the primordial solar nebula (see below), the abundances of the elements should be the same in both Jupiter and the Sun. Instead, there is a real spread in the values for different elements. On the other hand, except for methane and helium, the abundances of the gases from which the elemental abundances are derived depend on dynamical phenomena in Jupiter's atmosphere—principally condensation and vertical mixing. Thus the apparent enrichment in carbon may be an indication of a real difference in composition between Jupiter and the Sun.

Another difference is indicated by the presence of deuterium (D) on Jupiter. This heavy isotope of hydrogen has disappeared from the Sun as a result of nuclear reactions in the solar interior. Because no such reactions occur on Jupiter, the ratio of deuterium to hydrogen there should be identical to the ratio of those isotopes in the cloud of interstellar gas and dust that collapsed to form the solar system 4,600,000,000 years ago. Since deuterium may have been made primarily in the "big bang" that has been postulated to have begun the expansion of the universe, an accurate measurement of deuterium/hydrogen on Jupiter would allow the calibration of expansion models.

Temperature and pressure. The best profiles for the relation between changes in temperature and pressure in the Jovian atmosphere have resulted from the Pioneer and Voyager measurements. Both passed behind the planet as viewed from Earth, and the attenuation of the signal as

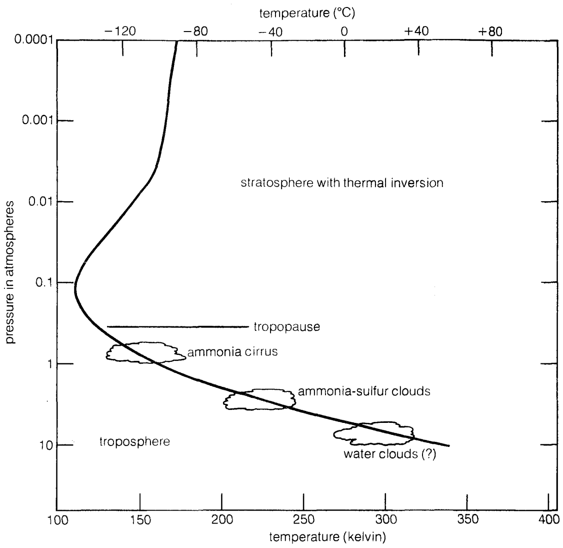


Figure 36: The structure of the atmosphere of Jupiter as deduced from Voyager measurements.

it passed through Jupiter's atmosphere before being completely extinguished provided a measure of the change in atmospheric density, which is a function of temperature and pressure. A self-consistent analysis of the data is in agreement with independent studies based on analyses of the thermal radiation escaping from the planet.

One of the profiles resulting from these studies is shown in Figure 36. From the figure it is possible to locate the positions of the principal cloud decks. It is interesting to notice that temperatures higher than the freezing point of water occur at pressures just a few times greater than the sea-level pressure on Earth. This is mainly a consequence of Jupiter's internal energy source, although some warming would occur just by the trapping of infrared radiation by the atmosphere in the so-called greenhouse effect.

Temperature inversion

The increase in temperature above the tropopause is known as an inversion because temperature normally decreases with height. The inversion is caused by the absorption of solar energy at these altitudes by gases and aerosol particles.

Other likely atmospheric constituents. The list of atmospheric abundances in Table 15 is certainly not complete. For example, although the noble gas neon has about the same cosmic abundance as nitrogen, it, like helium, is very difficult to detect by spectroscopic observations. Neon had not been observed in the Jovian atmosphere by the early 1980s, even though it should be as abundant as ammonia.

Complex organic molecules in Jovian atmosphere

The formation of complex organic molecules in Jupiter's atmosphere is of great interest in the study of the origin of life. The initial chemical processes leading to the formation of living organisms may have occurred on Earth at a time when the composition of the terrestrial atmosphere was very similar to the present Jovian atmosphere, allowing for an appropriate depletion of hydrogen and helium. The active Jovian cloud system is a source of electrical discharges, while solar ultraviolet radiation, precipitation of charged particles, and the internal energy of the planet are also available to drive chemical reactions in the Jovian atmosphere. Thus, Jupiter may well represent an enormous natural laboratory in which the initial steps toward the origin of life are being pursued again and again. A determination of the degree of complexity reached under such conditions constitutes one of the most fascinating problems confronting any program of space exploration.

Radio emission. Jupiter was the first planet found (in 1955) to be a source of radiation at radio wavelengths. The radiation was recorded at a frequency of 22 megahertz (*i.e.,* a wavelength of 13.6 metres, or 1.36 decametres) in the form of noise bursts with peak intensities sometimes great enough to make Jupiter the brightest source at this wavelength, except for the Sun during its most active phase. The bursts of radio noise from three distinct areas constituted the first evidence for a Jovian magnetic field. Subsequent observations at shorter (decimetre) wavelengths revealed that Jupiter is also a source of steady radio emission. It has become customary to refer to these two types of emission in terms of their characteristic wavelengths: decametre radiation (the erratic bursts) and decimetre radiation (the continuous source).

Synchrotron emission

The nonthermal component of the decimetric radiation is interpreted as synchrotron emission; that is, radiation emitted by very high-speed electrons moving in a magnetic field within a toroidal, or doughnut-shaped, region surrounding Jupiter—a phenomenon closely analogous to that of the terrestrial Van Allen belts. The maximum emission occurs at a distance of two planetary radii from the centre of the planet and has been detected at 178–5,000 megahertz. The position of the plane of the polarization (vibrations of the radio emission preferentially being in a plane) and intensity of the radio emission vary with the same period. Both effects are explained if the axis of the planet's magnetic field is inclined by about 10° to the rotational axis. The period of these variations is the period designated as System III (see Table 14).

The radio emission at decametre (10-metre) wavelengths has been studied from Earth in the accessible range of 3.5 to 39.5 megahertz (3.5×10^6 to 39.5×10^6 cycles per second). Freed from the low-frequency cutoff established by the Earth's ionosphere, the radio-wave experiment on the Voyager spacecraft was able to detect emissions from Jupiter down to 60 kilohertz (60×10^3 cycles per second), corresponding to a wavelength of five kilometres. The strength of the radio signal and the frequency of noise storms show a marked time dependence that led to the early detection of three "sources," or emitting regions. The System III coordinate system was initially defined through the periodicity of these sources.

Effect of Io on noise storms

The noise storms are greatly affected by the position of the satellite Io in its orbit. For one source, events are much more likely to occur when Io is 90° from superior geocentric conjunction (*i.e.,* 90° from the position in which Earth, Jupiter, and Io are in a straight line) than otherwise. The noise sources appear to be regions that lie in the line of sight toward the optical disk of the planet (unlike the nonthermal decimetric radiation).

The most promising explanation of the effect of the orbital motion of Io on noise storms relates the emission to a small region of space linked to Io by magnetic field lines (a flux tube) that move with Io. Electrons moving in spirals around the magnetic field lines could produce the observed radiation. Interaction between these electrons and the Jovian ionosphere can be expected and were, in fact, observed by the Voyager spacecraft.

The Jovian magnetosphere. The nonthermal radio emissions described above are the natural result of trapped charged particles interacting with Jupiter's magnetic field and ionosphere. Interpretation of these observations led to a remarkably accurate definition of the basic characteristics of the planet's magnetosphere that was supported by the direct exploration of fields and particles in the vicinity of Jupiter by the Pioneer and Voyager spacecraft. The basic magnetic field of the planet is dipolar in nature, generated by a hydromagnetic dynamo that is driven by convection within the electrically conducting outer layers of Jupiter's interior. The magnetic moment is 19,000 times greater than that of Earth, leading to a field strength at the equator of 4.3 gauss, compared with 0.3 gauss at the Earth's surface. The axis of the magnetic dipole is offset by 0.1 R_J (where R_J is Jupiter's equatorial radius of 71,400 kilometres) from Jupiter's rotational axis, to which it is indeed inclined by 10°. The orientation of the Jovian magnetic field is exactly opposite to the present orientation of the Earth's field, such that a terrestrial compass taken to Jupiter would point south (relative to the planet's vector of angular momentum).

Magnetic links with Io

The magnetic field dominates the region around Jupiter to a distance of about 10 R_J to a point between the orbits of its satellites Europa and Ganymede (see below *The satellites and rings*). Within this region the most striking activity is generated by the satellite Io, whose influence on the decametric radiation has been discussed above. An electrical current of approximately 5×10^6 amperes flows in the magnetic flux tube linking Jupiter and Io. This satellite is also the source of a torus (doughnut-shaped) cloud of ions that surrounds its orbit.

THE INTERIOR

The atmosphere of Jupiter constitutes only a very small fraction of the planet, much as the skin of a fruit compares with its contents. Nothing can be directly observed below this thin outer layer, so that indirect conclusions are drawn from the evidence in order to determine the composition of the interior of Jupiter.

The observed quantities with which the astronomer can work are the atmospheric temperature and pressure, mass, radius, shape, rate of rotation, heat balance, and perturbations of satellite orbits and spacecraft trajectories. From these follow the ellipticity and the departure from an ellipsoid of the planet, quantities that may also be predicted using theoretical descriptions, or models, for the internal distribution of material. Such models can then be tested by their agreement with the observations.

Jupiter's central temperature and pressure

The basic difficulty in constructing a model that will adequately describe the internal conditions for Jupiter is the absence of extensive laboratory data on the properties of hydrogen and helium at pressures and temperatures appropriate to the conditions near the centre of this giant planet. The central temperature is estimated to be close to 25,000 K, to agree with an internal source of heat that allows Jupiter to radiate about twice as much energy as it receives from the Sun. The central pressure is in the range of 50,000,000–100,000,000 atmospheres. The required extrapolation is not quite as dramatic as it may seem, since at such tremendous pressures hydrogen assumes a metallic state, the properties of which can be calculated with some confidence. The difficulty lies, first, in establishing the transition point at which the metallic state occurs and, second, in defining the properties of hydrogen between this transition point and the region where laboratory measurements are available. A third difficulty is posed by the solubility of helium in hydrogen at these pressures and temperatures, resulting in a multicomponent system.

Predominance of hydrogen and helium

Despite these problems, there has been a steady improvement in the precision of the models. Perhaps the most significant early conclusion from these studies was the realization that Jupiter cannot be composed entirely of hydrogen; if it were, it would have to be considerably larger than it is to account for its mass. On the other hand, hydrogen must predominate, constituting at least 78 percent of the planet by mass, no matter in what form—gas, liquid, or solid—it may occur. If it is assumed that the bulk of the rest of the planet is helium, the proportion of hydrogen to helium is on the order of 14:1, in close agreement with the atmospheric value and present ideas for the composition of the Sun (see Table 15). Current models agree on a transition from molecular to metallic hydrogen at approximately 0.75 R_J. It should be stressed that this is not a transition between a liquid and a solid but rather between two liquids with different electrical properties. In the metallic state the electrons are no longer bound, thus giving the hydrogen the conductivity of a metal. No solid surface exists in any of these models, although most contain a rock and ice core with a radius of 0.03–0.10 R_J (0.33–1.1 Earth radii).

The source of internal heat has not been resolved. The favoured explanation in the early 1980s invoked the gradual release of primordial heat left over from the planet's formation. In other words, the conversion of gravitational energy to thermal energy that initially led to a very hot Jupiter is still progressing with a very gradual contraction of the planet (see below *Theories of the origin of the Jovian system*).

THE SATELLITES AND RING

The first objects in the solar system discovered by means of a telescope were the four brightest satellites of Jupiter. Galileo, who first observed them in 1610, proposed that the satellites be named the Medicean stars, in honour of his patron, Cosimo II de' Medici, but they soon came to be known as the Galilean satellites in honour of their discoverer. Galileo regarded their existence as a fundamental argument in favour of the Copernican model of the solar system in which the planets orbit the Sun.

In order of increasing distance from the planet, these satellites are called Io, Europa, Ganymede, and Callisto, for legendary figures closely associated with Jupiter (Zeus) in Greek mythology. The names were assigned by the German astronomer Simon Marius, Galileo's contemporary and rival. There proved to be a certain further aptness in the choice of Io's name: Io—"the wanderer" (Greek *iōn*, "going")—has an indirect influence on the ionosphere of Jupiter. The fifth satellite was also discovered by telescope, while the other 11 known satellites were found photographically (JVI–JXIII) or in pictures obtained by the Voyager spacecraft (JXIV–JXVI). Roman numerals are assigned to the satellites in order of their discovery.

Satellite groups. Jovian satellite data are summarized in Table 16. The orbits of the inner eight satellites have low inclinations and low eccentricities; *i.e.,* they all are nearly circular. The orbits of the outer eight have much higher inclinations and eccentricities. The two innermost satellites (XIV and XVI) seem to be intimately associated with Jupiter's ring and may in fact be two of the sources of the fine particles within the ring itself. There are almost certainly additional members of all of these groups.

The Galilean satellites. Although approximate diameters and spectrophotometric characteristics of the four largest satellites of Jupiter had been determined from ground-based observations, it was the Voyager missions of 1979 that indelibly established these four bodies as worlds in their own right (see Figure 37). Before Voyager, it was known that Callisto and Ganymede were both as large or

Table 16: Satellites of Jupiter

roman numeral	name	discoverer	year of discovery	mean distance from Jupiter (km)	sidereal period (days)	orbital inclination (degrees)	orbital eccentricity	radius (km)	mass (kg)	mean density (g/cm³)
XVI	Metis	S. Synnott	1979	128,000	0.295	?	?	20	9.494×10^{16}	?
XV	Adrastea	D. Jewitt, E. Danielson	1979	129,000	0.298	?	?	12.5 × 10 × 7.5	1.899×10^{16}	?
V	Amalthea	E.E. Barnard	1892	181,000	0.498	0.40	0.003	135 × 83 × 75	7.215×10^{18}	?
XIV	Thebe	S. Synnott	1979	222,000	0.674	0.8	0.015	55 × 45	7.595×10^{17}	?
I	Io	Galileo, S. Marius	1610	422,000	1.769	0.04	0.004	1,815	8.886×10^{22}	3.55
II	Europa	Galileo, S. Marius	1610	671,000	3.551	0.47	0.009	1,569	4.785×10^{22}	3.04
III	Ganymede	Galileo, S. Marius	1610	1,070,000	7.155	0.21	0.002	2,631	1.481×10^{23}	1.93
IV	Callisto	Galileo, S. Marius	1610	1,883,000	16.689	0.51	0.007	2,400	1.075×10^{23}	1.83
XIII	Leda	C. Kowal	1974	11,094,000	238.72	26.07	0.148	8	5.696×10^{15}	?
VI	Himalia	C.D. Perrine	1904	11,480,000	250.566	27.63	0.158	93	9.494×10^{18}	?
X	Lysithea	S.B. Nicholson	1938	11,720,000	259.22	29.02	0.107	18	7.595×10^{16}	?
VII	Elara	C.D. Perrine	1905	11,737,000	259.653	24.77	0.207	38	7.595×10^{17}	?
XII	Ananke	S.B. Nicholson	1951	21,200,000	631	147	0.169	15	3.798×10^{16}	?
XI	Carme	S.B. Nicholson	1938	22,600,000	692	164	0.207	20	9.494×10^{16}	?
VIII	Pasiphae	P. Mellote	1908	23,500,000	735	145	0.378	25	1.899×10^{17}	?
IX	Sinope	S.B. Nicholson	1914	23,700,000	758	153	0.275	18	7.595×10^{16}	?

larger than the planet Mercury; that they and Europa had surfaces covered with water ice; that Io was surrounded by a torus of atoms and ions that included sodium, potassium, and sulfur; and that the inner two satellites have mean densities much larger than those of the outer two. This density gradient resembles that found in the solar system itself and seems to result from the same cause (see below *Theories of origin*). The implication is that Io and Europa have a rocky composition similar to that of our Moon, whereas approximately 60 percent of Ganymede and Callisto must be a much less dense substance, with water ice as the most likely candidate.

Callisto. The surface of this satellite is so dominated by impact craters that there are no smooth plains such as the dark maria on the Moon. In other words, there seem to be no areas on Callisto where subsequent internal activity has obliterated any of the record of early bombardment. This record was formed by impacting debris (comet nuclei and asteroidal material) primarily during the first 500,000,000 years after the formation of the solar system in much the same way that the craters on Earth's moon were produced.

Ganymede. Unlike Callisto, Ganymede, an equally icy satellite, reveals distinct patches of dark and light terrain. This contrast is reminiscent of what can be seen on the Moon, but the association of chronology with albedo (reflected brightness) is exactly reversed. The dark regions on Ganymede are the oldest areas, showing the heaviest concentration of craters. The light regions are younger, revealing a complex pattern of parallel and intersecting ridges and grooves in addition to unusually bright impact craters typically surrounded by systems of rays. Like Callisto, Ganymede also exhibits relatively low topographic relief, indicating the action of viscous flow of the icy surface.

Smoothness of Europa

Europa. The surface of Europa is totally different from that of Ganymede and Callisto, despite the fact that the infrared spectrum of this object indicates that it, too, is covered with ice. There are few impact craters on Europa—the number per unit area is about comparable to that on the continental regions of Earth. This indicates that the surface that can be seen is relatively recent. Some scientists think the surface is young enough to suggest that significant resurfacing is still taking place on the satellite. This resurfacing evidently consists of the outflow of water from the interior to form an instant frozen ocean. Its frozen surface is crisscrossed by a pattern of dark stripes and curved ridges. The relief is extremely low, with ridge heights perhaps a few hundred metres at most. Europa thus has the smoothest surface of any of the solid bodies examined in the solar system thus far.

Io. Seen through a telescope, Io looks reddish-orange, while the other satellites are neutral in tint. Io's infrared spectrum shows no evidence of the absorptions of water ice. Scientists expected Io's surface to look different from those of Jupiter's other moons, but the Voyager pictures revealed a landscape even more unusual than anticipated.

Io's volcanoes

Volcanic calderas, instead of impact craters, dot the surface of Io, and 10 volcanoes were observed in eruption when the two Voyager spacecraft flew by in 1979. This unprecedented level of activity makes Io the most tectonically active object in the solar system. The surface of the satellite is continually and completely replaced in spans of time considered short when compared with the age of the solar system. Various forms (allotropes) of sulfur appear to be responsible for the black, orange, and red areas on the satellite's surface, while solid sulfur dioxide is probably the main constituent of the white areas. Sulfur dioxide was detected as a gas near one of the active volcanic plumes by Voyager's infrared spectrometer and was identified as a solid in ultraviolet and infrared spectra obtained from Earth-orbital and ground-based observations. These identifications provide sources for the sulfur and oxygen ions observed in the Jovian magnetosphere.

The energy for this volcanic activity requires a special explanation, since radioactive heating is inadequate for a body as small as Io. The favoured explanation is based on the observation that orbital resonances with the other Galilean satellites perturb Io into a more eccentric orbit than it would assume if only Jupiter controlled its motion. The resulting tides developed by the contest between the other satellites and Jupiter may release enough energy to account for the observed volcanism.

Surface of Amalthea

Other satellites. The only other Jovian satellite that was close enough to the trajectories of the Voyager spacecraft in 1979 to allow surface features to be seen was Amalthea. So small that its gravitational field is not strong enough to deform it into a sphere, it has an irregular, oblong shape (see Table 16). Like Io, its surface exhibits a reddish colour that may result from a coating of sulfur compounds released by Io's volcanoes.

The ring. One of the tasks of the Pioneer 11 mission of 1974 was to monitor the charged particles around Jupiter. As the spacecraft sped toward its closest approach of 1.6 R_J, a sudden decrease in the density of charged particles was detected at a distance of 1.7–1.8 R_J. This led to the suggestion that a moon or a ring of material might be orbiting the planet at this distance. The existence of a ring was verified by the first Voyager spacecraft when it crossed the planet's equatorial plane, and the second spacecraft recorded additional pictures, including a series taken in the shadow of the planet looking back at the ring material. The ring is comprised of large numbers of micrometre-sized particles that produce strong forward scattering of incident sunlight. Submicrometric dust is also present, as indicated by a faint halo of material surrounding the ring plane. The halo particles are charged and move out of the equatorial plane in response to the rocking motion of the magnetic equatorial plane as the planet rotates.

Source of Jupiter's ring

The presence of such small particles requires a source. Indeed, the finest material extends all the way in to the planet itself. It seems likely that the source of this material is large boulders, or small moonlets, within the ring. Visible examples of what are presumably among the largest of such objects are satellites JXIV and JXVI. The ring particles are generated by impacts of micrometeoroids, cometary debris, and possibly volcanically produced material from Io. It seems plausible that the inner edge of Jupiter's ring is defined by the orbit of one of these moonlets, even as the outer edge appears to be defined by satellite JXIV.

THE GRAVITATIONAL INFLUENCE OF JUPITER

Resonances between the paths of Jupiter and the minor planets. Jupiter exerts a gravitational influence extending well beyond its own satellite system to affect minor planets of the solar system and comets as well. Among the minor planets with well-determined solar orbits, the distribution of their periods of revolution about the Sun shows striking gaps; there are no 5.9–, 4.8–, and 4-year periods, which correspond to $^1/_2$, $^2/_5$, and $^1/_3$, respectively, of the orbital period of Jupiter (see Table 14). These commensurabilities evidently lead to resonance-enhanced perturbations that prevent a body from remaining in an orbit of these durations. On the other hand, 19 minor planets are known with periods close to $^2/_3$ that of Jupiter, and none with periods slightly smaller or larger than this. This indicates another type of resonance perturbation, one acting to concentrate rather than disperse.

Trojan planets. Jupiter appears to "collect" minor planets, among which the Trojans form an important group. These bodies all have very nearly the same orbital period as Jupiter and are found at two points that are regions of gravitational stability, where the attraction of Jupiter and the Sun have equal effect (L_4 and L_5 equilibrium points of Lagrange's solution to the three-body problem; see MECHANICS). Perturbations by Saturn appear likely to cause membership of this group to change, a possibility that has led to the suggestion that the Trojans may be former satellites of Jupiter and not true minor planets.

The Jupiter family of comets. Jupiter also has a strong influence on the orbits of comets. A comet that originally has a very long period and an orbit that carries it beyond the planet Pluto may be converted to a short-period comet, with an aphelion of five astronomical units or less, as a result of a near encounter with Jupiter; in other words, the comet's farthest distance from the Sun is reduced to five or less times the mean distance of Earth from the Sun. A pronounced concentration of short-period comets, the paths of which cross the principal plane of the solar system at Jupiter's distance from the Sun (about five as-

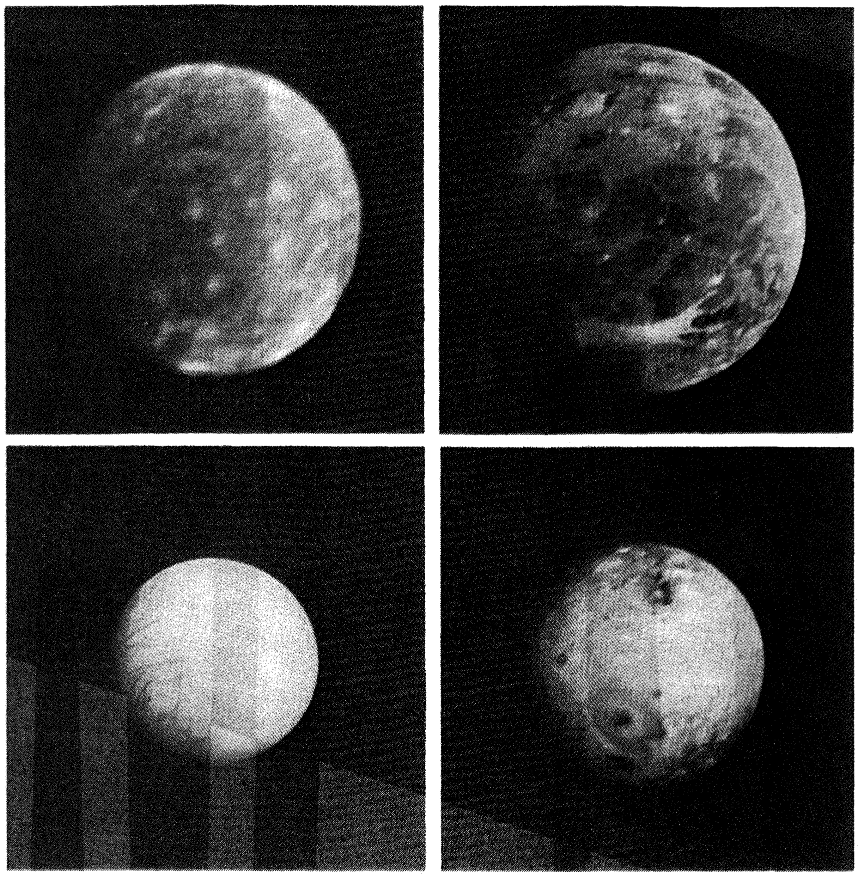


Figure 37: *The four Galilean satellites of Jupiter.*
(Top) Callisto and Ganymede. (Bottom) Europa and Io. Photographs taken by Voyager 1 between March 1 and 3, 1979.

By courtesy of the Jet Propulsion Laboratory/National Aeronautics and Space Administration

tronomical units), suggests that these comets have been strongly influenced by Jupiter's gravitational field. One may thus speak of a "Jupiter family" of comets, the best known member of which is Encke's Comet.

THEORIES OF THE ORIGIN OF THE JOVIAN SYSTEM

The origin of Jupiter and its satellites is part of the problem of the origin of the entire solar system. Current thinking favours the gradual development of the Sun and planets from a huge cloud of gas and dust containing gravitational instabilities. Details of the later evolutionary picture vary, however, especially on the role of magnetic fields and on the relative importance of condensation and accretion.

Early history of Jupiter. Given the planet's large proportion of hydrogen and its huge mass, it has been traditional to assume that Jupiter formed by condensation from the primordial solar nebula. This hypothesis implies that the elements should all be present on Jupiter in the same proportions that they occur in the Sun. The most recent evidence (Table 15) suggests that this may not be true.

Accretion of dust and ices

Current models for Jupiter's origin suggest that a core formed first as a result of the accretion of grains of dust and ices into accumulating planetesimals. The heavier elements in the core would have been present in solar proportions, but as the core grew and began to attract more volatile material, a nonsolar composition could have resulted. In particular, models have been proposed with a core containing high proportions of water ice and an envelope of hydrogen and helium. Although these light gases, relative to the heavier elements in the core, do not reach their solar abundance, the ratio of hydrogen to helium is solar in all of these models, as appears to be the case on Jupiter itself.

Early history of the satellites. The inner eight satellites of Jupiter are commonly thought to have originated in much the same way as the planet itself. Just as the primordial solar nebula is believed to have broken up into accreting planetesimals with a central condensation that became the Sun, the accumulation of material into a proto-planetary cloud at Jupiter's orbit ultimately led to the formation of the planet and its inner satellites. The analogy goes further. The high temperature of the forming planet apparently prevented volatile substances from condensing at the distances of the innermost satellites. Hence Ganymede and Callisto represent the volatile-rich outer bodies in this system.

The origin of the outer eight satellites, with their orbits of high eccentricities and inclinations, is thought to be quite different. They are members of the population of irregular satellites in the solar system and have most likely been captured by Jupiter.

(T.C.O.)

Saturn

Saturn, symbol ♄ in astronomy, is the sixth planet in order of distance from the Sun. It cannot approach the Earth closer than about 1,190,000,000 kilometres. Its brightness is due to its large size, inferior only to that of Jupiter. Saturn's equatorial diameter is 120,660 kilometres, but its globe is appreciably flattened, and the polar diameter is only 108,000 kilometres.

Saturn's ring system is in a class of its own. While Jupiter and Uranus also have rings, those of Saturn are striking, and a telescope of moderate power will show them excellently. There can be no doubt that Saturn is one of the most beautiful objects in the entire sky.

Saturn was the outermost of the planets known in ancient times. The earliest known observations of it, by the Babylonians, can be reliably dated to the mid-7th century BC, but it was probably noticed much earlier, as it is brighter than most stars. To the naked eye it appears yellowish; it moves slowly against the background of stars. The Greeks named it after Cronus, the original ruler of Olympus, whose Roman equivalent is the god Saturn.

Table 17: Planetary Data for Saturn

Distance from the Sun	mean 9.54 a.u. (1,472,000,000 km) maximum 10.07 a.u. (1,507,000,000 km) minimum 9.01 a.u. (1,347,000,000 km)
Sidereal period of revolution	10,759.20 days, or 29.46 years
Mean synodic period	378.1 days
Rotation period (mean)	10 hours 39 minutes 24 seconds
Mean orbital velocity	9.6 km/sec
Axial inclination	26°7′
Orbital inclination to the ecliptic	2°29′22″
Orbital eccentricity	0.056
Diameter (equatorial)	120,660 km
(polar)	108,000 km
Apparent diameter seen from Earth	maximum 21″ minimum 15″
Mass (Earth=1)	95.26
Volume (Earth=1)	744
Density (water=1)	0.7
Albedo	0.61
Satellites	at least 20

BASIC ASTRONOMICAL DATA

Mass, density, and surface gravity. The mass of Saturn is 95.26 times that of the Earth, or $^1/_{3,500}$ of that of the Sun, and the escape velocity—the velocity which, once attained, will enable an object to "coast" away from the planet—is 32.26 kilometres per second, more than three times that of Earth. The density of Saturn is surprisingly low—only 0.7 that of water—and its surface gravity is 1.16 times that of the Earth. Saturn's outer layers are made up of gas, and it is a world quite unlike our own.

Orbit. Among planetary orbits, that of Saturn is of fairly low eccentricity (0.056), though the difference in distance between perihelion (closest approach to the Sun) and aphelion (farthest retreat from the Sun) amounts to 160,000,000 kilometres. Because Saturn is so far from the Sun and the Earth, phase effects for its disk are negligible; that is, Saturn always appears to be full, or nearly so, in the sense of the Full Moon.

Synodic period

The mean synodic period, the interval between successive oppositions, when the Earth passes approximately between Saturn and the Sun, is 378.1 days, so that Saturn is well placed for observation during several months in each year. Opposition dates from 1982 to 1990 are April 8, 1982; April 21, 1983; May 3, 1984; May 15, 1985; May 27, 1986; June 9, 1987; June 20, 1988; July 2, 1989; and July 14, 1990.

Like all the outer planets, Saturn moves for the greater part of each year eastward against the starry background; its average rate is about 1° in eight days. As it approaches opposition, however, its motion seems to slow down and to stop altogether about 70 days before the opposition date. For a period that may be as little as 133 days or as great as 141 days, it then seems to move in a retrograde, or westward, direction before reaching another stationary point and resuming its eastward movement. This behaviour does not, of course, indicate any real alteration in motion. The apparent regression is due to the fact that the Earth, moving in a much smaller orbit at a greater velocity (average 29.8 kilometres per second as against only 9.6 kilometres per second for Saturn), is catching up with Saturn and passing it.

When at conjunction, on the far side of the Sun, Saturn cannot be observed, as it passes too close to the Sun in the sky. The inclination of the orbit to the ecliptic is 2°29′22″. During its synodic period of 378 days, Saturn advances for nearly eight months and retrogrades slowly for about 4½ months, so that on balance it progresses along its path by about 12° per year. Data for Saturn are given in Table 17.

HISTORY OF OBSERVATION

Early telescopic observations. The first telescopic observations of Saturn are thought to have been made by Galileo in July 1610. He saw the disk of the planet clearly, but his telescope gave a magnification of only 32 diameters and was not good enough to show the ring system in recognizable form; moreover, conditions were unfavourable because the rings were placed almost edgewise. Galileo thought that Saturn must be a triple planet and wrote that "Saturn is not one alone, but is composed of three, which almost touch one another." Two years later, he found to his surprise that the "companions" had vanished, so that Saturn appeared as a single object. The ring system was then edge-on to the Earth, so that it could not be seen at all in Galileo's telescope. The original aspect was again seen in the years following 1613, but Galileo was never able to interpret it correctly. Subsequently, various strange theories were proposed to explain the planet's unusual form; Hevelius of Danzig, for instance, believed Saturn to be elliptical in shape, with two appendages attached to the surface.

The problem was solved by the Dutch astronomer Christiaan Huygens, who began his observations in 1655. The telescopes that he used were much more powerful than Galileo's, and gave sharper definition, so that in a short time he was satisfied that "Saturn is surrounded by a thin, flat ring which nowhere touches the body" of the planet. His theory met with a surprising amount of opposition, but by 1665 it had been universally accepted, even though the nature of the ring system was not established until much later.

Modern Earth-based studies. Saturn's colour is yellowish; darker belts, parallel with its equator, are always seen. These belts are not nearly so conspicuous as those on Jupiter, nor do they show so much detail. Saturn's greater distance and smaller size make it less easy to study than Jupiter; moreover, the beauty of Saturn's ring system tends to divert attention from its disk, particularly when the rings are wide open, thereby hiding a considerable part of the globe.

Saturn's albedo (the proportion of incident sunlight it reflects) is 0.61. The planet's apparent magnitude, however, depends largely upon the angle at which the ring system is displayed, largely because the rings are more reflective than the disk. When the rings are wide open, the magnitude attains −0.3, so that, of the stars, only Sirius and Cano-

By courtesy of the Jet Propulsion Laboratory/National Aeronautics and Space Administration

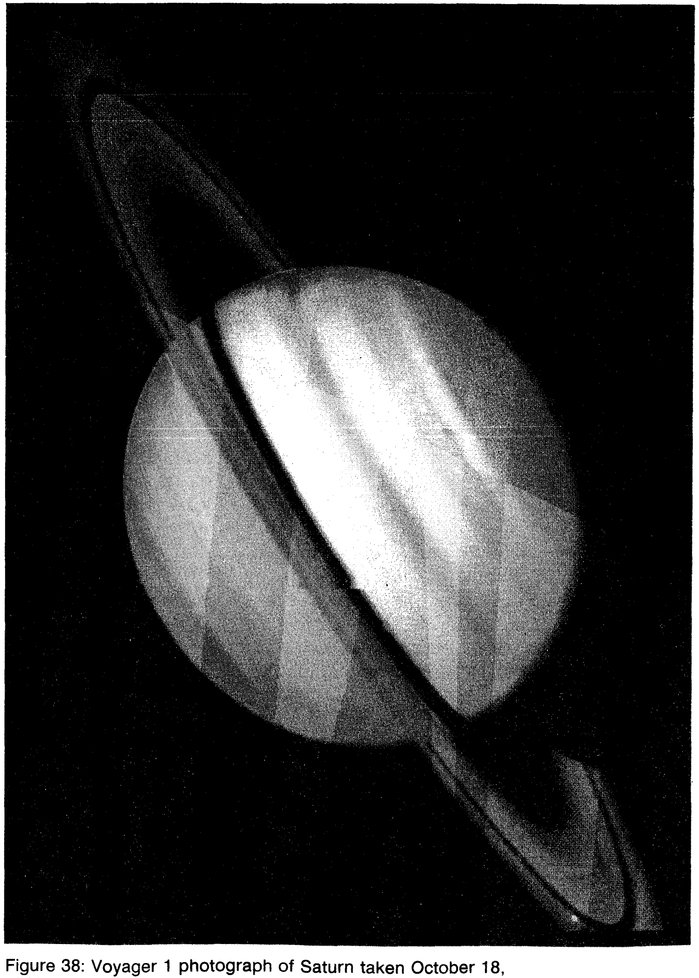


Figure 38: Voyager 1 photograph of Saturn taken October 18, 1980, and colour-enhanced to increase the visibility of features in Saturn's North Temperate Belt.

pus appear brighter than Saturn. At oppositions when the rings are edge on, as in 1980, the magnitude is as low as +0.8, though even at these times Saturn is still prominent.

Polar and equatorial zones

The equatorial zone appears creamy, sometimes almost white, in colour. The polar regions are always less brilliant; the belts, unlike those of Jupiter, do not show obvious colours. As with Jupiter, the two most prominent belts flank the equatorial zone; the other belts are much less conspicuous. Unless the rings are at a narrow angle, one or other of the equatorial belts will be covered.

Well-defined spots on Saturn are rare. The most prominent example sighted during the present century was that discovered on August 3, 1933, by the English amateur W.T. Hay, using a 15-centimetre refractor. The spot took the form of a large white oval patch in the equatorial zone, about one-fifth of the planet's diameter in length and with both ends well defined. During the next few weeks it lengthened rapidly, until by mid-September it had spread out so much that it could no longer be called a spot.

Space probes to Saturn. In September 1979 Pioneer 11 became the first unmanned space probe to rendezvous with Saturn. It had already passed by Jupiter, which had been its main objective, but valuable information about Saturn was obtained. The second probe was Voyager 1, which made its closest approach on November 12, 1980, when it passed only 124,200 kilometres above Saturn's clouds; its twin, Voyager 2, passed within 101,400 kilometres of Saturn's clouds on August 25, 1981.

Detailed views and a great amount of new information were obtained by the Voyager probes. Though Saturn's disk is blander than that of Jupiter because of a greater amount of overlying "haze," much detail was shown. A red spot was detected in the southern hemisphere, with brownish ovals in the northern hemisphere. Measurements of the locations of individual features at different times yielded data concerning the speed of the winds, and it seems that Saturn's circulation is different from that of Jupiter. The maximum westward velocities of Saturn occur near the centres of the darker regions, while on Jupiter they occur at the poleward interfaces between the belts and bright zones. In the polar latitudes, the large-scale light and dark bands break down into small-scale features, giving the impression of waves and eddies. The wind velocity at the cloud tops seems to be about 1,400 kilometres per hour, or twice that of Jupiter.

COMPOSITION

Possible iron and rock core

According to a recent theory, Saturn has an iron and rock core that extends out 13,000 to 14,000 kilometres from its centre and is so compressed that it contains 15 to 20 times the mass of the Earth. Surrounding the core is a layer of electrically conductive metallic hydrogen in liquid form, outside of which is an envelope of hydrogen and helium. There is a magnetic field 1,000 times stronger than that of the Earth, though far weaker than that of Jupiter; the magnetic axis is almost coincident with the axis of rotation, and the magnetosphere extends out beyond the orbit of the largest satellite, Titan. There are fairly strong radiation belts, but these, again, are much weaker than those of Jupiter.

The temperature at the cloud tops of Saturn is approximately −180° C. A theory that the planet might be self-luminous has been disproved. Saturn, like Jupiter, however, releases about twice the amount of energy it receives from the Sun, indicating the existence of an internal source of heat. While Jupiter emits energy from the gravitational contraction that occurred when the planet was formed about 4,600,000,000 years ago, it is unlikely that the heat source of Saturn is similar. Saturn is smaller and less massive and has a lower overall density (less than that of water); therefore, any heat remaining from the gravitational contraction of Saturn would have been dissipated long ago. Instead, the heat source of Saturn may result from the separation of hydrogen and helium in the planet's outer layers, with the heavier helium sinking through the liquid hydrogen middle layer.

THE SATELLITES

Saturn has at least 20 known satellites, details of which are given in Table 18. An additional satellite, named Themis, was reported by W.H. Pickering in the early 20th century, but it has not been recovered and probably does not exist. Several small satellites have been detected since 1977; they are extremely faint.

The satellites Calypso and Telesto move in orbits close to the inner and outer edges of Ring F and presumably keep this ring stable. Janus and Epimetheus move in the same orbit and must periodically approach each other closely, though clearly they do not collide. The satellite designated 1980S6 moves in the same orbit as Dione.

Phoebe, the outermost satellite, was not within range of the Voyager 1 probe, and little is known about it; it may be a captured asteroid. Iapetus is unique in having one hemisphere of high albedo and the other of low albedo; this explains why, as seen from the Earth, it appears much brighter when west of Saturn than when east of the planet. As with all the other satellites, except Phoebe, Iapetus has a captured, or synchronous, rotation; *i.e.*, its axial rotation period is equal to the time taken to complete one orbit, as is true of our Moon with respect to the Earth. Rhea, Dione, Tethys, and Mimas have icy, cratered surfaces; one huge crater on Mimas has more than one-third the diameter of the satellite itself. There is also a very large crater on Tethys. Enceladus has a surface that is smooth over wide areas; other regions have young, well-formed, and relatively small craters.

Dense atmosphere of Titan

Titan, by far the largest of Saturn's satellites, is in a class of its own because of its dense atmosphere. Data obtained from Voyager 1 indicate that the atmospheric pressure at Titan's surface is 1.5 to two times that of the Earth at sea level. The actual surface of the satellite is obscured by the top of an orange-coloured layer of what may be called "photochemical smog." Titan's atmosphere is made up mainly of nitrogen with smaller amounts of methane and cyanide; it is possible that the intensely cold surface is covered, at least in part, by oceans of liquid methane.

Table 18: Selected Satellites of Saturn

satellite	distance from centre of Saturn (kilometres)	period (days)	orbital eccentricity	orbital inclination (degrees)	diameter (kilometres)*
Atlas	137,670	0.602	0.000	0.3	40 × 20
Calypso	294,660	1.888	—	—	34 × 22 × 22
Telesto	294,660	1.888	—	—	34 × 28 × 26
Epimethius	151,422	0.694	0.009	0.33	140 × 120 × 100
Janus	151,472	0.694	0.007	0.14	220 × 200 × 160
Mimas	185,520	0.942	0.020	1.53	392
Enceladus	238,020	1.370	0.004	0.00	500
Tethys	294,660	1.888	0.000	1.86	1,060
Dione	377,400	2.737	0.002	0.02	1,120
1980S6	377,400	2.737	0.005	0.0	36 × 32 × 30
Rhea	527,040	4.518	0.001	0.35	1,530
Titan	1,221,830	15.945	0.029	0.33	5,150
Hyperion	1,481,100	21.277	0.104	0.43	410 × 260 × 220
Iapetus	3,561,300	79.330	0.028	14.72	1,460
Phoebe†	12,952,000	550.48	0.163	177	220

*The diameters of some of the smaller satellites remain somewhat uncertain. †Phoebe moves around Saturn in a retrograde direction.

THE RINGS

The rings of Saturn are much more prominent than those of Jupiter or Uranus, and they are different in nature. Saturn's rings are made up of icy or ice-covered particles, each ring moving around Saturn in its own independent orbit. The ring plane remains in an almost fixed position with reference to the stars, but as seen from the Earth and the Sun the tilt of the rings is continually changing. Twice in each Saturnian revolution—at alternate intervals of 13 years nine months and 15 years nine months—the plane of the rings passes through the Sun. A few months before and after each such occasion, the ring plane passes through the Earth, which is near the Sun as viewed from Saturn. During the shorter interval, when Saturn passes through perihelion, the south pole of Saturn is tilted toward the Sun, so that the southern aspect of the rings is illuminated and the northern hemisphere of the planet is partly obscured. When the northern face is displayed, and the southern hemisphere of the globe is partly obscured, Saturn passes through aphelion; it is then moving slowly in its orbit, resulting in the inequality of the two intervals between successive edge-on presentations. The ring system is thin, so that at times of edge-on presentation the rings almost disappear as seen from the Earth.

Two of the three main rings—called A (outer) and B—are bright. In 1675 G.D. Cassini, at the Paris Observatory, discovered the gap between them, known as Cassini's Division. The "ring" seen by Huygens was a combination of the two, and his telescope was not good enough to show the division. Ring B is much the brighter of the two. In 1837 J.F. Encke, at Berlin, discovered a less prominent division in Ring A, known as Encke's Division. In 1850 a dusky inner ring, Ring C, was discovered by W.C. Bond and G.P. Bond at Harvard and independently by W.R. Dawes in England; this is generally known as the Crepe (Crêpe) Ring. The outer diameter of Ring A is 272,300 kilometres, the inner diameter 239,600 kilometres; for Ring B the outer and inner diameters are 234,200 kilometres and 181,100 kilometres, respectively. The inner diameter of Ring C is 149,300 kilometres, so that it extends to 17,000 kilometres above the cloud tops.

Dimensions of the main rings

Additional rings have since been discovered. A dusky ring, closer in than Ring C, was reported by telescopic observers; Voyager 1 results indicate that a ring, designated Ring D, does exist there, but whether it is genuinely observable from Earth is dubious. Outside Ring A is the clumpy and braided Ring F; Rings E and G were also shown on the Voyager 1 images.

The Voyager probes showed that the rings are much more complex than had been thought. Each contains hundreds of components, and there are several distinct narrow rings inside the Cassini Division. After examining photographs transmitted by Voyager 2, scientists estimated that the ring system is composed of thousands of rings and smaller "ringlets." It was once thought that the main divisions in the ring system were due to the cumulative perturbation effects of Saturn's satellites, but this explanation is clearly inadequate, and other forces must be involved. (P.Mo.)

Uranus

Uranus, symbol ⛢ in astronomy, is the seventh planet from the Sun in distance. Although less than half the size of Jupiter, it is still the third largest planet of the solar system. Its diameter is four times that of the Earth. Uranus is faintly visible to the naked eye, but it has been recognized as a planet for only 200 years. Physically, Uranus is almost a twin of Neptune, being only 3 percent larger and actually 13 percent less massive than Neptune. These two, with Jupiter and Saturn, constitute the Jovian planets characterized by large size, low density, and compositions consisting primarily of the very light elements.

EARLY OBSERVATIONS

The planets out through Saturn have been known since antiquity; Uranus was the first major body of the solar system to be discovered after the invention of the telescope. The English astronomer William Herschel undertook a survey of all stars down to the eighth magnitude—*i.e.,* those about 10 times fainter than stars visible to the naked eye—using a 15-centimetre telescope of his own manufacture at his observatory in Bath. On March 13, 1781, he found "a curious either nebulous star or perhaps a comet," distinguished from the stars by its clearly visible disk. Its lack of any trace of tail and its slow motion led to the suggestion that the observations were consistent with a planet moving in a nearly circular orbit. Within a year, independent studies established that the orbit was planetary, with a radius exceeding 18 astronomical units. The object was thus twice as far from the Sun as Saturn, and its discovery at once doubled the size of the known solar system.

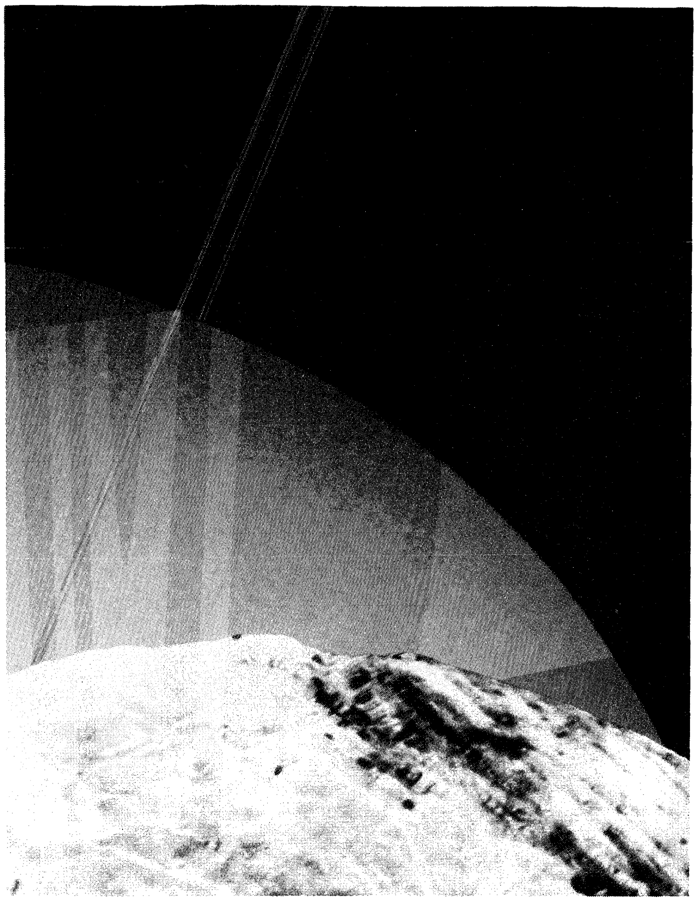


Figure 39: Montage of Voyager 2 photographs taken in January 1986 that simulates a view of Uranus and rings as if seen over the horizon of Miranda, one of the satellites of Uranus.

By courtesy of the Jet Propulsion Laboratory/National Aeronautics and Space Administration

Three names were seriously considered for the new planet. Herschel proposed first to call it Georgium Sidus (in English, Georgian Planet) after his sovereign patron, King George III of England. The name Georgian appeared intermittently in England for over 50 years, especially in the *Nautical Almanac.* In France the name Herschel was also used occasionally until the mid-19th century. But already in the year of discovery Johann Elert Bode of Germany had suggested Uranus, as the father of Saturn in Roman mythology, who was, in turn, the father of Jupiter. This name was eventually universally accepted.

Progressive improvements in knowledge of the planet's motion made possible the projection of its orbit farther into the past, in a search for prediscovery observations for which accurate coordinates had been noted by observers who assumed the planet to be a star. The earliest of 22 such prediscovery observations proved to have been made in 1690, nearly a century before Herschel's discovery.

BASIC ASTRONOMICAL DATA

Uranus' orbit. The orbit of Uranus departs only slightly from a circle, the deviation corresponding to an elliptical orbit with eccentricity of 0.047. The mean distance from the Sun is 19.18 a.u.; the eccentricity brings Uranus to a perihelion (or point nearest the Sun) of 18.3 a.u. and an aphelion (point farthest from the Sun) of 20.1 a.u. The orbit is very slightly inclined, by 0.77°, to the ecliptic plane (the Earth's orbit).

Uranus completes one revolution around the Sun in 84.01 years. In the nearly 300-year span over which at least some observations exist, the planet has thus completed about three and a half revolutions. This is sufficient to test rigorously any proposed orbit of Uranus, including a careful accounting for the perturbations (gravitational influences) exerted by the other known planets of the solar system. But determinations of precise orbits for Uranus proved vexing from the beginning. By 1830 the observed motion departed from a best fitting ellipse by an intolerable 15″ of arc. Several astronomers suggested that perturbations by a hitherto unknown planet, lying outside the orbit of Uranus, might explain the anomalies. Computations, first in 1843–45 by an English mathematician and astronomer, John Couch Adams, and in 1845–46 by a French astronomer, Urbain-Jean-Joseph Le Verrier, led to precise predictions of the position of the hypothetical planet. The resulting discovery of Neptune in 1846 provided a stunning vindication of gravitational theory.

Appearance and spectrum. Under good conditions, Uranus shows a perceptible blue-green disk, through even a 10-centimetre telescope. This astronomically unusual colour is caused by absorption of red light in its atmosphere. Experienced visual observers report faint markings on the disk, in particular, grayish belts. Photographs taken by the U.S. Voyager 2 spacecraft during its flyby of Uranus (the probe passed within 81,000 kilometres of the planet's cloud tops) in 1986 revealed that such features are similar to the latitudinal cloud belts on Saturn though far less conspicuous. Several individual cloud formations resembling thunderheads on Earth also were detected by the probe but only between 25° and 50° latitude.

Faint markings

On the stellar-magnitude scale, in which larger numbers correspond to fainter brightness, Uranus over the past century had an average visual brightness at mean opposition (that is, when closest to the Earth) of about 5.8 magnitudes (the faintest stars visible to the naked eye on a very clear, totally dark night are about 6.5). The observed brightness can change by half a magnitude from the mean, depending on the Earth–Uranus and Uranus–Sun distances and also possibly on real change of reflectivity. Recent accurate photoelectric measures give a mean-opposition (photoelectric-visual) magnitude of about 5.5.

Uranus' rotation. In the decades 1910–30, several sets of visual observations of Uranus indicated rapid brightness variations. These were interpreted as arising from the rotation of a spotted planet, and a period of rotation of 10.82 hours was inferred. At the same time, spectrographic observations of the tilt of spectral lines in sunlight reflected from Uranus were claimed to be consistent with a rotation period of about 10.75 hours; for the next half century all reference works accepted this rotation period for Uranus. Modern analyses of these older data, however, have shown that they cannot be used to determine the correct value for rotation. Half a dozen attempts with very large telescopes were made in the 1970s to derive the Uranus rotation period using modern spectroscopic equipment; these gave widely divergent values ranging from 13 to 24 hours.

Recent rotation calculations

Voyager 2's discovery in 1986 of a comparatively strong magnetic field around Uranus enabled astronomers to make more accurate estimates of its rotation period. An analysis of field data (*e.g.,* radio emission from the field) indicated that Uranus rotates once every 17.3 hours or so.

The rotation of Uranus leads to an appreciable reduction in the effective force of gravity at the equator, hence to a substantial equatorial bulging. The equatorial diameter is about 3 percent greater than the polar diameter.

Mass, diameter, density, and composition. The orbital period and orbital radius of a satellite suffice to establish the mass of its primary, a result first derived by Newton. Recent determinations give Uranus a mass of 8.6978×10^{25} kilograms. This is about 14.5 times greater than that of the Earth and about 22,900 times less than that of the Sun.

The most accurate value for the size of Uranus comes from stellar occultation data. Timing the interval required for the planet, in its orbital motion, to pass in front of a star yields an equatorial radius of 25,700 kilometres, corresponding to four times the radius of Earth.

With the mass and radius of Uranus given above, its average density is 1.23 times the density of water; for comparison, that of Saturn is 0.70 and that of Earth is 5.52. So low a mean density requires Uranus to be composed of materials far lighter than those that make up the Earth; otherwise, its greater mass would lead to compression producing greater central pressures and densities than the Earth's. By the same reasoning, its average composition must be somewhat heavier than that of Saturn. Theoretical models of Jupiter and Saturn are consistent with a total composition similar to the basic solar chemical abundance (roughly 70 percent hydrogen by mass, most of the remainder helium). But in Uranus these two lightest elements must be accompanied by appreciable amounts of the next heavier common elements, especially carbon, oxygen, and nitrogen. Data transmitted by Voyager 2 has provided strong evidence that beneath Uranus' thick atmosphere lies an immense "ocean" of water and smaller amounts of methane and ammonia. This ocean, possibly more than 8,000 kilometres deep, appears to be heated to several thousand kelvins. The liquid, however, does not boil away because of the tremendous pressure from the weight of the overlying atmosphere. In all likelihood, this vast superheated ocean surrounds a core of heavy, rocky material (*e.g.,* silicate compounds) in a molten state. Investigators theorize that slow churning within the ocean and perhaps the central core generates Uranus' magnetic field.

THE ATMOSPHERE

Since the 1860s the spectrum of Uranus has been known to contain deep absorption bands toward the red. In 1932 these were shown theoretically by Rupert Wildt to be almost certainly due to methane (CH_4). Laboratory and high-resolution spectroscopic studies confirmed this. The photographic spectrum of Uranus is dominated by methane, corresponding to about 3.5 kilometre atmospheres of the gas (a kilometre atmosphere is the amount of a pure gas that would be encountered in a one-kilometre path length if the gas were at 0° C and one earth-atmosphere pressure). Like those of Jupiter and Saturn, Uranus' atmosphere must, nevertheless, consist primarily of hydrogen and helium; but, because both these gases are nearly invisible spectroscopically at low temperatures, the methane dominates. Weak hydrogen quadrupole lines are found in the spectrum of Uranus, in amounts consistent with several hundred kilometre atmospheres of this gas in the visible part of the atmosphere. Other common gases, primarily water and ammonia, would be frozen out of the upper atmosphere and be present in quantity only at depths to which the line of sight from Earth cannot penetrate.

The temperature of a planet can be determined by direct measurement of the infrared radiation that it emits. This technique indicates that the effective temperature of the visible outer atmosphere of Uranus is about 60 K. Data derived from Voyager 2 indicates that temperature increases with depth. Uranus radiates nearly 1.3 times as much heat as it receives from the Sun. Like the other three Jovian planets, it in effect generates its own heat albeit in modest amounts.

Temperature of Uranus

THE SATELLITES

Uranus has 15 known satellites, the most of any planet except for Jupiter and Saturn. In 1787 William Herschel found two of the satellites, later called Oberon and Titania. Two more satellites, Ariel and Umbriel, were discovered in 1851 at Liverpool. The names for these four satellites were apparently proposed by the astronomer Sir John Herschel, the son of William Herschel. The last of the five major Uranian satellites to be found, Miranda, less than 10″ of arc from the planet, was first photographed in 1948 at the McDonald Observatory. Ten substantially smaller satellites were discovered by Voyager 2 inside the orbit of Miranda, the innermost of the major Uranian moons.

Uranus' satellite system is compact and closely coplanar with the planet's equator. But, whereas the equators and satellite orbits of all the other planets have only modest inclinations to the plane of the ecliptic, in the case of Uranus these are all curiously tilted, by an extreme 98°. Thus, the rotation of the planet and revolution of its satellites are essentially perpendicular and even slightly retrograde to the orbital revolution of Uranus about the Sun.

Table 19: Major Satellites of Uranus

name	date found	discoverer	telescope size (in.)	satellite diameter (km)	orbit period (days)	magnitude (V_0)
I Ariel	1851	W. Lassell	24	1,330	2.52	14.4
II Umbriel	1851	W. Lassell	24	1,110	4.14	15.3
III Titania	1787	W. Herschel	19	1,600	8.71	13.9
IV Oberon	1787	W. Herschel	19	1,630	13.46	14.2
V Miranda	1948	G.P. Kuiper	82	484	1.41	16.5

THE RINGS

Uranus occulted a relatively bright star on March 10, 1977. Several observers of this rare event independently noted additional occultations corresponding to a hitherto unknown system of rings. Two additional rings were later discovered by Voyager 2, bringing the total to 11 rings. The entire ring system is thought to extend from about 41,600 to 50,700 kilometres from Uranus' centre. The rings are only about one to 10 kilometres wide, except for the one designated epsilon, which is between 30 and 100 kilometres wide. They consist chiefly of boulder-sized chunks roughly one metre in diameter. The rings have among the lowest reflectivities of any known material in the solar system, similar to the reflectivity of very dark carbonaceous chondritic meteoric material. They lie essentially in the plane of the planet's equator and of the satellite orbits but have very slight tilts and ellipticities, which suggest that they may have been formed fairly recently in the history of the solar system. (H.Sm./Ed.)

Low reflectivities

Neptune

Neptune, symbol ♆ in astronomy, is the eighth planet from the Sun and, until Pluto was discovered in 1930, was the outermost known planet in the solar system. Its mean distance from the Sun is almost exactly 30 astronomical units. Because part of Pluto's orbit is closer to the Sun than its orbit, Neptune periodically becomes, in fact, the most distant planet from the Sun. Such is the case as Pluto moved inside Neptune's orbit in 1979 and will remain there until early 1999. Neptune, being only as bright as a star of the eighth magnitude, is invisible without optical aid.

EARLY OBSERVATIONS AND DISCOVERY

The detection of Neptune through its action upon Uranus is one of the fascinating stories in the history of astronomy and a striking example of the precision of the theory of planetary motions. Before its actual discovery by the English astronomer Sir William Herschel in 1781, Uranus had been observed as a fixed star on at least 17 other occasions, first by the English astronomer John Flamsteed in 1690. The French astronomer Alexis Bouvard, in 1820, constructed tables of the motions of Jupiter, Saturn, and Uranus; the observed positions of Jupiter and Saturn were represented satisfactorily, but he found it impossible to do this for Uranus. During the next two decades there was continued divergence between the theory of perturbations, describing how the known planets should alter slightly one another's motions, and the observations. The idea of an ultra-Uranian planet occurred to Bouvard, the German astronomer Friedrich Wilhelm Bessel, the American astronomer William J. Hussey, and others.

Bessel, in 1842, announced his intention of investigating the problem of the motion of Uranus. Unfortunately, he died before much had been done. Meanwhile an English astronomer, John Couch Adams, and a French astronomer, Urbain-Jean-Joseph Le Verrier, were also tackling the problem energetically.

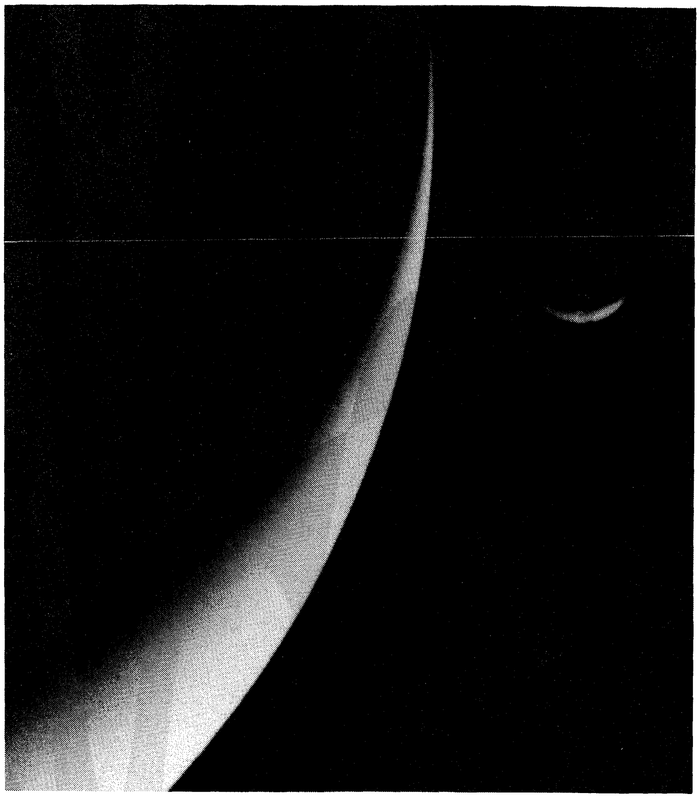


Figure 40: The crescents of Neptune and its largest satellite, Triton (background), photographed from the Voyager 2 space probe in August 1989. The image was taken at a distance of some 4.86 million kilometres from the planet.

By courtesy of the National Aeronautics and Space Administration

Adams, at St. John's College, Cambridge, began serious work in 1843 and by September 1845 had communicated his results to James Challis, one of his colleagues, and the following month to the Astronomer Royal, Sir George Biddell Airy. Adams said that his calculations would account for the irregularities in the motion of Uranus by the action of an exterior planet whose orbital elements were tabulated. The position of the exterior planet, as derived from Adams' work, was later found to be so close to the true value that two evenings of assiduous searching should have been sufficient to locate it. Adams also demonstrated that the observations of Uranus from 1780 onward were in good agreement with his theory. Instead of looking for the unknown planet, however, Airy did no more than to ask, in a reply, whether the assumed perturbations would also explain the known error in the calculated distance of Uranus from the Sun.

Meanwhile Le Verrier in France, working on similar lines, had published the results of his thorough mathematical investigations in a communication to the French Academy of Sciences on November 10, 1845. A second memoir was presented by Le Verrier on June 1, 1846.

The close agreement between the elements found by the two investigators is clear from the tabulated values (see

Table 20: Physical Data for Neptune

Mean distance from Sun	4,500,000,000 km (30.1 a.u.)
Eccentricity of orbit	0.009
Inclination of orbit to ecliptic	1°46′34″
Sidereal period of revolution	60,190 days, or 164.8 years
Rotation period	16 hours 3 minutes
Albedo	0.84
Mean synodic period	367.49 days
Mean orbital velocity	5.44 km/sec
Mean daily motion	0° .005981
Inclination of equator to ecliptic	28°8′
Mass (Earth = 1)	17.2
Diameter	49,530 km
Density	1.64 gm/cc
Apparent magnitude	+7.8
Satellites	8

Table 21) and led Airy to suggest to Challis, on July 9, 1846, that a search for the unknown planet be made with the Northumberland telescope, then at Cambridge Observatory. Challis commenced on July 29, sweeping each region twice, so that a comparison of the two sets of observations would then reveal the unknown planet by its motion. It subsequently transpired that Challis observed the planet on August 4 but failed to compare his observations with those of the previous night.

On August 31, Le Verrier's final paper was presented to the French Academy, and on September 18 he wrote to the German astronomer Johann Gottfried Galle at Berlin suggesting that he search for the planet, with the hope of distinguishing it from a star by its disklike appearance. On September 23, the planet was located on the first attempt by Galle and another German astronomer, Heinrich Louis d'Arrest, not by its disklike appearance but rather by the absence of an observed starlike body from one of the new star charts prepared at the observatory.

Table 21: Orbital Data According to Le Verrier and Adams

element	Le Verrier	Adams	
		hypothesis I	hypothesis II
Semimajor axis (a.u.)	36.154	38.38	37.27
Eccentricity	0.1076	0.16103	0.12062
Longitude of perihelion	284°45′	315°57′	299°11′
Mean longitude	318°47′	325°08′	323°02′
Epoch	1847, Jan. 1	1846, Oct. 1	1846, Oct. 1
True longitude	326°32′	328°	329°

Controversy over the credit for Neptune's discovery

The question whether Le Verrier should receive sole credit was warmly discussed by French and British astronomers; Adams and Le Verrier took no part in the controversy. The French physicist D.-F.-J. Arago went so far as to suggest that the new planet be named Le Verrier, but this was received with such little favour outside France that he withdrew it, proposing Neptune instead.

The mathematical methods employed in the solution of the problem and, in particular, the choice of the value of the semimajor axis of the orbit have been questioned. At the date of its discovery, Neptune was only 1° from the place predicted by Le Verrier and approximately 2½° from the position predicted by Adams. However, it has since been shown that it was fortunate that the search was made around the middle of the 19th century, because, if the orbit worked out by Adams had been used for a search in 1770, then Neptune would have been nearly 30° from its predicted position and not much less if Le Verrier's orbit had been used. Figure 41 shows just how sharply the actual and predicted positions would have diverged after 1850–60.

BASIC ASTRONOMICAL DATA

Appearance. The great distance and relative faintness of Neptune make it exceedingly difficult for any details to be observed from the Earth, even with large telescopes. It was not until the U.S. Voyager 2 probe passed within 5,000 kilometres of Neptune's cloud tops on August 25, 1989, that scientists were able to study the physical features of the planet.

Neptune is seen by light reflected from the Sun, and its greenish blue hue is due to absorption of the red, orange, and yellow solar rays by its extensive atmosphere. Strong absorption bands in the infrared region of the spectrum had been identified in 1932 as being due to methane (CH_4); some unidentified bands at a wavelength of about 7500 angstroms, near the deep red end of the visible spectrum, were found by the U.S. astronomer G.P. Kuiper in 1949.

Atmosphere. Neptune's atmosphere consists of hydrogen and helium mixed with methane. Images and data transmitted by Voyager 2 revealed that winds of about 300 metres per second (700 miles per hour) blow westward around the planet. Several enormous storm systems were also detected by the spacecraft. The largest of these systems, the so-called Great Dark Spot (or GDS), is as wide as the Earth and resembles Jupiter's Great Red Spot. Located about 20° latitude, this whirlwind turns in a counterclockwise direction. Wispy clouds of methane hover over the southern rim of the Great Dark Spot some 50 kilometres above a layer of hazy methane stratus clouds that envelops much of Neptune. A second giant storm system shaped somewhat like an eye appears to the south of the Great Dark Spot.

Great Dark Spot

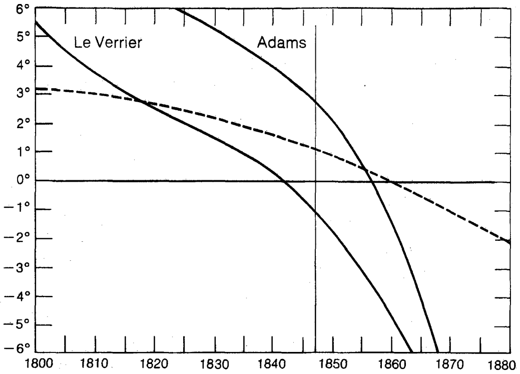


Figure 41: Curves of amounts by which orbits of Adams, Le Verrier, and circular-orbit solution (dashed curve) differ from true position of Neptune. The vertical marks the date of discovery, 1846 (see text).

From R.A. Lyttleton, *Mysteries of the Solar System* (1968); Clarendon Press, Oxford

Magnetic field. Voyager 2 discovered that Neptune has a complex magnetic field. While the field is not as strong as those surrounding the other giant outer planets and Earth, it is still capable of trapping galactic cosmic-ray and solar wind particles—*e.g.*, energetic protons and electrons—in a roughly doughnut-shaped region termed a magnetosphere. The magnetic field rotates with the planet, and, as it does so, accelerates the electrons to high speeds approaching that of light. This causes them to emit radio energy, which is released in beams whose intensity varies regularly with the field's (and hence the planet's) period of rotation.

Voyager 2 detected auroras in Neptune's atmosphere. These luminous phenomena are associated with the charged solar wind particles captured by a planet's magnetic field. The auroral displays on Neptune are spread over a fairly wide region rather than being confined around the magnetic poles, as on the Earth.

The dipole axis of Neptune's magnetic field is tilted 47° from the rotation axis and is displaced from the planetary centre by some 10,000 kilometres. With the exception of Uranus' magnetic field, which is similarly tilted, those of the other planets generally coincide with their rotation axis. (Neptune's axis of rotation is tilted roughly 29° to the plane of the planet's orbit around the Sun.) The tilted magnetic fields of Neptune and Uranus have led some investigators to speculate that the fields of both planets result from the convection of electrically conducting fluids near their surface rather than within a molten metallic core, as in the case of the Earth. In effect, the electric dynamo of Neptune and Uranus operates around their core instead of in it.

Rotation. Spectroscopic observation of Neptune in 1928 led astronomers to believe that the planet's rotation period was about 15 hours. Although this value was subsequently refined, investigators remained uncertain as to its preciseness. Voyager 2's measurements of radio emissions from Neptune's magnetic field enabled them to determine that the planet actually completes one orbit in 16 hours and 3 minutes.

Composition, diameter, and density. Like the other giant outer planets, Neptune has no solid surface. It consists

Table 22: Elements of Neptune's Orbit

Mean anomaly	133° 44′ 09″.783
Perihelion	270° 03′ 30″.833
Ascending node	131° 16′ 41″.893
Inclination	1° 46′ 33″.651
Eccentricity	0.009
Semimajor axis	29.9871290465

chiefly of water and molten rock, resembling its neighbour Uranus.

Neptune has an equatorial diameter of about 49,530 kilometres. Analysis of Voyager 2 data suggests that Neptune is the densest of all the giant planets, having an estimated density of 1.64 grams per cubic centimetre.

SATELLITES AND RINGS

Satellites. It had long been known that Neptune had two satellites, Triton and Nereid. Triton, the larger and brighter of the two, was discovered visually by W. Lassell, using a 60-centimetre telescope at Liverpool, England, in 1846, within a month of the discovery of the planet itself. It was originally assumed that Triton revolved around Neptune in the same direction as the planet's rotation, implying that Neptune's north pole pointed south of the ecliptic, like that of Uranus (in those days no retrograde satellites were known). In 1928, however, careful spectroscopic work at Lick Observatory, California, showed that Neptune rotated in the direct sense unlike Uranus; thus Triton moved around Neptune in the retrograde direction, opposite to that of the planet's spin. The orbit of Triton is circular and was seen edge on from the Earth in 1953. Nereid was discovered photographically in 1949 by G.P. Kuiper, with the aid of the 208-centimetre McDonald telescope in Texas.

Triton's retrograde orbit

Voyager 2 provided a wealth of information about Triton and sighted six additional satellites. Triton proved to be smaller than initially thought: its diameter is only 2,705 kilometres, about 1,000 kilometres less than originally estimated using ground-based observations. Triton is composed of ice and rock and appears to have a very thin atmosphere of nitrogen and methane. It may be the coldest object in the entire solar system, having a surface temperature of −240° C or less.

Triton has an unusual landscape, one that has been likened to the rind of a cantaloupe. High-resolution images transmitted by Voyager 2 cameras showed that its equatorial zone has enormous canyons and narrow ridges that cut across one another along various segments. There are vast expanses of ice flows, presumably frozen water mixed with methane and ammonia. Some of these are pitted by meteorite craters reminiscent of those on the Moon. Triton's southern polar cap is covered by nitrogen frost, over which numerous feathery, dark streaks can be seen drifting northward. It has been proposed that these streaks, along with various features of Triton's terrain, are manifestations of a phenomenon referred to as "icy volcanism." According to this theory, seasonal heating of the surface causes some of the thick frost cover to evaporate and break up. The decrease in the volume of icy matter results in a reduction in the pressure exerted on subsurface reservoirs of nitrogen and smaller amounts of methane. When the confining pressure drops to a certain point, the liquid, heated by some internal heat source, vaporizes and discharges through fissures in the overlying ice layer. Exposure to the below-freezing air causes the gases to condense to ice crystals. Some of these crystals, primarily of methane, are carried by winds blowing across the polar caps, thereby producing the dark streaks. The long thin ridges mentioned above probably formed where the liquid nitrogen and methane oozed onto the surface from cracks in the surface and quickly froze. The flows of ice on the broad plains, which resemble volcanic lava flows, are thought to have formed in a similar way. In this case, however, ices composed of a mixture of ordinary water, methane, and most likely ammonia may have poured out from the interior of Triton early in its history and spread across wide areas. It is explained that mixed ices can flow at substantially lower temperatures than can ice consisting of pure water.

The other satellites of Neptune are much smaller than Triton. One of the newly discovered moons, simply designated 1989 N1, is the second largest, measuring about 400 kilometres in diameter, while Nereid, at 340 kilometres, is third. The remaining moons range from about 50 to 200 kilometres in diameter (see Table 23). Voyager 2 images of 1989 N1 indicate that it is irregularly shaped and that its surface is heavily pocked with impact craters. The shapes and surfaces of the other Neptunian moons are similar.

Table 23: Satellite Data

satellite name or designation	distance from centre of Neptune (in kilometres)	diameter (in kilometres)
Nereid	5,513,400	340
Triton	354,800	2,705
1989 N1	117,600	400
1989 N2	73,600	190
1989 N4	62,000	150
1989 N3	52,500	180
1989 N5	50,000	80
1989 N6	48,000	54

Rings. Neptune has a system of four rings encircling it. The rings are composed of dust particles, most likely the debris ejected from the satellites during meteorite impacts. The outermost ring has a dense core with three segments consisting of numerous small "moonlets" (no more than 10 kilometres in diameter). It is surrounded by a diffuse dust halo that extends about 50 kilometres wide. The ring is orbited by one of the smaller satellites, 1989 N4, which is thought to confine the ring's constituent particles. The third ring inward toward Neptune also has a so-called shepherd satellite, the moon designated as 1989 N3. The other two rings of Neptune are very diffuse and broad.

Rings of dust particles

(G.E.T./Ed.)

Pluto

Pluto, symbol ♇ in astronomy, is the ninth and most distant of the known planets in the solar system. The principal orbital and physical data of Pluto and its satellite, Charon, are shown in Table 24.

DISCOVERY OF PLUTO AND CHARON

Before Pluto's discovery in 1930, several astronomers had interpreted the motions of Uranus and Neptune as revealing gravitational perturbations beyond those accountable for by all of the planets known at that time. Their calculations were similar to those that led John Couch Adams and Urbain-Jean-Joseph Le Verrier in the 1840s to predict correctly the position of Neptune. Also among those who thus predicted a planet beyond Neptune was Percival Lowell, to whom credit for a prediction was generally given. Not only were his predicted orbital elements close to those of Pluto but he also initiated the effort that led ultimately to its discovery by his successors at the Lowell Observatory in Flagstaff, Arizona. Although Lowell's *Memoir on a Trans-Neptunian Planet* (*Memoirs of the Lowell Observatory*, vol. 1) was not published until 1915, the year before his death, he and his colleagues carried out intermittent searches between 1905 and 1916 with a variety of telescopes of limited capability. On the basis of experience gained while conducting these early efforts, a fast wide-field telescope was designed specifically for the trans-Neptunian planet search.

Percival Lowell's contributions

After the telescope was completed in 1929, the project was placed in the hands of Clyde William Tombaugh. He painstakingly compared pairs of photographic plates of selected sky regions, taken several days apart, looking for images of moving objects. The project was carefully planned so that, among the millions of stellar and starlike images, other moving bodies such as the numerous asteroids could be easily recognized and not be incorrectly identified as the sought-for planet. On February 18, 1930, Tombaugh discovered the trans-Neptunian planet on plates that he had taken on January 21, 23, and 29, 1930. Announcement of the discovery was not made until March 13, 1930, the anniversary of Lowell's birth, but in the interval the Lowell Observatory astronomers were able to assemble further supporting evidence to confirm the discovery. After long deliberation over many suggested names for the new planet, the name Pluto was chosen because it could be represented by the symbol ♇, made up of Percival Lowell's initials.

Even while the announcement of this significant discovery was drawing the attention of the entire astronomical world to the new planet, certain grave doubts were being expressed as to whether it indeed was Lowell's trans-Neptunian planet. The difficulty was that Pluto was much fainter—about one-tenth as bright—than had been expected, and its telescopic image had no measurable diameter. It was therefore difficult to resist the conclusion that Pluto was altogether too small and, hence, of insufficient mass to have caused the large perturbations in Uranus' and Neptune's motions that had led Lowell and others to predict its existence and orbital motion. Soon after Pluto's discovery, Ernest William Brown demonstrated mathematically that Lowell's prediction was, in fact, dynamically invalid. Then, while he was associated with the Lick Observatory, Mt. Hamilton, California (1932–36), E.C. Bower, among others, convincingly demonstrated that Pluto's mass was probably less than one-half that of the Earth and that the resulting perturbations to the motions of Uranus and Neptune must have been significantly smaller than the probable errors of the earliest observations of these bodies. Modern data have since confirmed this conclusion. Thus, Pluto, at first considered as the second planet, after Neptune, to have been discovered through prediction, is actually the second planet, after Uranus, to have been discovered by accident. Or perhaps more properly, it is the first planet to have been discovered through a systematic, comprehensive search.

In 1978 the existence of a satellite of Pluto was established by James W. Christy and Robert S. Harrington of the U.S. Naval Observatory, who discovered that photographic images of Pluto periodically appeared slightly elongated. The satellite, named Charon, orbits the planet with a period of 6.39 days, in synchronism with the known period of variation in Pluto's brightness. Charon's orbit is probably circular, with a radius of about 20,000 kilometres.

BASIC ASTRONOMICAL DATA

Pluto's orbit. Pluto moves in a markedly eccentric orbit about the Sun with a period of about 248 years. The orbit is more eccentric than that of any other planet and has a very high inclination, about 17°, to the mean plane of the solar system. Because of the high eccentricity of the orbit, Pluto is closer to the Sun than is Neptune at times around perihelion as, for example, in the interval 1979–99.

Conjunctions with Neptune and Uranus

The near commensurability of the periods of revolution of the three outermost planets brings the planets into successive conjunctions at about the same places in space. At intervals of approximately 500 years, during which Pluto makes two revolutions and Neptune three, these two planets come closest to each other and are then separated by only 18 astronomical units; this phenomenon last occurred in 1891 and will occur again in 2391. Similarly, Pluto and Uranus have close conjunctions, although in two series. For example, in the years 1854, 2107, and 2360, Pluto and Uranus are approximately 29 a.u. apart, and in the years 1713, 1966, and 2220 they are only 14 a.u. apart.

Mass, diameter, and density. The difficult problem of determining the mass of Pluto was resolved with the discovery of Pluto's satellite and the determination of its orbital period around, and its mean distance from, the planet. Before the discovery of Charon, estimates of Pluto's mass ranged from about $^1/_5$ that of Earth to seven times that of Earth, with indeterminate uncertainties. Pluto's mass is now believed to be only about $^1/_{500}$ that of the Earth, with an uncertainty of 20 percent. While these later figures are subject to improvement as refined studies progress, there is no doubt that they are more reliable than those made before the discovery of Charon.

Following upon the determination of the combined mass of Pluto and Charon, the chief problem is the measurement of their diameters. This is not to say that other unknowns—such as their chemical compositions, the physical nature of their surfaces, or their origins—are less interesting or important. It is simply an acknowledgment that, without reliable determinations of Pluto's and Charon's diameters and mean densities, many of these other problems will remain obscured and their answers speculative.

Pluto's diameter

The diameter of Pluto has been the subject of several investigations, but it was not determined with confidence until the late 1980s. At that time, several groups of astronomers, using the most advanced telescopes and detectors, determined that the planet measures from about 2,200 to 2,300 kilometres across. In effect, Pluto is much smaller than the Moon.

Table 24: Orbital and Physical Data for Pluto

Mean distance from Sun	39.53 a.u.
Least distance from Sun	29.81 a.u.
Greatest distance from Sun	49.26 a.u.
Eccentricity	0.246
Inclination tò ecliptic	17.123°
Period of revolution	248.5 years
Time of perihelion	1989.8
Visual magnitude at mean opposition	15.0
Period of axial rotation	6.3867 days
Satellites	one known; 1978 P1 (Charon)
Diameter (estimated)	Pluto: 2,200–2,300 kilometres Charon: about 1,200 kilometres
Mass	Pluto and Charon together: 0.0025 terrestrial mass Pluto probably about 10 times as massive as Charon
Density	probably between 0.4 and 0.9 gram per cubic centimetre
Composition	uncertain, may consist of a large proportion of frozen methane
Inclination of Charon's Orbital plane with respect to Pluto's orbital plane around the Sun	111°

A secure measurement of the diameters of celestial bodies can be made if the bodies occult a star in the course of their apparent motion across the sky and if the occultation is observed from two observatories located at different latitudes on the Earth. In 1980, such an occultation occurred in the Pluto–Charon system, but it was evidently observed only at the South African Astronomical Observatory. The occulting body apparently was Charon, and the results show that Charon is at least 1,200 kilometres in diameter. In 1965, numerous North American observatories had been alerted to a possible occultation by Pluto–Charon, but no such occultation was observed.

On the basis of the data given above for current estimates of the masses and diameters of Pluto and Charon, it appears that the mean density of Pluto and Charon lies between 0.4 and 0.9 grams per cubic centimetre. This low value of density has led some investigators to propose that these bodies may be made up of a large proportion of solid methane, instead of the denser, rocky materials found in the terrestrial planets.

THE SURFACES OF THE PLANET AND ITS SATELLITE

Because the telescopic image of Pluto is generally indistinguishable from that of a faint star, all efforts made by the early 1980s to detect surface features by conventional methods were unsuccessful. That the surface is not uniform has been known since 1955, when M.F. Walker and R.H. Hardie at the Lowell Observatory found that the brightness of the planet varied regularly by about 12 percent within a period of 6.39 days. The variation in apparent brightness is now interpreted as due to the combined effects of irregular surface reflection properties and an axial rotation of the planet with a period of 6.39 days. Studies since 1955 have shown that the range of brightness variation has gradually increased; analysis of this increase has in turn allowed for an estimation of the orientation of Pluto's spin axis in space.

Variations in apparent brightness

Information about the surfaces of both Pluto and Charon remains limited. Observations of the Infrared Astronomical Satellite (IRAS) reveal that Pluto has a dark band along its equator and polar caps of frozen methane, which vary over time. The infrared spectrum of Charon, on the other hand, indicates that water ice rather than methane frost covers its surface.

POSSIBLE ATMOSPHERE

Whether Pluto has an atmosphere is not known with any degree of assurance, but certain conditions relating to a possible atmosphere can be stated in general terms. It can be estimated that the average surface temperature on the sunlit portion of Pluto's surface is about 60 K when the planet is at perihelion and 47 K at aphelion. The dark side of the planet probably has a temperature of about 20 K. At such temperatures, virtually all materials are frozen, or at least liquefied, except for a few gases such as hydrogen and helium. These gases are so light in molecular weight, and Pluto's gravity is so low, that it is unlikely that they could be retained in free form. There is some possibility, however, that frozen methane might sublimate slightly on the sunlit side of Pluto to produce an atmosphere of low density. A clearer picture of the atmospheric and surface conditions of Pluto awaits improvements in the resolution of spectrophotometric techniques and further exploration by spacecraft.

COMPARISON WITH OTHER PLANETS

Pluto differs greatly from the major planets in its general physical properties—the other outer planets have great mass, large size, extensive atmosphere, and very rapid axial rotation—and has much higher orbital eccentricity and inclination. Thus, Pluto probably had an origin and evolution different from that of the major planets. Because Pluto's physical properties appear more akin to those of the satellites, especially Triton, and because Pluto's orbit nearly intersects Neptune's, it has been proposed by R.A. Lyttleton and others that Pluto may in fact be a runaway satellite originally bound to Neptune.

Similarity to satellites

Whether there is any validity to the apparent perturbations shown in the motions of Uranus and Neptune that led to the prediction and ultimate discovery of Pluto remains open to question. There is little doubt that many of the earliest observed positions of these planets were adversely affected by errors due to primitive techniques. In recent years, however, the entire subject was again under review with renewed vigour and the advantage of an enormously wider range of computational techniques than was available to the investigators of the early 1900s. It is not inconceivable that current studies may point to the existence of another planet that is as yet undiscovered.

(R.H.H./Ed.)

OTHER CONSTITUENTS OF THE SOLAR SYSTEM

Asteroids

Asteroids are rocky bodies about 1,000 kilometres or less in diameter that orbit the Sun primarily between the orbits of Mars and Jupiter. Because of their small size and large numbers relative to the nine major planets, asteroids are also called minor planets. The two designations are frequently used interchangeably, though dynamicists, astronomers who study individual objects with dynamically interesting orbits or groups of objects with similar orbital characteristics, generally use the term minor planet, whereas those who study the physical properties of such objects usually refer to them as asteroids. The term planetoid is sometimes used as well.

Nomenclature

HISTORICAL SURVEY OF MAJOR ASTEROID DISCOVERIES

Early observations. The first asteroid was discovered on Jan. 1, 1801, by Giuseppe Piazzi at Palermo. At first Piazzi thought that he had discovered a comet; however, after the orbital elements of the object had been computed it became clear that the object moved in a planetlike orbit between the orbits of Mars and Jupiter. Due to illness, Piazzi was only able to observe the object until February 11, and, as no one else was aware of its existence, it was not reobserved before it moved into the daytime sky. The short arc of observations did not allow computation of an orbit of sufficient accuracy to predict where the object would reappear when it moved back into the night sky, and so it was "lost." There matters might have stood were it not for the fact that this object was located at the heliocentric distance predicted by the Titius-Bode law (also known as Bode's law) of planetary distances proposed in 1766 by the German astronomer Johann D. Titius and popularized by his compatriot Johann E. Bode, who used the scheme to advance the notion of a "missing" planet between Mars and Jupiter. The discovery of Uranus in 1781 by the German-born British astronomer William Herschel at a distance that closely fit the distance predicted by the Titius-Bode law was taken as strong evidence of its correctness. Some astronomers were so convinced that during an astronomical conference in 1796 they agreed to undertake a systematic search. Ironically, Piazzi was not a party to this attempt to locate the missing planet. Nonetheless, Bode and others, on the basis of the preliminary orbit, believed that Piazzi had found and then lost it. This led the German mathematician Carl Friedrich Gauss to develop in 1801 a method for computing the orbit of an asteroid from only a few observations, a technique that has not been significantly improved since. Using Gauss's predictions, the German astronomer Franz von Zach recovered Ceres on Jan. 1, 1802. Piazzi named this object Ceres after the ancient Roman grain goddess and patron goddess of Sicily, thereby initiating a tradition that continues to the present day: asteroids are named by their discoverers (in contrast to comets, which are named for their discoverers).

Identification of Ceres

The discovery of three more faint (compared with Mars and Jupiter) objects in similar orbits over the next six years (Pallas, Juno, and Vesta, respectively) complicated this elegant solution to the missing-planet problem and gave rise to the surprisingly long-lived, though no longer generally accepted, idea that the asteroids were remnants of a planet that had exploded.

Following this flurry of activity, the search for the planet appears to have been abandoned until 1830 when Karl L. Hencke renewed it. In 1845 he discovered the fifth asteroid, which he named Astraea.

Modern research. There were 88 known asteroids by 1866, when the next major discovery was made: Daniel Kirkwood, a U.S. astronomer, noted that there were gaps in the distribution of asteroid distances from the Sun (see below *Distribution and Kirkwood gaps*). The introduction of photography to the search for new asteroids by the German astronomer Max Wolf in 1891, by which time 322 asteroids had been identified, accelerated the discovery rate. By the end of the 19th century, 464 had been found. The asteroid designated 323 Brucia, detected by Wolf in 1891, was the first to be discovered by means of photography.

Application of photography

The first measurements of the sizes of asteroids were made in 1894 and 1895 by the U.S. astronomer Edward E. Barnard, who used a filar micrometer (an instrument normally employed for visual measurement of the separations of double stars) to estimate the diameters of the first four asteroids. Barnard's results established that Ceres was the largest asteroid, with an estimated diameter of nearly 800 kilometres. These values remained the best available until new techniques were introduced during the early 1970s (see below *Size and albedo*). The first four asteroids came to be known as "the big four," and, because all other asteroids were much fainter, they were believed to be considerably smaller as well.

In 1918 the Japanese astronomer Kiyotsugu Hirayama recognized clustering in three of the orbital elements of various asteroids (semimajor axis, eccentricity, and inclination). He speculated that objects sharing these elements had been formed by explosions of larger parent asteroids and called such groups of asteroids "families."

The idea that Jupiter was responsible for interrupting the formation of a planet from the swarm of planetesimals accreting near a heliocentric distance of 2.8 a.u. was introduced in 1944 by O.J. Schmidt. In 1951 the Estonian astronomer Ernest J. Öpik calculated the lifetimes of asteroids with orbits that passed close to those of the major planets and showed that most such asteroids were destined to collide with a planet or be ejected from the solar system on time scales of a few hundred thousand to a few million years. Since the age of the solar system is approximately 4,500,000,000 years, this meant that the asteroids seen today in such orbits must have entered them recently and implied a source for the asteroids. Öpik believed this source to be comets that had been captured by the planets and that had lost their volatile material through repeated passages inside the orbit of Mars.

The mass of Vesta was deduced by the German-born U.S. astronomer Hans G. Hertz in 1966 from measurements of its perturbations on the orbit of 197 Arete. The first mineralogical determination of the surface composition of an asteroid was made in 1969 by Thomas McCord, John B. Adams, and Torrence V. Johnson of the United States, who used a technique known as spectrophotometry to identify the mineral pyroxene in the surface material of 4 Vesta. In 1970 the first reliable albedos (reflectivities) and diameters of asteroids were determined by two groups of U.S. astronomers—Joseph F. Veverka, Benjamin H. Zellner, and their colleagues, who used a technique based on polarization measurements, and David A. Allen and Dennis L. Matson, who employed infrared radiometry.

Development of classification systems. In 1975 Zellner, together with Clark R. Chapman and David D. Morrison, grouped the asteroids into three broad taxonomic classes, which they designated C, S, and M (see below *Composition*). They estimated that about 75 percent belonged to class C, 15 percent to class S, and 5 percent to class M. The remaining 5 percent were unclassifiable in their system due to either poor data or genuinely unusual properties. Furthermore, they noted that the S class dominated the population at the inner edge of the asteroid belt, whereas the C class was dominant in the middle and outer regions of the belt. In 1982 other U.S. astronomers, Jonathan C. Gradie and Edward F. Tedesco, expanded this taxonomic system and recognized that the asteroid belt consisted of rings of differing taxonomic classes with the S, C, P, and D classes dominating the populations at heliocentric distances of approximately 2, 3, 4, and 5 a.u., respectively (see below *Composition;* see also Figure 42).

ORBITS OF ASTEROIDS

Asteroid designations

Because of their large numbers (about 3,500 identified as of 1987), asteroids are assigned numbers as well as names. The numbers are assigned consecutively after accurate orbital elements have been determined. (For example, Ceres is officially known as 1 Ceres, Pallas as 2 Pallas, and so forth.) The discoverers have the right to choose a name for their discoveries as soon as they are numbered. Nowadays the names selected are submitted to the International Astronomical Union for approval.

Prior to the mid-20th century, asteroids were sometimes assigned numbers before accurate orbital elements had been determined, and so some numbered asteroids cannot be located now. These objects are referred to as "lost" asteroids. As of 1987, asteroids 719 Albert, 724 Hapag, and 878 Mildred were still lost.

The Minor Planet Center at the Harvard-Smithsonian Center for Astrophysics in Cambridge, Mass., maintains computer files for all measurements of asteroid positions. The Institute for Theoretical Astronomy in Leningrad publishes each year the *Ephemerides of Minor Planets,* which contains the orbital elements of all numbered asteroids, together with their opposition dates and ephemerides.

Deviations from circular orbits

Although most asteroids travel in fairly circular orbits, there are some notable exceptions. One of the most extreme of these is 3200 Phaethon, discovered by the Infrared Astronomical Satellite (IRAS) in 1983. (It was the first asteroid to be discovered by a spacecraft.) Phaethon approaches to within 0.14 a.u. of the Sun, well within Mercury's perihelion distance of 0.31 a.u. Phaethon's aphelion (2.4 a.u.) is in the main asteroid belt. This object is the parent body of the Geminid meteor stream and, since the parent bodies of all other meteor streams identified to date are comets, it is considered by some to be a defunct comet. Another asteroid, 944 Hidalgo, is also thought by some to be a defunct comet because of its unusual orbit. This object, discovered in 1920 by Walter Baade at the Bergedorf Observatory near Hamburg, has a perihelion distance of 2.02 a.u. at the inner edge of the main asteroid belt and an aphelion distance of 9.68 a.u. just beyond the orbit of Saturn at 9.54 a.u. Finally, there is the case of 2060 Chiron, discovered in 1977 by Charles Kowal at the Palomar Observatory near San Diego, Calif. This object is classified as an asteroid but is considered by some to be an inactive cometary nucleus. It travels in an orbit that lies wholly outside of the asteroid belt, having a perihelion distance of 8.43 a.u. (between the orbits of Jupiter and Saturn) and an aphelion distance of 18.8 a.u., which nearly reaches the orbit of Uranus at 19.2 a.u. Chiron is the most distant body in the solar system besides some comets that moves in what is known as a chaotic orbit. Astronomers believe that it will eventually collide with a planet or be permanently ejected from the solar system (see also below *Groups of comets and other unusual cometary objects*).

Cause of the Kirkwood gaps

Distribution and Kirkwood gaps. About 95 percent of the known asteroids move in orbits between those of Mars and Jupiter. These orbits, however, are not uniformly distributed but, as shown in Figure 42, exhibit "gaps" in the distribution of their semimajor axes. These so-called Kirkwood gaps are due to resonances with Jupiter's orbital period. An asteroid with a semimajor axis of 3.3 a.u., for example, makes two circuits around the Sun in the time it takes Jupiter to make one and is thus said to be in a one-to-two (written 1/2) resonance orbit with Jupiter. Consequently, once every two orbits Jupiter and an asteroid in such an orbit would be in the same relative positions, and such an asteroid would experience a force in a fixed direction. Repeated applications of this force would eventually change the semimajor axes of asteroids in such orbits, thus creating a gap at that distance. Gaps occur at 1/4, 2/7, 1/3, 2/5, 3/7, and 1/2 resonances, while concentrations occur at the 2/3 (Hilda group), 3/4 (Thule), and 1/1 (Trojan group) resonances. An adequate explanation of why some resonances produce gaps and others produce concentrations has yet to be found.

Families. Within the main asteroid belt are groups of asteroids that cluster in certain orbital elements (semimajor axis, eccentricity, and inclination). Such groups are called

From J. Gradie and E. Tedesco, *Science*, vol. 216, p. 1406 (June 25, 1982); © American Association for the Advancement of Science

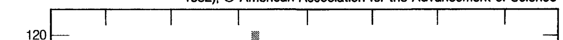


Figure 42: (Top) Number distribution (*N*) in semimajor axis (*a*) for asteroids with diameters greater than about 50 kilometres. The fractions below the centre indicate the locations of the major orbital period resonances with Jupiter. The deep Kirkwood gaps occur within the main belt near the 1/3, 2/5, and 3/7 resonances, whereas the concentrations are found at the 2/3, 3/4, and 1/1 resonances. The main belt is confined to the region between the 1/4 and 1/2 resonances. The Kirkwood gap at the 2/7 resonance is too narrow to be displayed at the scale plotted. (Bottom) Fractional number distribution (*f*) of the major asteroid classes from Table 25. The sum of the various classes found at any given semimajor axis must equal 1. The F and R classes do not appear because they do not constitute more than 2 percent of the total number of asteroids found at any given semimajor axis.

families and assigned the name of the lowest numbered asteroid in the family. Asteroid families are thought to be formed when an asteroid is disrupted in a catastrophic collision. Theoretical studies indicate that such catastrophic collisions between asteroids are common enough to account for the number of families observed. About 40 percent of all known asteroids belong to such families.

The three largest families (Eos, Koronis, and Themis) have been determined to be compositionally homogeneous. If the asteroids belonging to them are considered to be fragments of a single parent body, then their parent bodies must have had diameters of 200, 90, and 300 kilometres, respectively. The smaller families have not been as well studied because their numbered members are fewer and smaller (and hence fainter). Nevertheless, it is known that some of the smaller families are compositionally inhomogeneous and that, at least in some cases, what are observed are pieces of a geochemically differentiated parent body. It is theorized that some of the Earth-crossing asteroids and meteorites reaching the terrestrial surface are fragments produced in collisions similar to those that produced the asteroid families.

Outer-belt asteroids. The vast majority of asteroids have orbital periods between three years and six years—*i.e.,* between one-fourth and one-half of Jupiter's orbital period (see Figure 42). These asteroids are said to be main-belt asteroids. Besides the few asteroids with unique orbits, some of which were noted above, there are a number of groups that fall outside the main belt. Those that have orbital periods greater than one-half that of Jupiter are called outer-belt asteroids. There are four such groups: the Cybeles, Hildas, and Thule, named after the lowest numbered asteroid in each group, as well as the Trojan group, so called because all of its members are named after characters from Homer's epic work on the Trojan War, the *Iliad.* There are about 65 known Cybeles asteroids, 40 known Hildas, one known Thule, and 45 known Trojans.

Asteroid groups that fall outside the main belt

Trojan asteroids. In 1772 the French mathematician and astronomer Joseph-Louis Lagrange predicted the existence and location of two groups of asteroids located near the (L_4 and L_5) equilateral triangular stability points of a three-body system formed by the Sun, Jupiter, and asteroid. These are two of the five stable points in the circular, restricted three-body problem. (The other three stable points are located along a line passing through the Sun and Jupiter. Because of the presence of other planets, principally Saturn, the Sun–Jupiter–Trojan asteroid system is not a true three-body system, and so these other three points are not stable and no asteroids have been found near them. In fact, most of Jupiter's Trojan asteroids do not move in the plane of its orbit but rather in orbits inclined by up to 25° and at longitudes that differ by as much as 40° from the longitudes of the true Lagrangian points.)

In 1906 Max Wolf discovered 588 Achilles near the Lagrangian point following Jupiter in its orbit. Within a year August Kopff had discovered two more: 617 Patroclus, located near the following Lagrangian point, and 624 Hector near the preceding Lagrangian point. It was decided to name these asteroids after the participants in the Trojan War as given in the *Iliad* and, furthermore, to name those near the preceding point after Greek warriors and those near the following point after Trojan warriors. With the exception of two "spies" (Hector, the lone Trojan in the Greek camp, and Patroclus, the lone Greek in the Trojan camp), this tradition has been maintained.

The term Trojan has since been applied to any object occupying the equilateral Lagrangian points of other pairs of bodies. Searches have been made for Trojans of the Earth, Saturn, and Neptune, as well as of the Earth–Moon system, but so far none have been found. It is doubtful whether truly stable orbits could exist near these Lagrangian points because of perturbations by the major planets.

Although only about 45 Trojans have been numbered, photographic surveys have shown that there are about 900 such asteroids with diameters greater than 15 kilometres. Curiously, about 700 of these are located near the preceding Lagrangian point and only 200 near the following Lagrangian point. The reason for this disparity is unknown.

Inner-belt asteroids. There is only one known group of inner-belt asteroids—namely, the Hungarias. The Hungaria asteroids have a mean semimajor axis of 1.91 a.u.; thus, their orbital periods are less than one-fourth that of Jupiter (see Figure 42). Hungarias have nearly circular orbits (the mean eccentricity is 0.08) but large inclinations, the mean inclination being 22.9°. Two dynamic families (the Hungaria family and the Strackea family) exist within the Hungaria group. Because of the low eccentricities of their orbits, the mean perihelion distance of a Hungaria asteroid is 1.76 a.u. Accordingly, a typical Hungaria cannot pass close to Mars, whose aphelion distance is 1.67 a.u. A few Hungarias, however, have perihelion distances a few hundredths of 1 a.u. less than Mars's aphelion distance and so are shallow "Mars crossers" (see below) as well.

The Hungaria group

Near-Earth asteroids. Asteroids that can pass inside the orbit of Mars are said to be near-Earth asteroids. The near-Earth asteroids are subdivided into several classes. The most distant—those that can cross the orbit of Mars but that have perihelion distances (q) greater than 1.3 a.u.—are dubbed Mars crossers. This group is further subdivided into two groups: shallow Mars crossers ($1.58 \leq q < 1.67$ a.u.) and deep Mars crossers ($1.3 < q < 1.58$ a.u.).

Classes of near-Earth asteroids

The next most distant group of near-Earth asteroids are the Amors ($1.017 < q \leq 1.3$ a.u.). Amor asteroids have perihelion distances greater than the Earth's present aphelion distance (Q) of 1.017 a.u. and therefore do not at present cross the planet's orbit. Because of strong gravitational perturbations produced by their close approaches to the Earth, however, the orbital elements of all the Earth-approaching asteroids, except the shallow Mars crossers, change appreciably on a time scale of a few years or tens of years. For this reason about half the known Amors, including 1221 Amor, are part-time Earth crossers. Only asteroids that cross the orbits of planets, such as the Earth-approaching asteroids and objects like 944 Hidalgo and 2060 Chiron, suffer significant changes in their orbital elements on time scales shorter than many millions of years. Hence, the outer-belt asteroid groups (Cybeles, Hildas, Thule, and the Trojans) do not interchange members.

There are two groups of near-Earth asteroids that deeply cross the Earth's orbit on an almost continuous basis. The first of these to be discovered were the Apollo asteroids, 1862 Apollo being detected by the German astronomer Karl Wilhelm Reinmuth in 1932 but lost shortly thereafter and not rediscovered until 1978. Apollo asteroids have semimajor axes (a) that are greater than or equal to 1.0 a.u. and perihelion distances that are less than or equal to 1.017 a.u.; thus, they cross the Earth's orbit when near the perihelia of their orbits. For the other group of Earth-crossing asteroids, named Atens after 2062 Aten discovered in 1976 by Eleanor F. Helin of the United States, $a < 1.0$ a.u. and $Q \geq 0.983$ a.u., the present perihelion distance of the Earth. These asteroids cross the Earth's orbit when near the aphelia of their orbits.

As of 1986, the number of known Aten, Apollo, and Amor asteroids were six, 40, and 42, respectively. Of these 88 objects, 26 were discovered before 1961 (eight of which have since been lost), whereas 51 were discovered between 1975 and early 1986 (none of which have been lost). It is estimated that there are roughly 100 Atens, 700 Apollos, and 1,000 Amors that have diameters larger than about one kilometre.

THE NATURE OF ASTEROIDS

Rotation and shape. Asteroid rotation periods and shapes are determined primarily by monitoring their changing brightness on time scales of hours to days. Short-period fluctuations in brightness caused by the rotation of an irregularly shaped and/or spotted body (a spotted body being a spherical object with albedo differences) give rise to a light curve (a graph of brightness versus time) that repeats at regular intervals corresponding to an asteroid's rotation period. The range of brightness variation is more difficult to interpret but is closely related to an asteroid's shape and/or spottedness.

Rotation periods have been determined for more than 400 asteroids. They range from 2.3 hours to 48 days, but the majority (more than 80 percent) lie between four hours and 20 hours. Periods longer than a few days may actually be due to precession caused by an unseen satellite. The mean rotation period is roughly 10 hours for the entire sample. The largest asteroids (those with diameters greater than about 175 kilometres), however, have a mean rotation period close to seven hours, whereas this value is about 10 hours for smaller asteroids. The largest asteroids may have preserved their primordial rotation rates, but the smaller ones have almost certainly had theirs modified by subsequent collisions. The difference in rotation rates between the larger and the smaller asteroids is believed to stem from the fact that large asteroids retain all of their collisional debris from minor collisions, whereas smaller asteroids retain more of the debris ejected in the direction opposite to that of their spins, causing a loss of angular momentum and thus a reduction in speed of rotation.

Major collisions can completely disrupt smaller asteroids. The debris from such collisions comprise still smaller asteroids, which can have virtually any shape or spin rate. Thus, the fact that no rotation periods shorter than about 2.5 hours have been observed implies that the material of which asteroids are made is not strong enough to withstand the centripetal forces that such rapid spins would produce.

For mathematical reasons it is impossible to distinguish between the rotation of a spotted sphere and an irregular shape of uniform reflectivity on the basis of observed brightness changes alone. Nevertheless, the fact that opposite hemispheres of most asteroids appear to have albedos differing by no more than a few percent suggests that their brightness variations are due mainly to changes in their projected, illuminated, visible, cross-sectional areas. Hence, in the absence of evidence to the contrary, it is generally accepted that variations in reflectivity contribute little to the observed rotational light-curve amplitude. Asteroid 4 Vesta is a notable exception to this generalization because it is known that the difference in reflectivity between its opposite hemispheres is alone sufficient to account for its entire light-curve amplitude.

Asteroid light-curve amplitudes range from zero to a factor of 6.5 in the case of the Apollo asteroid 1620 Geographos. A light-curve amplitude of zero is caused by viewing an asteroid along one of its rotational poles, while the 6.5 to one variation in brightness is believed to result from either of two possibilities: Geographos is a cigar-shaped object that is viewed along a line perpendicular to its rotational axis, or it is a pair of objects nearly in contact that orbit each other around their centre of mass. The mean rotational amplitude for asteroids is about a factor of 1.3. These data, together with the assumptions mentioned above, allow astronomers to estimate asteroid shapes.

Asteroids exhibit a wide range of shapes (see Figure 43). Some of them, such as 1 Ceres, 2 Pallas, and 4 Vesta, are nearly spherical, whereas others, like 15 Eunomia, 107 Camilla, and 511 Davida, are quite elongated. Still others, as, for example, 624 Hektor and 1580 Betulia (not shown in Figure 43, but whose proposed shape is that of a kidney bean), may have bizarre shapes.

Application of thermal radiometry

Size and albedo. The most widely used technique for determining the sizes of asteroids is that of thermal radiometry. This technique makes use of the fact that the infrared radiation (heat) emitted by an asteroid must balance the solar radiation it absorbs. By using a so-called thermal model to balance the intensity of infrared radiation with the intensity at visual wavelengths, investigators are able to arrive at the diameter of the asteroid. Several other techniques—polarimetry, speckle interferometry, and radar—also are used, but they are limited to brighter, larger, and/or closer asteroids.

The only technique that measures the diameter directly (*i.e.,* without having to "model" the actual observations) is that of stellar occultation. In this method, investigators measure the length of time that a star disappears due to the passage of an asteroid between the Earth and the star. Then, using the known distance and the rate of motion of the asteroid, they are able to determine the latter's diameter uniquely. The results of this method have made it possible to reliably calibrate the indirect techniques, thermal radiometry in particular. As a consequence, asteroid diameters obtained by means of such techniques are now thought to have uncertainties of less than 10 percent. Because passages of asteroids in front of stars are rare and best applied to fairly spherical asteroids and because only one cross section is measured, the majority of asteroid sizes have been obtained using indirect techniques.

By the mid-1980s about 20 asteroids, including 1 Ceres, 2 Pallas, and 3 Juno, had had their diameters determined by means of the stellar-occultation method, whereas about 600 had had theirs measured with the indirect techniques, principally thermal radiometry.

Asteroid 1 Ceres, with a diameter of 940 kilometres, is the largest, followed by 2 Pallas at 535 km and 4 Vesta at 500 km. The fourth largest asteroid, 10 Hygiea, has a diameter of 430 km. There are two asteroids with diameters between 300 and 400 km and about 24 with diameters between 200 and 300 km. In sum, there are about 30 asteroids with diameters greater than 200 km. The smallest known asteroids are members of the Earth-crossing groups, since these asteroids can approach the Earth to within a few hundredths of 1 a.u. The smallest observed Earth-approaching asteroids measure less than 0.5 km across. It has been estimated that there are 250 asteroids larger than 100 km in diameter and perhaps 1,000,000 with diameters greater than 1 km.

Albedo

A parameter closely related to size is albedo, or reflectivity. This property also provides compositional information. Albedo is the ratio between the amount of light actually reflected and that which would be reflected by a uniformly scattering disk of the same size. Snow has an albedo of approximately 1.0 and coal an albedo of about 0.05.

An asteroid's albedo and diameter are closely related. For example, if 1 Ceres and 4 Vesta could both be observed at the same distance, Vesta would be the brighter of the two by about 10 percent. Vesta's diameter, however, is less than 55 percent that of Ceres. It appears brighter because its albedo is around 0.35, compared with 0.09 for Ceres.

The asteroids fall into three albedo groups: low (0.02–0.065), moderate (0.066–0.23), and high (greater than 0.23). About 80 percent of the known asteroids are low-albedo objects, and most of them are located in the outer half of the main asteroid belt and among the outer-belt populations. Roughly 19 percent of all known asteroids belong to the moderate-albedo group, the vast majority of which are found in the inner half of the main belt. The high-albedo asteroids make up the remaining 1 percent of the asteroid population. For the most part, they occupy

From K. Beatty, B. O'Leary, and A. Chaikin (eds.), *The New Solar System*, 2nd ed.

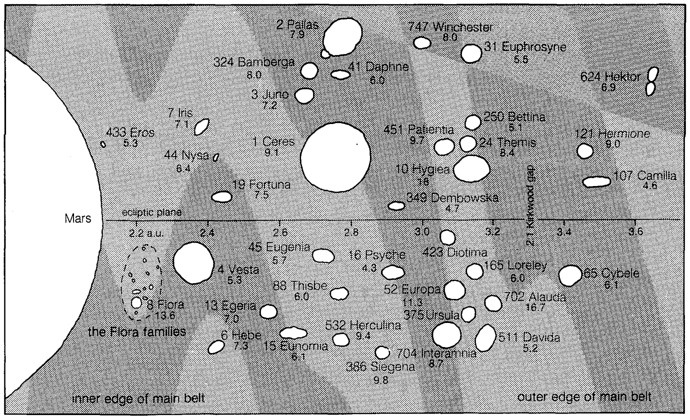


Figure 43: This representation of notable asteroids includes all those exceeding 200 kilometres in diameter. The asteroids are depicted in their correct relative sizes and shapes (the limb of Mars is included for comparison), and their rotational pole orientations (if known) are indicated. They are positioned at their correct relative distances from the Sun. Asteroids located near the top or bottom of the diagram travel in relatively eccentric and/or inclined orbits, whereas those near the ecliptic move in fairly circular, noninclined orbits. Rotational periods (in hours) are indicated beneath the names of the asteroids. Among the special smaller asteroids shown are the members of the Flora families larger than 15 kilometres across.

the same regions of the main asteroid belt as the moderate-albedo objects and make up the majority of the Hungaria group as well.

Mass and density. Because asteroid masses are low, they have a negligible effect on the orbits of the major planets. The masses of the three largest asteroids, however, have been measured by noting their effect on the orbits of other asteroids that they happen to approach closely at regular intervals. Since the diameters of these three largest asteroids also have been determined and since their shapes are either spherical or ellipsoidal, their volumes are known as well. Knowledge of the mass and volume of an asteroid allows its density to be computed.

The mass of the largest asteroid, Ceres, is 1.2×10^{24} grams (0.0002 the mass of the Earth). The masses of the second and third largest asteroids, Pallas and Vesta, are only 0.18 and 0.23 times the mass of Ceres. The mass of the entire asteroid belt is roughly three times that of Ceres. Most of the mass in the asteroid belt is concentrated in the larger asteroids. The 10th largest asteroid has only $^1/_{60}$ the mass of the largest, and about 90 percent of the total mass is contained in asteroids with diameters exceeding 100 kilometres. Ninety-one percent of the total mass of the asteroids is located in the main belt, 8 percent is in the outer belt, and the remainder is distributed among the inner belt and planet-crossing asteroid populations.

The densities of the largest asteroids as compared with those of the inner planets

The densities of Ceres, Pallas, and Vesta are 2.7, 2.7 and 4.2 grams per cubic centimetre, respectively. These compare with 5.4, 5.2, and 5.5 g/cm^3 for Mercury, Venus, and the Earth, respectively; 3.95 g/cm^3 for Mars; and 3.34 g/cm^3 for the Moon. The densities of Ceres and Pallas are similar to those of a class of meteorites known as carbonaceous chondrites, which contain a larger fraction of volatile material than do ordinary terrestrial rocks and hence have a somewhat lower density (see below *Specific asteroidal source regions for recovered meteorites*). The density of Vesta is similar to those of the rocky planets. Insofar as Ceres, Pallas, and Vesta are typical of asteroids in general and the first two are members of the most common asteroid compositional class, it can be concluded that asteroids are rocky bodies.

Composition. The combination of spectral reflectance measurements (measures of the amount of reflected sunlight at various wavelengths between about 0.3 and 1.1 micrometres) and albedos is used to classify asteroids into various taxonomic groups. If sufficient spectral resolution is available (*i.e.,* if about 25 approximately equally spaced wavelengths were sampled), these measurements also can be used to infer the composition of the surface reflecting the light. This can be done by comparing the asteroid data with that obtained in the laboratory using meteorites or terrestrial rocks or minerals.

Basic taxonomic classes

By the mid-1980s, high-resolution (24-wavelength) spectral reflectances had been obtained for about 300 asteroids, and low- (three-wavelength) or moderate- (eight-wavelength) resolution data had been secured for roughly 1,000, with albedos available for about 500 of these. Table 25 summarizes the nine taxonomic classes into which the asteroids can be divided on the basis of such data.

Asteroids of the C and F classes have low albedos and spectral reflectances similar to those of carbonaceous chondritic meteorites and carbon-rich terrestrial materials (*e.g.,* lampblack, or amorphous carbon), magnetite, and montmorillonite. Some C-class asteroids are known to have hydrated minerals on their surfaces, whereas Ceres probably has water present as a layer of permafrost. S-class asteroids have moderate albedos and spectral reflectances similar to the stony-iron meteorites, and they are known to contain significant amounts of silicates, including the minerals olivine, pyroxene, and feldspar on their surfaces. M-class asteroids are moderate-albedo objects, may have significant amounts of nickel-iron metal in their surface material, and exhibit spectral reflectances similar to the nickel-iron meteorites. P- and D-class asteroids have low albedos and no known meteorite or naturally occurring mineralogical counterparts, but they may contain a large fraction of carbon polymers in their surface material. R-class asteroids are very rare; only one (the high-albedo asteroid 349 Dembowska) has been identified for certain. The surface material of Dembowska has been identified as being most consistent with an olivine-rich composition that has no known meteoritic analogue. The E-class asteroids have the highest albedos and spectral reflectances that match those of the enstatite chondrite meteorites. Vesta, though having a high albedo and unique spectral reflectance, is sometimes referred to as a U-class asteroid. Its reflectance properties closely match those of one particular type of basaltic achondritic meteorite, the eucrite. The match is so good that some believe that Vesta is the parent body of the eucrite meteorites; that is to say, the eucrites exhibited in museums are chips from the surface of Vesta that were knocked off during a major cratering event. (For additional information about achondrite meteorites, see below *Types of meteorites.*)

The C-class asteroids are the most common, both numerically and by mass; 66 percent of the larger asteroids (those with diameters greater than 100 kilometres) and nearly 75 percent of the mass of such objects belong to this class. The next most common taxonomic class is S class, in which 15 percent of the asteroids and 7.5 percent of the mass are found. About 8 percent by number and 2.5 percent by mass are in the D class, and approximately 4 percent by number and 2 and 3 percent, respectively, by mass are in the P and M classes. The F class contains 2 percent by number and 1.8 percent by mass, and the remaining classes (E, R, and U) account for less than 1 percent of the population by number or mass. The unique asteroid Vesta contains about 8 percent of the mass of the entire asteroid belt.

The distribution of the taxonomic classes throughout the asteroid belt is highly structured, as can be seen from Figure 42. Some believe this variation with distance from the Sun means that the asteroids formed at or near their present locations and that a detailed comparison of the chemical composition of the asteroids in each region will provide constraints on models for the conditions that may have existed within the contracting solar nebula at the time the asteroids were formed.

Table 25: Summary of Asteroid Taxonomic Classes

class	albedo*	spectral reflectivity (0.3–1.1 micrometres)
C	low	neutral, slight absorption at wavelengths of 0.4 micrometre or shorter
D	low	very red at wavelengths of 0.7 micrometre or longer
F	low	flat
P	low	featureless, sloping up into red†
S	moderate	reddened, typically with an absorption band between 0.9 and 1.0 micrometre
M	moderate	featureless, sloping up into red†
E	high	featureless, sloping up into red†
R	high	very red with a deep absorption band between 0.9 and 1.0 micrometre
U	varies	unclassifiable in this system‡

*Low = visual geometric albedo between 0.02 and 0.065; moderate = visual geometric albedo between 0.066 and 0.23; and high = visual geometric albedo greater than 0.23.

†Classes E, M, and P are spectrally indistinguishable at these wavelengths and require an independent albedo measurement for unambiguous classification. ‡Includes Pallas and other unique or rare objects.

THE ORIGIN AND EVOLUTION OF THE ASTEROIDS

Available evidence indicates that the asteroids are the remnants of a "stillborn" planet. It is thought that at the time the planets were forming from the low-velocity collisions among asteroid-sized planetesimals, one of them grew at a high rate and to a size larger than the others. In the final stages of its formation this planet, Jupiter, gravitationally scattered large planetesimals, some of which may have been as massive as the Earth is today. These planetesimals were eventually either captured by Jupiter or another of the trans-Jovian planets or ejected from the solar system. While they were passing through the inner solar system, however, such large planetesimals strongly perturbed the orbits of the planetesimals in the region of the asteroid belt, raising their mutual velocities to the average five kilometres per second they exhibit today. These high-velocity collisions ended the accretionary collisions by transforming them into catastrophic disruptions. Only objects larger than about 500 kilometres in diameter could have survived collisions with objects of comparable size at collisional velocities of five kilometres per second. Since that time, the asteroids have been collisionally evolving so that, with the exception of the very largest, most present-day asteroids are either remnants or fragments of past collisions.

Evidence of differentiation

While breaking larger asteroids down into smaller ones, collisions expose deeper layers of asteroidal material. If asteroids were compositionally homogeneous, this would have no noticeable result. Some of them, however, have become differentiated since their formation. This means that some asteroids, originally formed from so-called primitive material (*i.e.,* material of nonvolatile solar composition), were heated, perhaps by short-lived radionuclides or solar magnetic induction, to the point where their interiors melted and geochemical processes occurred. In certain cases, temperatures became high enough for iron to form. Being denser than other materials, the iron then sank to the centre, forming an iron core and forcing basaltic lavas onto the surface. At least one asteroid with a basaltic surface, Vesta, survives to this day. Other differentiated asteroids were disrupted by collisions that stripped away their crusts and mantles and exposed their iron cores. Still others may have had only their crusts partially stripped away, which exposed surfaces such as those visible today on the E- and R-class asteroids.

Collisions were responsible for the formation of the Hirayama families and at least some of the planet-crossing asteroids. A number of the latter enter the Earth's atmosphere, giving rise to sporadic meteors. Larger pieces survive passage through the atmosphere, and some of these end up in museums and laboratories as meteorites (see below *Relationship of meteoroids to asteroids and comets* and *Meteorites, asteroids, and the early solar system*). The very largest produce craters such as Meteor Crater in Arizona in the southwestern United States, and one may even have been responsible for the extinction of the dinosaurs some 65,000,000 years ago. This extinction may have been triggered by an explosion resulting from the impact of an asteroid measuring about 10 kilometres in diameter. (Some investigators believe that a cometary body rather than an asteroid may have caused such an explosion.) Fortunately, collisions of this sort are rare. According to current estimates, three asteroids of one-kilometre diameter collide with the Earth every 1,000,000 years.

(E.F.T.)

Comets

GENERAL CONSIDERATIONS

Differences between cometary nuclei and asteroids

Basic features. The traditional definition of a comet is a nebulous body with a "hairy" tail that makes a transient appearance in the sky. The word comet comes from the Greek *komētēs* meaning "hairy one," a description that fits the bright comets noticed by the ancients. Many comets, however, do not develop tails. Moreover, comets are not surrounded by any nebulosity during most of their lifetime. The only permanent feature of a comet is its nucleus, which is a small body that may be seen as a stellar image in large telescopes when tail and nebulosity do not exist, particularly when the comet is still far away from the Sun. Two characteristics differentiate the cometary nucleus from a very small asteroid—namely, its orbit and its chemical nature. A comet's orbit is more eccentric; therefore, its distance to the Sun varies considerably. Its material is more volatile. When far from the Sun, however, a comet remains in its pristine state for eons without losing any volatile components due to the deep cold of space. For this reason, astronomers believe that pristine cometary nuclei may represent the oldest and best-preserved material in the solar system.

During a close passage near the Sun, the nucleus of a comet looses water vapour and other more volatile compounds, as well as dust dragged away by the sublimating gases. It is then surrounded by a transient dusty "atmosphere" that is steadily lost to space. This feature is the coma, which gives a comet its nebulous appearance. The nucleus surrounded by the coma makes up the head of the comet. When even closer to the Sun, solar radiation usually blows the dust of the coma away from the head and produces a dust tail, which is often rather wide, featureless, and yellowish. The solar wind, on the other hand, drags ionized gas away in a slightly different direction and produces a plasma tail, which is usually narrow with nods and twists and has a bluish appearance.

Designations. In order to classify the chronological appearance of comets, the *Astronomische Nachrichten* ("Astronomical Reports") introduced in 1870 a system of preliminary and final designations that is still used today with only minor modifications. The preliminary designation classifies comets according to their order of discovery, using the year of discovery followed by a lowercase letter in alphabetical order, as in 1987a, 1987b, 1987c, and so forth. Comets are reclassified as soon as possible—usually a few years later—according to their chronological order of passage at perihelion (closest distance to the Sun); a Roman numeral is used in this case, as in 1987 I, 1987 II, 1987 III, and so on. Since the discovery may have taken place at any time either before or after perihelion passage, the two chronologies are not necessarily in the same order, and even the year may change in the final designation. The official designation generally includes the name(s) of its discoverer(s)—with a maximum of three names—preceded by a P/ if the comet is on a periodic orbit. If a person discovers several comets, an Arabic numeral is used after his name, as in 1867 II Tempel 1 and 1873 II Tempel 2. The discoverer's rule has not always been strictly applied: comets P/Halley, P/Lexell, P/Encke, and P/Crommelin have been named after the astronomers who proved their periodic character. Some comets become bright so fast that they are discovered by a large number of persons at almost the same time. They are given an arbitrary impersonal designation such as Brilliant Comet (1882 II), Southern Comet (1947 XII), or Eclipse Comet (1948 XI). Finally, comets may be discovered by an unusual instrument without direct intervention of a specific observer, as in the case of the Earth-orbiting Infrared Astronomical Satellite (IRAS). Its initials are used as if it were a human observer, as in 1983 VII IRAS-Araki-Alcock.

HISTORICAL SURVEY OF COMET OBSERVATIONS AND STUDIES

Early observations. In ancient times, without interference from streetlights or urban pollution, comets could be seen by everyone. Their sudden appearance—their erratic behaviour against the harmonious order of the heavenly motions—was interpreted as an omen of nature that awed people and was used by astrologers to predict flood, famine, pestilence, or the death of kings. The Greek philosopher Aristotle (4th century BC) thought that the heavens were perfect and incorruptible. The very transient nature of comets seemed to imply that they were not part of the heavens but were merely earthly exhalations ignited and transported by heat to the upper atmosphere. Although the Roman philosopher Seneca (1st century AD) had proposed that comets could be heavenly bodies like the planets, Aristotle's ideas prevailed until the 14th century AD. Finally, during the 16th century the Danish nobleman Tycho Brahe established critical proof that comets are heavenly bodies. He compared the lack of diurnal par-

The studies of Tycho Brahe

allax of the comet of 1577 with the well-known parallax of the Moon (the diurnal parallax is the apparent change of position in the sky relative to the distant stars due to the rotation of the Earth). Tycho deduced that the comet was at least four times farther away than the Moon, establishing for the first time that comets were heavenly bodies.

The impact of Newton's work. The German astronomer Johannes Kepler still believed in 1619 that comets travel across the sky in a straight line. It was the English physicist and mathematician Isaac Newton who demonstrated in his *Principia* (1687) that, if heavenly bodies are attracted by a central body (the Sun) in proportion to the inverse-square of its distance, they must move along a conic section (circle, ellipse, parabola, or hyperbola). Using the observed positions of the Great Comet of 1680, he identified its orbit as being nearly parabolic.

Newton's friend, the astronomer Edmond Halley, endeavoured to compute the orbits of 24 comets for which he had found accurate enough historical documents. Applying Newton's method, he presupposed a parabola as an approximation for each orbit. Among the 24 parabolas, three were identical in size and superimposed in space. The three relevant cometary passages (1531, 1607, and 1682) were separated by two time intervals of 76 and 75 years. Halley concluded that the parabolas were actually the end of an extremely elongated ellipse. Instead of three curves open to infinity, the orbit is closed and brings the same comet periodically back to the Earth. As a consequence, it would return in 1758, he predicted. Observed on Christmas night, 1758, by Johann Georg Palitzsch, a German amateur astronomer, the comet passed at perihelion in March 1759 and at perigee (closest to the Earth) in April 1759. The perihelion date of 1759 had been predicted by Alexis-Claude Clairaut, a French astronomer and physicist, with an accuracy of one month. Clairaut's work contributed much to the acceptance of Newton's theory on the Continent. With this, the until-then anonymous comet came to be called Halley's comet (or, in modern nomenclature, Comet P/Halley).

Passages of Comet P/Halley. Since 1759, Comet Halley has reappeared three more times—in 1835, 1910, and 1986. Its trajectory has been computed backward, and all of its 30 previous passages described in historical documents over 22 centuries have been authenticated. Comet Halley's period has irregularly varied between 74.4 years (from 1835 to 1910) and 79.6 years (from AD 451 to 530). These variations, which have been accurately predicted, result from the changing positions of the giant planets, mainly Jupiter and Saturn, whose variable attractions perturb the trajectory of the comet. The space orientation of the orbit has been practically constant, at least for several centuries. Since its returns are not separated by an integer number of years, however, the comet encounters the Earth each time on a different point of its orbit around the Sun; thus, the geometry of each passage is different and its shortest distance to the planet varies considerably. The closest known passage to the Earth, 0.033 a.u., occurred on April 9, AD 837.

The 1986 passage of Comet Halley

The perigee distance of most of Comet Halley's historical passages has been between 0.20 and 0.50 a.u. The last perigee, on April 11, 1986, took place at 0.42 a.u. from the Earth (Figure 44). By contrast, the comet passed at only 0.14 a.u. from the Earth in 1910. Seen from closer range, it was brighter and had a longer tail than on its return in 1986. This is one reason why the 1986 passage proved so disappointing to most lay observers. Yet, a far more important factor had to do with geometry: in the latitudes of the major Western countries, the comet was hidden by the southern horizon during the few weeks in April 1986 when it was at its brightest. Moreover, the night sky of most Western countries is brightly and constantly illuminated by public and private lights. Even in the absence of moonlight, the nighttime sky is pervaded by a milky glare that easily hides the tail of a comet.

Each century, a score of comets brighter than Comet Halley have been discovered. Yet, they appear without warning and will not be seen again. Many are periodic comets like Comet Halley, but their periods are extremely long (millennia or even scores or hundreds of millennia), and they have not left any identifiable trace in prehistory. Bright Comet Bennett 1970 II will return in 17 centuries, whereas the spectacular Comet West 1976 VI will reappear in about 500,000 years. Among the comets that can easily be seen with the unaided eye, Comet Halley is the only one that returns in a single lifetime. Approximately 100 comets whose periods are between three and 200 years are known, however. Unfortunately they are or have become too faint to be readily seen without the aid of telescopes (see below *Periodic comets*).

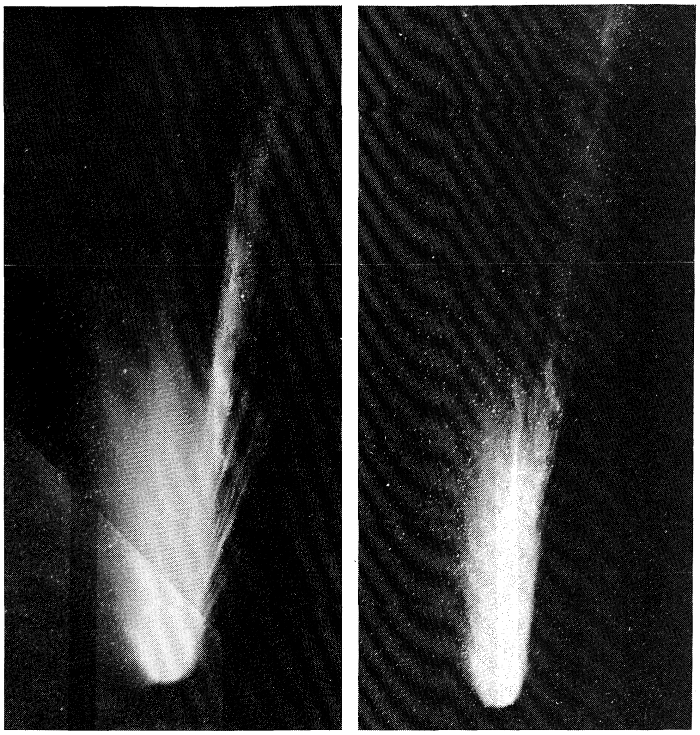


Figure 44: Comet Halley photographed on March 8 and 9, 1986, by the 1-metre Schmidt telescope of the European Southern Observatory at La Silla, Chile.

By courtesy of the European Southern Observatory

Modern cometary research. During the 19th century it was shown that the radiant (*i.e.,* spatial direction) of the spectacular meteor showers of 1866, 1872, and 1885 coincided well with three known cometary orbits that happened by chance to cross the Earth's orbit at the dates of the observed showers. The apparent relationship between comets and meteor showers was interpreted by assuming that the cometary nucleus was an aggregate of dust or

By courtesy of the Department of Astronomy, University of Michigan, Ann Arbor

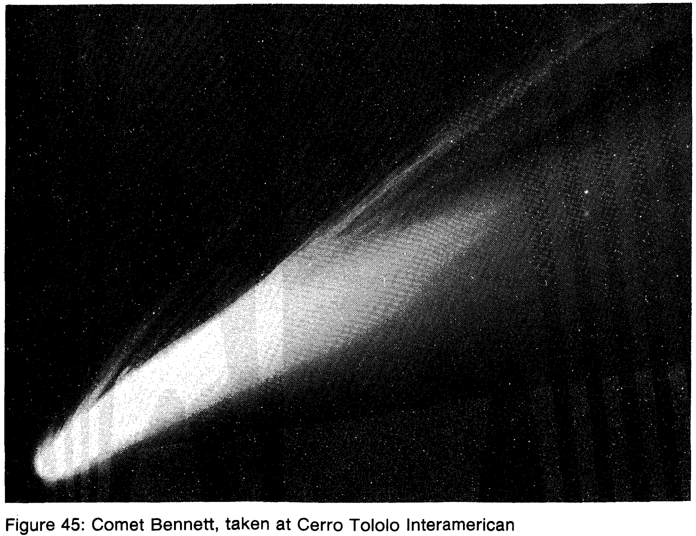


Figure 45: Comet Bennett, taken at Cerro Tololo Interamerican Observatory, Chile, March 16, 1970.

sand grains without any cohesion. (This conception of the cometary nucleus became known as the "sandbank" model [see below *Cometary models*].) Meteor showers were explained by the spontaneous scattering of the dust grains along a comet's orbit, and the cometary nucleus began to be regarded only as the densest part of a meteor stream. At the end of the 19th and the beginning of the 20th century, spectroscopy revealed that the reflection of sunshine by the dust was not the only source of light in the tail; it showed the discontinuous emission that constitutes the signature of gaseous compounds. More specifically, it revealed the existence in the coma of several radicals—molecular fragments such as cyanogen (CN) and the carbon forms C_2 and C_3, which are chemically unstable in the laboratory because they are very reactive in molecular collisions. Spectroscopy also enabled investigators to detect the existence of a plasma component in the cometary tail by the presence of molecular ions, as, for example, those of carbon monoxide (CO^+), nitrogen (N_2^+), and carbon dioxide (CO_2^+). The radicals and ions are built up by the three light elements carbon (C), nitrogen (N), and oxygen (O). Hydrogen (H) was added when the radical CH was discovered belatedly on spectrograms of Comet Halley taken in 1910. The identification of CH was proposed by the American astronomer Nicholas Bobrovnikoff in 1931. In 1941, the Belgian astronomer Pol Swings and his coworkers identified three new ions: CH^+, OH^+, and CO_2^+. The emissions of the light elements hydrogen, carbon, oxygen, and sulfur and of carbon monoxide were finally detected when the far ultraviolet spectrum (which is absorbed by the Earth's atmosphere) was explored during the 1970s with the help of rockets and satellites. This included the very large halo (10^7 kilometres) of atomic hydrogen (the Lyman-alpha emission line) first observed in comets Tago-Sato-Kosaka 1969 IX and Bennett 1970 II.

Spectroscopic discoveries

Although the sandbank model was still seriously considered until the 1960s and 1970s by a small minority (most notably the British astronomer Raymond A. Lyttleton), the presence of large amounts of gaseous fragments of volatile molecules in the coma suggested to Bobrovnikoff the release by the nucleus of a bulk of unobserved "parent" molecules such as H_2O, CO_2, and NH_3 (ammonia). In 1948, Swings proposed that these molecules should be present in the nucleus in the solid state as ices.

In a fundamental paper, the American astronomer Fred L. Whipple set forth in 1950 the so-called dirty snowball model, according to which the nucleus is a lumpy piece of icy conglomerate wherein dust is cemented by a large amount of ices—not only water ice but also ices of more volatile molecules. This amount has to be substantial enough to sustain the vaporizations for a large number of revolutions. Whipple noted that the nuclei of some comets at least are solid enough to graze the Sun without experiencing total destruction, since they apparently survive unharmed. (Some but not all Sun-grazing nuclei split under solar tidal forces.) Finally, argued Whipple, the asymmetric vaporization of the nuclear ices sunward produces a jet action opposite to the Sun on the solid cometary nucleus. When the nucleus is rotating, the jet action is not exactly radial. This explained the heretofore mysterious nongravitational force identified as acting on cometary orbits. In particular, the orbital period of P/Encke mysteriously decreased by one to three hours per revolution (of 3.3 years), whereas that of P/Halley increased by some three days per revolution (of 76 years). For Whipple, a prograde rotation of the nucleus of P/Encke and a retrograde rotation of that of P/Halley could explain these observations. In each case, a similar amount of some 0.5 to 0.25 percent of the ices had to be lost per revolution to explain the amount of the nongravitational force. Thus, all comets decay in a matter of a few hundred revolutions. This duration is only at most a few centuries for Encke and a few millennia for Halley. At any rate, it is millions of times shorter than the age of the solar system.

Dirty snowball model

The observed comets, however, have obviously survived until now. If they have existed for a very long time, they must have been stored in an extremely cold place far away from the Sun before recently coming into the inner solar system where they could be seen from Earth. A reply to such a suggestion had already been anticipated in 1932 by the Estonian-born astronomer Ernest J. Öpik, who proposed the possible existence of a large cloud of unobservable comets surrounding the solar system. Nearly 20 years later, the Dutch astronomer Jan Hendrick Oort established the existence of such a cloud of comets by indirect reasoning based on observations. Since the appearance of his theory in 1950, this enormous cloud of comets has come to be called the Oort cloud.

The Oort cloud

Oort showed by statistical arguments that a steady flux of a few "new" comets are observed per year (new meaning that they had never been through the solar system before; see below *Periodic comets*). This flux comes from the fringe of the Oort cloud. He identified it by looking at the distribution of the original values of the total energies of cometary orbits. These energies are in proportion to a^{-1}, with a being the semimajor axis of the cometary orbit. The original value of a refers to the orbit when the comet was still outside of the solar system, as opposed to the osculating orbit, which refers to the arc observed from the Earth after it has already been modified by the perturbations of the giant planets. Passages through the solar system produce a rather wide diffusion in orbital energies (in a^{-1}). In 1950 Oort accounted for only 19 accurate original orbits of long-period comets. The very fact that 10 of the 19 orbits were concentrated in a very narrow range of a^{-1} established that most of them had never been through this diffusion process due to the planets. The mean value of a for these new comets suggested the distance they were coming from: about 10^5 a.u. This distance is also the place where perturbations resulting from the passage of nearby stars begin to be felt. The distance coincidence suggested to Oort that stellar perturbations were indeed the mechanism by which comets were sent into the planetary system.

Subsequent work by the American astronomer Brian G. Marsden and his coworkers confirmed the existence of the Oort cloud. Their list of approximately 90 original orbits crammed within an extremely narrow range of a^{-1} corroborated Oort's initial effort. The mean aphelion distance of this list of new comets implies, however, that the Oort cloud margin is only at some 40,000 to 50,000 a.u., which makes the standard mechanism of stellar perturbations much less effective than Oort had believed. Comets must therefore come down from the Oort cloud in several steps, penetrating first into the outer solar system where the perturbations of Uranus and Neptune are weak enough not to remove them from the action of passing stars except after several revolutions.

During the late 1980s astronomers explored new ideas with which to determine how the outer perturbations on the Oort cloud could increase. Dark molecular clouds, for example, may be substituted for stars as major perturbing agents. The hypothesis that there exists some extra undetected matter (like black dwarfs) in the disk of the Galaxy has also been used. Then, the total mass distribution in the galactic disk may be large enough to induce tidal forces in the Oort cloud that would change cometary orbits.

MOTION AND DISCOVERY OF COMETS

Types of orbits. In the absence of planetary perturbations and nongravitational forces, a comet will orbit the Sun on a trajectory that is a conic section with the Sun at one focus. The total energy E of the comet, which is a constant of motion, will determine whether the orbit is an ellipse, a parabola, or a hyperbola. The total energy E is the sum of the kinetic energy of the comet and of its gravitational potential energy in the gravitational field of the Sun. Per unit mass, it is given by $E = \frac{1}{2}v^2 - GMr^{-1}$, where v is the comet's velocity and r its distance to the Sun, with M denoting the mass of the Sun and G the gravitational constant. If E is negative, the comet is bound to the Sun and moves in an ellipse. If E is positive, the comet is unbound and moves in a hyperbola. If E = 0, the comet is unbound and moves in a parabola.

Total energy of a comet

In polar coordinates written in the plane of the orbit, the general equation for a conic section is

$$r = q(1 + e)\,(1 + e\cos\theta)^{-1},$$

where r is the distance from the comet to the Sun, q the perihelion distance, e the eccentricity of the orbit, and θ an angle measured from perihelion. When $0 \leq e < 1$, $E < 0$ and the orbit is an ellipse (the case $e = 0$ is a circle, which constitutes a particular ellipse); when $e = 1$, $E = 0$ and the orbit is a parabola; and when $e > 1$, $E > 0$ and the orbit is a hyperbola.

Orbital elements

In space, a comet's orbit is completely specified by six quantities called its orbital elements. Among these are three angles that define the spatial orientation of the orbit: *i,* the inclination of the orbital plane to the plane of the ecliptic; Ω, the longitude of the ascending node measured eastward from the vernal equinox; and *ω*, the angular distance of perihelion from the ascending node (also called the argument of perihelion). The three most frequently used orbital elements within the plane of the orbit are *q,* the perihelion distance in astronomical units; *e,* the eccentricity; and *T,* the epoch of perihelion passage.

Identifying comets and determining their orbits. Up to the beginning of the 19th century, comets were discovered exclusively by visual means. Many discoveries are still made visually with moderate-sized telescopes by amateur astronomers. Although comets can be present in any region of the sky, they are often discovered near the western horizon after sunset or near the eastern horizon before sunrise, since they are brightest when closest to the Sun. Because of the Earth's rotation and direction of motion in its orbit, discoveries before sunrise are more likely, as confirmed by discovery statistics. At discovery, a comet may still be faint enough not to have developed a tail; therefore, it may look like any nebulous object—*e.g.,* an emission nebula, a globular star cluster, or a galaxy. The famous 18th-century French comet hunter Charles Messier (nicknamed "the ferret of comets" by Louis XV for his discovery of 21 comets) compiled his well-known catalog of "nebulous objects" so that such objects would not be mistaken for comets. The final criterion remains the apparent displacement of the comet after a few hours or a few days with respect to the distant stars; by contrast, the nebulous objects of Messier's catalog do not move. After such a displacement has been undisputably observed, any amateur wishing to have the comet named for himself must report his claim to the nearest observatory as soon as possible.

Photographic discovery of comets

Most comets are and remain extremely faint. Today, a larger and larger proportion of comet discoveries are thus made fortuitously from high-resolution photographs, as, for instance, those taken during sky surveys by professional astronomers engaged in other projects.

The faintest recorded comets are approaching the limit of detection of large telescopes (those that are two metres or more in diameter). That is to say, they are of the 22nd–23rd magnitude, or 10^6 to 10^7 times fainter than the limit of the naked eye. Several successive photographic observations of these faint moving objects are necessary to ensure identification and simultaneous calculation of a preliminary orbit. In order to determine a preliminary orbit as quickly as possible, the eccentricity $e = 1$ is assumed since some 90 percent of the observed eccentricities are close to one, and a parabolic motion is computed. This is generally sufficient to ensure against "losing" the comet in the sky.

The best conic section representing the path of the comet at a given instant is known as the osculating orbit. It is tangent to the true path of the chosen instant, and the velocity at that point is the same as the true instantaneous velocity of the comet. Nowadays, high-speed computers make it possible to produce a final ephemeris (table of positions) that is not only based on the definitive orbit but also includes the gravitational forces of the Sun and of all significant planets that constantly change the osculating orbit. In spite of this fact, the deviation between the observed and the predicted positions usually grows (imperceptibly) with the square of time. This is the signature of a "neglected" acceleration, which comes from a nongravitational force (see above). Formulas representing the smooth variation of the nongravitational force with heliocentric distance are now included for many orbits. The best formula assumes that water ice prevails and controls the vaporization of the nucleus.

Nongravitational forces

COMETARY STATISTICS

The *Catalog of Cometary Orbits,* compiled by Marsden, remains the standard reference for orbital statistics. Its 1986 edition lists 1,187 computed orbits from 239 BC to AD 1985; only 91 of them were computed using the rare accurate historical data from before the 17th century. More than 1,000 are therefore derived from cometary passages during the last three centuries. The 1,187 cometary apparitions of Marsden's catalog involve only 748 individual comets; the remainder represents the repeated returns of periodic comets.

Short- and intermediate-period comets

Periodic comets. The periodic comets are usually divided into short-period comets (those with periods of less than 200 years) and long-period comets (those with periods of more than 200 years). Of the 135 short-period comets, 85 have been observed at two or more perihelion passages. In 1986, four of these comets had been definitely lost, and three more were probably lost, presumably because of their decay in the solar heat. Some authors have found it advantageous to change the definition of short-period comets by diminishing their longest-period cutoff to 20 years. This leaves 116 short-period comets (new style) in the *Catalog;* the 19 others having periods between 20 and 200 years are called intermediate-period comets. These two new classes are separated by a period gap of close to 20 years. The average short-period comet has a seven-year period, a perihelion distance of 1.5 a.u., and a small inclination (13°) on the ecliptic. All short-period comets (new style) revolve in the direct sense around the Sun, just as the planets do. The intermediate-period comets have on average a larger inclination of the ecliptic, and four of them turn around the Sun in the retrograde. The most famous of the latter is P/Halley (30 appearances); the others are P/Tempel-Tuttle (four appearances), P/Pons-Gambart, and P/Swift-Tuttle (the last two with only one appearance each). Ten of the 19 intermediate-period comets have been observed during a single appearance.

Long-period comets

The comets with long-period orbits are distributed at random in all directions of the sky, and roughly half of them turn in the retrograde direction. Of the 613 comets of long period contained in the *Catalog,* 179 have osculating elliptic orbits, and 112 have osculating orbits that are very slightly hyperbolic. Finally, 322 are listed as having parabolic orbits, but this is rather fallacious because either it has not been possible to detect unequivocal deviations from a parabola on the (sometimes very short) arc along which the comets have been observed or, more simply, the final calculations have never been made. The parabola is always assumed first in the preliminary computation as it is easier to deal with. If the osculating orbit is computed backward to when the comet was still far beyond the orbit of Neptune and if the orbit is then referred to the centre of mass of the solar system, the original orbits almost always prove to be elliptic. (The centre of mass of the solar system is different from the centre of the Sun primarily because of the position of massive Jupiter.) Twenty-two original orbits remain (nominally) slightly hyperbolic beyond the orbit of Neptune, but 19 remain not significantly different from a parabola. Even the three that are significantly different near 50 a.u. are likely to become elliptic when they are 50,000 or 100,000 a.u. from the Sun. The reason is that, though the mass of the Oort cloud remains uncertain, it should be added to the mass of the inner solar system to compute the orbits. The smallest possible mass of the Oort cloud is likely to transform the orbits into ellipses. It is thus reasonable to believe that all observed comets were initially on elliptic orbits bound to the solar system. Accordingly, all parabolic and nearly parabolic comets are thought to be comets of very long period.

The future orbit of a long-period comet is obtained when the osculating orbit is computed forward to when the comet will be leaving the planetary system (beyond the orbit of Neptune) and is referred to the centre of mass of the solar system. Because of the planetary perturbations, slightly more than half of the future orbits become strongly elliptic, whereas slightly less than half become strongly hyperbolic. Roughly half of the long-period comets are thus "captured" by the solar system on more strongly bound orbits; the other half are ejected forever out of the system.

"New" comets

Among the very-long-period comets, there is a particular class that Oort showed as having never passed through the planetary system before (see above), notwithstanding the fact that their original orbits were elliptic, which implies repeated passages. This paradox vanishes when it is understood that their perihelia were outside of the planetary system before their first appearance but that their orbits have been perturbed near aphelia (either by stellar or dark interstellar-cloud passages or by galactic tides) in such a way that their perihelia were lowered into the planetary system. The first passage of a "new" comet is usually brighter than an average passage (a large fraction of the famous bright historical comets were such new comets). This is possibly explained by the presence of more volatile gases and of a larger component of very fine dust. The most volatile gases may have disappeared during subsequent passages, and the finest dust may have agglomerated into larger dust grains that reflect less light for the same production rate. About 90 comets have been identified as new in long-period orbits. If the same proportion exists in the poorly computed parabolic orbits, the total must be close to 170 new comets in Marsden's catalog, but 80 of them have not been identified.

Groups of comets and other unusual cometary objects. Some comets travel in strikingly similar orbits, only the time of perihelion passages being appreciably different. Members of such a group of comets are thought to be fragments from a larger comet that was tidally disrupted earlier by the Sun or in some cases by the differential jet action of nongravitational forces on a fragile nucleus. Many such breakups have been observed historically. Slight differences in the resultant velocities—though they occur very gently—are sufficient to cause cometary fragments to separate along orbits close to but distinct from each other, particularly as far as their total energy is concerned. A very slight variation in a^{-1} introduces an orbital period that may vary by several years, and when the cometary fragments return they will go through perihelion at widely separated epochs. The best-known example is the famous group of "Sun-grazing" comets (also called the Kreutz group), which has 12 definite members (plus one probable) with perihelion distances between 0.002 and 0.009 a.u. (less than half a solar radius). Their periods are scattered from 400 to 2,000 years, and their last passages occurred between 1880 and 1970. The most famous fragment of the group is Comet Ikeya-Seki 1965 VIII.

The "Sun-grazing" comet group

Comet P/Schwassmann-Wachmann 1, which has a period of 15 years, is in a quasi-circular and somewhat unstable orbit between Jupiter and Saturn, with a perihelion q that equals 5.45 a.u. and an aphelion of 6.73 a.u. It can be observed every year for several months when opposite to the Sun in the sky. Without any visible tail, it has irregular outbursts that make its coma grow in size for a few weeks and become up to 1,000 times as bright as normal. Another unusual object is the so-called asteroid 2060 Chiron, which has a similar orbit between Saturn and Uranus. Though classified as an asteroid, its icy nucleus of 350 kilometres would seem to suggest that it is a giant comet provisionally parked in a quasi-circular but unstable orbit. Within a few thousand years, Chiron might be perturbed enough by Saturn to become a spectacular comet. Two bright comets, Morehouse 1908 III and Humason 1962 VIII, exhibited a peculiar tail spectrum in which the ion CO^+ prevailed in a spectacular way, possibly because of an anomalous abundance of a parent molecule (carbon monoxide, carbon dioxide, or possibly formaldehyde [CH_2O]) vaporizing from the nucleus. Finally, Comet Halley is the brightest and therefore the most famous of all short- and intermediate-period comets as the only one that returns in a single lifetime and can be seen with the naked eye.

THE NATURE OF COMETS

The nucleus. As previously noted, the traditional picture of a comet with a hazy head and a spectacular tail applies only to a transient phenomenon produced by the decay in the solar heat of a tiny object known as the cometary nucleus. In the largest telescopes, the nucleus is never more than a bright point of light at the centre of the cometary head. At substantial distances from the Sun, the comet seems to be reduced to its starlike nucleus. The nucleus is the essential part of a comet because it is the only permanent feature that survives during the entire lifetime of the comet. In particular, it is the source of the gases and dust that are released to build up the coma and tail when a comet approaches the Sun. The coma and tail are enormous: typically the coma measures 100,000 kilometres or more in diameter, and the tail extends about 100,000,000 kilometres in length. They scatter and continuously dissipate into space but are steadily rebuilt by the decay of the nucleus, whose size is usually in the range of 10 kilometres.

The evidence on the nature of the cometary nucleus remained completely circumstantial until March 1986, when the first close-up photographs of the nucleus of Comet Halley were taken during a flyby by the *Giotto* spacecraft of the European Space Agency (Figure 46). Whipple's basic idea that the cometary nucleus was a monolithic piece of icy conglomerate (see above) had been already well supported by indirect deductions in the 1960s and '70s and had become the dominant though not universal view. The final proof of the existence of such a "dirty snowball," however, was provided by the photographs of Comet Halley's nucleus.

If there was any surprise, it was not over its irregular shape (variously described as a potato or a peanut), which had been expected for a body with such small gravity (10^{-4}). Rather, it was over the very black colour of the nucleus, which suggests that the snows or ices are indeed mixed together with a large amount of sootlike stuff (*i.e.*, carbon and tar in fine dust form). The very low geometric albedo (2 to 4 percent) of the cometary nucleus puts it among the darkest objects of the solar system. Its size is thus somewhat larger than anticipated: the roughly elongated body measures 15 by eight kilometres and has a total volume of some 500 cubic kilometres. Its mass is rather uncertain, estimated in the vicinity of 10^{17} grams, and its bulk density is very small, ranging anywhere from 0.1 to 0.8 gram per cubic centimetre. The infrared spectrometer on board the Soviet Vega 2 spacecraft estimated a surface temperature of 300 to 400 K for the inactive "crust" that seems to cover 90 percent of the nucleus. Whether this crust is only a warmer layer of outgassed dust or whether

By courtesy of H.U. Keller and H. Reitsema; copyright Max-Planck-Institut fur Aeronomie, Lindau, W.Ger.

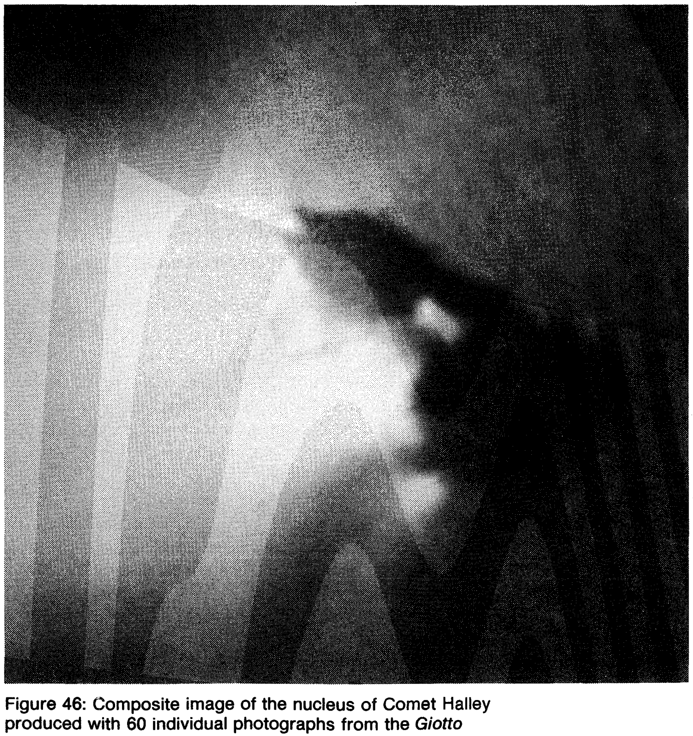


Figure 46: Composite image of the nucleus of Comet Halley produced with 60 individual photographs from the *Giotto* spacecraft. The photographs were taken on March 13 and 14, 1986, at various distances with the Halley Multicolour Camera carried on board the probe.

the dust particles are really fused together by vacuum welding under contact is still open to speculation.

Chemical composition of Comet Halley's nucleus

The 10 percent of the surface of Halley's nucleus that shows signs of activity seems to correspond to two large and a few smaller circular features resembling volcanic vents. Large sunward jets of dust originate from the vents; they are clearly dragged away by the gases vaporizing from the nucleus. This vaporization has to be a sublimation of the ices that cools them down to no more than 200 K in the open vents. The chemical composition of the vaporizing gases, as expected, is dominated by water vapour (more than 80 percent of the total production rate). The next most abundant volatile (close to 10 percent) appears to be carbon monoxide (CO), though it could come from the dissociation of another parent molecule (*e.g.*, CO_2 or CH_2O). Following CO in abundance is carbon dioxide (nearly 5 percent). Methane (CH_4) and ammonia (NH_3), on the other hand, seem to be close to the 1 percent level, and the percentage of carbon disulfide (CS_2) is even lower; at that level, there also must be unsaturated hydrocarbons and amino compounds responsible for the molecular fragments observed in the coma. This is not identical to—though definitely reminiscent of—the composition of the volcanic gases on the Earth, which also are dominated by water vapour but contain much more CO_2 than CO and occasionally some methane and sulfur in the molecular species, S_2, as well. The major difference may stem from the different temperature involved—often near 1,300 K in terrestrial volcanoes, as opposed to 200 K for cometary vaporizations. This may make the terrestrial gases closer to thermodynamic equilibrium. The dust-to-gas mass ratio is uncertain but is possibly in the vicinity of 0.4 to 1.1.

The dust grains are predominantly silicates. Mass spectrometric analysis by the *Giotto* spacecraft revealed that they contain as much as 20–30 percent carbon, which explains why they are so black. There also are grains composed almost entirely of organic material (molecules made of atoms of hydrogen, carbon, nitrogen, and oxygen).

There is some uncertainty concerning the rotation of Halley's nucleus. Two different rotation rates of 2.2 days and 7.3 days have been deduced by different methods. Both may exist, one of them involving a tumbling motion, or nutation, that results from the irregular shape of the nucleus, which has two quite different moments of inertia along perpendicular axes.

Scientific knowledge of the internal structure of the cometary nucleus was not enhanced by the flyby of Comet Halley, and so it rests on weak circumstantial evidence from the study of other comets. Earlier investigations had established that the outer layers of old comets were processed by solar heat. These layers must have lost most of their volatiles and developed a kind of outgassed crust, which probably measures a few metres in thickness. Inside the crust there is thought to exist an internal structure that is radially the same at any depth. Arguments supporting this view are based on the fact that cometary comas and tails do not become essentially different when comets decay. Since they lose more and more of their outer layers, however, the observed phenomena come from material from increasingly greater depths. These arguments are specifically concerned with the dust-to-gas mass ratio, the atomic and molecular spectra, the splitting rate, and the vaporization pattern during fragmentation.

Variation in size of nuclei

Before the *Giotto* flyby of Comet Halley, other cometary nuclei had never been resolved optically. For this reason, their albedo (fraction of incident light reflected) had to be assumed first in order to compute their sizes. Techniques proposed to deduce the albedo yielded only that of the dusty nuclear region made artificially brighter by light scattering in the dust. In 1986 the albedo of Comet Halley's nucleus was found to be very low (A = 2 to 4 percent). If this value is typical for other comets, then 11 of 18 short-period comets studied would be between six and 10 kilometres in diameter; only seven of them would be somewhat outside these limits. Comet Schwassmann-Wachmann 1 (see above) would be a giant with a diameter of 96 kilometres; 10 long-period comets would all have diameters close to 16 kilometres (within 10 percent). Since short-period comets have remained much longer in the solar system than comets having very long periods, the smaller size of the short-period comets might result from the steady fragmentation of the nucleus by splitting. Yet, the albedo may also diminish with aging. At the beginning, if the albedo were close to that of a slightly less dirty snow (A = 10 percent), the nuclear diameter of long-period comets would come very close to that of the largest of the short-period comets. The diameters of new comets also have been shown to be rather constant and most likely measure close to 10 kilometres. Of course, these are mean "effective" diameters of unseen bodies that are all likely to be very irregular.

The region around the nucleus, up to 10 or 20 times its diameter, contains an amount of dust large enough to be partially and irregularly opaque or at least optically thick. It scatters substantially more solar light than is reflected by the black nucleus. Dust jets develop mainly sunward, activated by the solar heat on the sunlit side of the nucleus. They act as a fountain that displaces somewhat the centre of light from the centre of mass of the nucleus. This region also is likely to contain large clusters of grains that have not yet completely decayed into finer dust; the grains are cemented together by ice.

Molecular collision zone

The gaseous coma. The coma, which produces the nebulous appearance of the cometary head, is a short-lived, rarefied, and dusty atmosphere escaping from the nucleus. It is seen as a spherical volume having a diameter of 10^5 to 10^6 kilometres, centred on the nucleus. The coma gases expand at a velocity of about 0.6 kilometres per second. This velocity can be measured from the motion of expanding "halos" triggered by outbursts in the nucleus, from the speed required to produce the Greenstein effect (see below), and from the fluid dynamics required to drag dust particles away at those places where they are observed in the dust tails. This expansion velocity varies somewhat with heliocentric distance r: $v = 0.58r^{-0.5}$ (in kilometres per second, when r is in astronomical units). The light of the spherical coma comes mainly from molecular fragments that have been produced by the dissociation of unobserved "parent molecules" in a zone on the order of 10^4 kilometres around the nucleus. This also is the approximate size of the zone where molecular collisions continue to occur; beyond that zone, the gas becomes too rarefied for such interaction to occur. The zone simply expands radially without molecular collisions into the vacuum of space. The parent molecules (*e.g.*, those of water vapour, carbon dioxide, and hydrogen cyanide [HCN]) are generally not observed because they do not fluoresce in visible light. So far, only a few have been observed at millimetre or centimetre wavelengths by radio telescopes; many more are needed if they are to be regarded as the source of the various radicals and ions that have been detected (Table 26).

If the mixture of original parent molecules has been frozen out of thermodynamic equilibrium in the nuclear ices, many chemical reactions can still take place in the molecular collision zone. At the usually cold temperature of vaporization, the kinetics of fast ion-molecular reactions would prevail. The reactions might reshuffle the original molecules present in the nucleus into new parent species, which would be the ones subsequently photodissociated into observed fragments by solar light. (This complex situation is still far from being completely understood.) In turn, the observed fragments, after having absorbed and reemitted photons from the solar light several times, would photodissociate or photoionize, which make them disappear from sight at the fuzzy limit of the light-emitting coma (typically 2–5 × 10^5 kilometres). A composite list of all observed species in cometary comas and tails is given in Table 26. It is based mainly on observations of the bright comets of the 1960s, '70s, and '80s, including spacecraft results from Comet Halley.

Table 26: Observed Chemical Species in Comets

Organic	C, C_2, C_3, CH, CN, CO, CO_2, CS, HCN, CH_3CN, HCO, H_2CO
Inorganic	H, NH, NH_2, O, OH, H_2O, S, S_2, NH_3, NH_4
Metals	Na, K, Ca, V, Mn, Fe, Co, Ni, Cu
Ions	C^+, CH^+, CO^+, CO_2^+, N_2^+, O^+, OH^+, H_2O^+, H_3O^+, S^+, S_2^+, H_2S^+, CS_2^+,
Dust	silicates, organic compounds

The organic radicals given in the Table were seen in cometary heads as visual or ultraviolet emission lines or bands. The exceptions were water vapour, along with hydrogen cyanide and methyl cyanide (CH_3CN); these species, which could be called parent molecules, were observed as pure rotation lines at radio frequencies. The metals—except for sodium (Na), which is observed in many comets—were seen as visual lines in Sun-grazing comets alone. They are assumed to result from the vaporization of dust grains by solar heat. Sodium is a volatile metal that is not unlikely to vaporize easily from dust grains at large distances from the Sun (more than 1 a.u.). The ions were seen in the visual or ultraviolet emission lines or bands at the onset of the plasma tail or detected by spacecraft. The silicate signature was found in infrared emission bands at the onset of dust tails. The occurrence of the silicate elements, as well as the presence of a rather large amount of organic compounds, was confirmed by the mass spectrometric analysis of dust grains during the *Giotto* flyby of Comet Halley.

Swings effect

An extremely weak coma appeared in 1984 when Comet Halley still was 6 a.u. from the Sun; this is the best case study to date. Rarely have comas been detected beyond 3 or 4 a.u., where they are still quite small; they grow to a maximum near 1.5 a.u. and seem to contract as they approach closer to the Sun. This effect comes from the more rapid decay in solar light (by photoionization or photodissociation) of the visible radicals that emit the coma light. The discrete emission of light by cometary atoms, radicals, or ions is due to the selective absorption of sunlight followed by its reemission either at the same wavelength (resonance) or at a different wavelength (fluorescence). In 1941, Pol Swings explained the peculiar appearance of some of the molecular bands in comets by the irregular spectral distribution of the exciting solar radiation owing to the presence of Fraunhofer lines (dark, or absorption, lines) in this radiation. The temporal variations that occur in the molecular bands as a comet approaches the Sun were explained quantitatively by the variable shift in the apparent wavelengths of the solar Fraunhofer lines due to the variable radial velocity of the comet. This is the so-called Swings effect. Later, the American astronomer Jesse Greenstein explained, by a differential Swings effect, the observed differences in the molecular bands in front of and behind the nucleus: the radial expansion velocity of the coma introduces a different shift forward and backward. This differential Swings effect is often referred to as the Greenstein effect (Figure 47).

Exceptions to the resonance-fluorescence mechanism are known and are exemplified by the case of the emission of the "forbidden" red doublet of atomic oxygen at wavelengths of 6300 and 6364 angstroms. Such an emission cannot be excited by direct absorption of sunlight but is produced directly by the photodissociation of H_2O into $H_2 + O$ (in the ^1D state) and, in an accessorial manner, of CO_2 into $CO + O$ (in the ^1D state). The ^1D state is an excited state of the oxygen atom that decays spontaneously into the ground state by emitting the forbidden red doublet, provided that it had not been quenched earlier by molecular collisions.

Halo of atomic hydrogen

The large atomic hydrogen halo detected up to 10^7 kilometres from the nucleus is nothing more than a large coma visible in ultraviolet (Lyman-alpha line). It is two orders of magnitude larger than the comas that can be seen in visible light only because the hydrogen atoms, being lighter, move radially away 10 times faster and are ionized 10 times more slowly than the other radicals.

Cometary tails. The tails of comets are generally directed away from the Sun. They rarely appear beyond 1.5 or 2 a.u. but develop rapidly with shorter heliocentric distance. The onset of the tail near the nucleus is first directed toward the Sun and shows jets curving backward like a fountain, as if pushed by a force emanating from the Sun. The German astronomer Friedrich Wilhelm Bessel began to study this phenomenon in 1836, and Fyodor A. Bredikhin of Russia developed, in 1903, tail kinematics based on precisely such a repulsive force that varies as the inverse square of the distance to the Sun. Bredikhin introduced a scheme for classifying cometary tails into three types, depending on whether the repulsive force was more than 100 times the gravity of the Sun (Type I) or less than one solar gravity (Types II and III). Subsequent research showed that Type-I tails are plasma tails (containing observed molecular ions as well as electrons not visible from ground-based observatories), and Types II and III are dust tails, the differences between them being attributable to a minor difference in the size distribution of the dust grains. As a result of these findings, the traditional classification formulated by Bredikhin is no longer considered viable and is seldom used. Most comets (but not all) simultaneously show both types of tail: a bluish plasma tail, straight and narrow with twists and nods; and a yellowish dust tail, wide and curved, which is often featureless (see Figure 48).

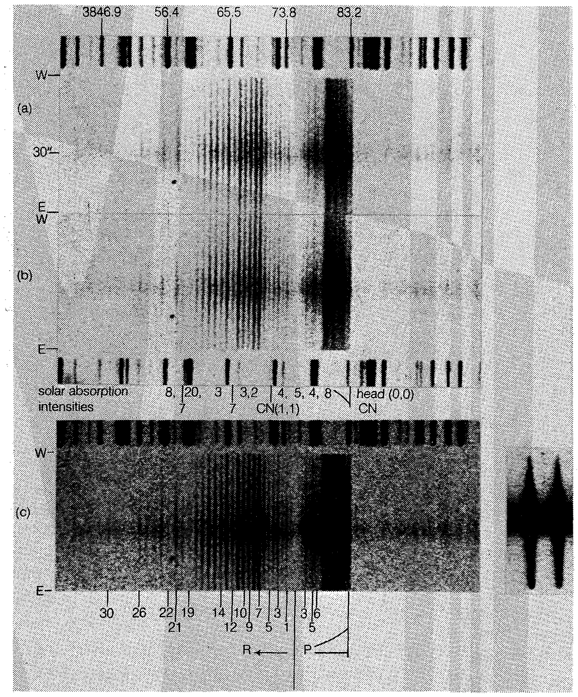


Figure 47: High-dispersion spectrum of the cyanogen (CN) bands near 3880 angstroms in Comet Mrkos, showing band irregularities (Swings and Greenstein effects; see text).

By courtesy of J.L. Greenstein, Palomar Observatory, California Institute of Technology

Plasma tail

The plasma tail has its onset in a region extremely close to the nucleus. The ion source lies deep in the collision zone (typically 10^3 kilometres). It is likely that charge-exchange reactions compete with the photoionization of parent molecules, but the mechanism that produces ions is not yet quantitatively understood. In 1951 the German astronomer Ludwig Biermann predicted the existence of the solar wind (see above) in order to account for the rapid accelerations observed in plasma tails as well as their aberration (*i.e.,* deviation from the direction directly opposite to the Sun). The cometary plasma is blown away by the magnetic field of the solar wind until it reaches its own velocity—nearly 400 kilometres per second. This action explains the origin of the large forces postulated by the Bessel–Bredikhin theory. Spectacular changes observed in the plasma tail, such as its sudden total disconnection, have been explained by discontinuous changes in the solar wind flow (*e.g.,* the passage of magnetic sector boundaries).

In 1957 the Swedish physicist Hannes Alfven predicted the draping of the magnetic lines of the solar wind around the cometary ionosphere. This phenomenon was detected by the International Cometary Explorer spacecraft, launched by the National Aeronautics and Space Administration (NASA), when it passed through the onset of the plasma tail of Comet P/Giacobini-Zinner on Sept. 11, 1985. Two magnetic lobes separated by a current-carrying neutral sheet were observed as expected. A related feature known as the ionopause was detected by the *Giotto* space probe during its flyby of Comet Halley in 1986. The ionopause is a cavity without a magnetic field that

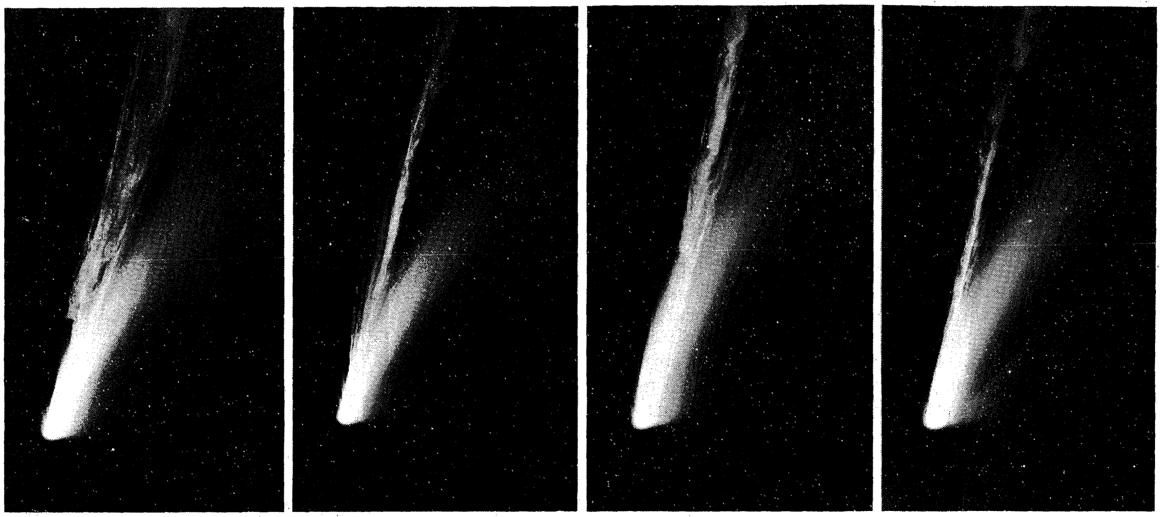


Figure 48: Comet Mrkos, photographed Aug. 22, 24, 26, 27, 1957. The straight tail with prominent streamers and irregularities, best seen at upper left in all photographs, is formed of ionized molecules; the more uniform, curved tail is composed primarily of small solid particles.

By courtesy of Mount Wilson Observatory

contains only cometary ions and is separated from the solar wind by a sharp discontinuity. Halley's ionopause lies about 4,000 to 5,000 kilometres from the nucleus of the comet. An analysis of all the encounter data indicates that a complete understanding of cometary interaction with the solar wind has not yet been achieved. It is well understood, however, that the neutral coma remains practically spherical. The solar wind is so rarefied that there are no direct collisions of its particles with the neutral particles of the coma, and, as these particles are electrically neutral, they do not "feel" the magnetic field.

Dust tail

The source of the dust tail is the dust dragged away by the vaporizing gases that emanate from the active zones of the nucleus, presumably from vents like those observed on Comet Halley's nucleus (Figure 46). The dust jets are first directed sunward but are progressively pushed back by the radiation pressure of sunlight. The repulsive acceleration of a particle varies as $(sd)^{-1}$ (with linear size s and density d). For a given density, it thus varies as s^{-1}, separating widely the particles of different sizes in different parts of the tail. Studying the dust tail isophotes of varying brightnesses therefore yields the dust grain distribution. This distribution may peak for very fine particles near 0.5 micrometre (μm), assuming a density of two, as in the case of Comet Bennett; however, it falls off with s^{-n} (with n ranging from three to five) for larger particles. This mechanism neglects particles much smaller than the mean wavelength of sunlight. Because such particles do not reflect light, they do not feel its radiation pressure. (They are not detected from ground-based observations anyway.)

One of the major results of the *Giotto* flyby of Halley's nucleus was the detection of abundant particles much smaller than the wavelength of light, indicating that the size distribution does not peak near 0.5 μm but seems rather to grow indefinitely with a slope close to a^{-2} for finer and finer particles down to possibly 0.05 μm (10^{-17} gram). The dust composition analyzers on board the *Giotto* and Vega spacecraft revealed the presence of at least three broad classes of grains. Class 1 contains the light elements hydrogen, carbon, nitrogen, and oxygen only (in the form of either ices or polymers of organic compounds). The particles of class 2 are analogous to the meteorites known as CI carbonaceous chondrites but are possibly slightly enriched in carbon and sulfur. Class 3 particles are even more enriched in carbon, nitrogen, and sulfur; they could be regarded as carbonaceous silicate cores (like those of class 2) covered by a mantle of organic material (similar to that of class 1) that has been radiation-processed. Most of the encounter data were excellent for elemental analyses but poor for determining molecular composition, because most molecules were destroyed by impact at high encounter velocity. Hence, there still remains much ambiguity regarding the chemical nature of the organic fraction present in the grains.

Meteors are extraterrestrial particles of sand-grain or small-pebble size that become luminous upon entering the upper atmosphere at very high speeds. Meteor streams have well-defined orbits in space. More than a dozen of these orbits have practically the same orbital elements as the orbits of the identical number of short-period comets (see below *Orbits of meteoroids and meteor showers*). Fine cometary dust consists primarily of micrometre- or submicrometre-sized particles that are much too small to become visible meteors (they are more like cigarette smoke than dust). Moreover, they are scattered in the cometary tail very far from the comet orbit. The size distribution of cometary dust grains, however, covers many orders of magnitude; a small fraction of them may reach 0.1 millimetre to even a few centimetres. Because of their large size, the dust grains are almost not accelerated by the radiation pressure of sunlight. They remain in the plane of the cometary orbit and in the immediate vicinity of the orbit itself, even though they separate steadily from the nucleus. They sometimes become visible as an anti-tail—*i.e.,* as a bright spike extending from the coma sunward in a direction opposite to the tail (Figure 49). This phenomenon occurs as a matter of geometry: it takes place for only a few days when the Earth crosses the plane of the cometary orbit. At such a time, this plane is viewed through the edge and all large grains are seen accumulated along a line. The same grains scatter farther and farther away from the nucleus until some are along the entire cometary orbit. When the Earth's orbit intersects such an orbit (an event that occurs year after year at the same calendar date), these large grains produce meteor showers.

Anti-tail

Extremely fine cometary grains also may penetrate the Earth's atmosphere, but they can be slowed down gently without burning up. Some have been collected by NASA's U-2 aircraft at very high altitudes. Grains of this kind are known as Brownlee particles and are believed to be of cometary origin (Figure 50). Their composition is chondritic, though they show somewhat more carbon and sulfur than the CI carbonaceous chondrites, and their structure is fluffy with many pores. Similar grains were found in space during the space probe exploration of Comet Halley.

COMETARY MODELS

As previously noted, the sandbank model of the cometary nucleus fell into disregard by the late 1950s and early 1960s and was supplanted by the dirty snowball (or icy conglomerate) concept. Much circumstantial evidence supported the latter, but confirmation was lacking until 1986 when the *Giotto* spacecraft returned detailed, close-up photographs of Comet Halley's nucleus. Yet, while

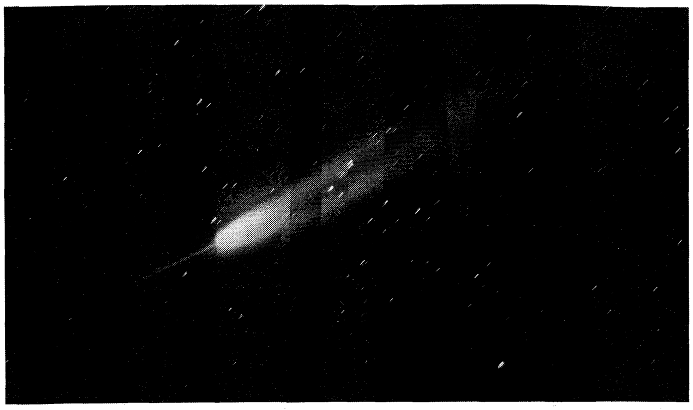


Figure 49: Comet Arend-Roland photographed on April 25, 1957. The prominent anti-tail extending from the coma appears to precede the comet, though it actually trails from behind.

By courtesy of Lick Observatory, University of California

these photographs corroborated the general idea of the model, they revealed that "dirty snowball" was in fact a misnomer because snow (even when dirty) is suggestive of something white or at least gray in colour. In actuality, the cometary nucleus proved to be pitch-black owing to the large amount of very fine, black sootlike particles intermixed with the volatile ices (see above).

Many variations of the icy conglomerate model have been proposed since the early 1980s, as, for example, the fractal model, rubble-pile model, and icy-glue model. These names, however, suggest only slightly different types of accretion of primordial particles; they all share common features—namely, irregular shape, heterogeneous mixture, and very low density because of cavities and pores. The existence of a crust or dust mantle of a different nature had already been proposed before the 1986 spacecraft encounter with Comet Halley for two reasons. First, cosmic-ray processing of the outer layers had been described by L. Shul'man of the Soviet Union (1972) and later advocated by Fred Whipple and Bertram Donn of the United States, while the outgassing of the outer layers by solar heat had also been assumed since the proposal of Whipple's model (1950). Second, detailed models of the formation and disruption of such mantles due to solar-radiation processing of the upper layers had been studied by Devamitta Asoka Mendis of the United States (1979) and M. Horanyi of Hungary (1984).

Elemental abundances of the cometary nucleus

An average heuristic model for the elemental abundances of the cometary nucleus was developed by the American astronomer Armand H. Delsemme in 1982. Delsemme computed the H:C:N:O:S ratios from ultraviolet and visual observations of atomic and molecular species in bright comets detected during the 1970s and deduced the abundances of metals from the chondritic composition of cometary dust. In this model, hydrogen was depleted by a factor of 10^3 with respect to solar or cosmic abundances, and carbon was depleted by a factor of four in the gaseous fraction. The results of the 1986 study of Comet Halley confirmed the average chemical model and showed that the carbon missing in the gas was actually present in the dust. Except for hydrogen (and presumably helium), it appears that all elements are roughly in cosmic proportions in comets in spite of their extremely low gravity (10^{-4} times that of the Earth). This emphasizes the pristine nature of comets. Unlike most bodies of the solar system, comets obviously have never been severely processed by any heating episode since their formation. If the accretion of comets occurred at very low temperatures, near absolute zero (0 K), the water ice in a newly formed comet must be amorphous. Idealized models show that the transition to cubic ice might be the cause of sudden flare-ups between 3 and 6 a.u.

By courtesy of D. Brownlee, University of Washington; photograph, M. Wheelock

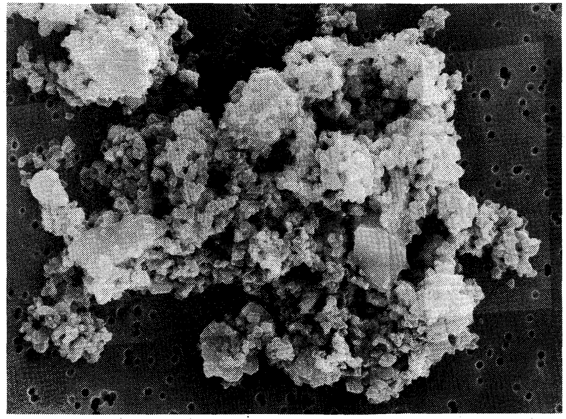


Figure 50: Electron micrograph of chondritic interplanetary dust particle (18.3 micrometres in width) of possible cometary origin. The particle was collected in the Earth's atmosphere by a NASA U-2 research plane.

ORIGIN AND EVOLUTION OF COMETS

Evolution of orbits

All observed comets make up an essentially transient system that decays and disappears almost completely in less than 1,000,000 years. Since they all pass through the solar system, planetary perturbations eject a fraction of them into deep space on hyperbolic orbits and capture another fraction on short-period orbits. In turn, those that have been captured decay rapidly in the solar heat. Fortunately, there is a permanent source of new comets that maintains the steady state—namely, the outer margin of the Oort cloud. As explained above, these so-called new comets are those Oort-cloud comets whose perihelia have been brought down into visibility—*i.e.,* into the inner planetary system where they display their spectacular decay through comas and tails. Comets within the bulk of the Oort cloud are unobservable, not only because they do not develop comas and tails but also because they are too far away.

Formation of the Oort cloud. Any modern theory about cometary origins must first explain the origin of the Oort cloud. None of the comets observed today left the Oort cloud more than 3,000,000 or 4,000,000 years ago. The Oort cloud is, however, gravitationally bound to the solar system, which it follows in its orbit around the Milky Way Galaxy. Therefore, it is likely that the Oort cloud has existed for a long time. The most probable hypothesis is that it was formed at the same time as the giant planets by the very process that accreted them. The Soviet astronomer V.S. Safronov developed this accretionary theory of the planetary system mathematically in 1972. According to his model, the planets originated from a disk or a ring of dust around the Sun, and cometary nuclei are nothing more than primordial planetesimals that accreted first and became the building blocks of the planets. From the accreted mass of the giant planets, Safronov predicted the correct order of magnitude of the mass of the Oort cloud, which was built up by those planetesimals that missed colliding with the planetary embryos and were thrust far away by their perturbations. In effect, the Oort cloud in this theory becomes the necessary consequence and the natural by-product of the accretion of the giant planets.

The Oort cloud as a by-product of the accretion of the giant planets

Later in the 1970s the American astronomer A.G.W. Cameron developed a much more massive model of the protostar nebula, in which the comets accreted in a circular ring at some 1,000 a.u. from the Sun, which is far beyond the present limits of the planetary system. The primeval circular orbits were then transformed into the elongated ellipses present in the Oort cloud by mass loss of the primitive solar nebula. Both the Cameron and Safronov models put the origin of comets together with that of the solar system some 4,500,000,000 years ago. Plausibility is given to the general idea of accretion from dust disks by the existence of such disks around many young stars—a fact established by infrared observations in the 1980s and confirmed visually in at least one case (β Pictoris). Further support is found in clues derived from meteorites.

Since the early 1980s, new ideas have been explored to determine whether the Oort cloud could be much younger than the solar system or at least periodically replenished. The role of the massive and dense molecular clouds that exist in interstellar space has been reexamined in different ways. Could comets have accreted in these clouds directly from interstellar grains? Mechanisms for later capturing them into the Oort cloud cannot be very effective, but the efficiency is not capital and some possibilities have been proposed. Since the solar system itself was probably formed from the gravitational collapse of such a molecular cloud, it seems more likely that either comets or the interstellar grains that were going to accrete into comets followed suit during gaseous collapse and were put into the Oort cloud at the same time that the planets were being formed. Elemental isotopic ratios deduced from the Comet Halley flyby have not brought about any conspicuous anomalies that could be attributed to matter coming from outside the solar system. So far, observational clues all favour the idea of cometary matter deriving from the same primeval reservoir as the stuff of the solar system, but it must be recognized that the evidence remains weak.

Possible pre-solar-system origin of comets. Telltales based on the chemical constitution of cometary nuclei as well as on the evolution of their orbits suggest that the origin of comets goes back beyond that of the planets and their satellites. Two scenarios are among the likeliest possibilities. In the first, comets had already accreted in all dense molecular clouds of the Milky Way Galaxy by the agglomeration of interstellar grains covered by a frost of organic molecules that cemented them together. Later, such a cloud collapsed to form the solar system. In the second scenario, dense molecular clouds were not able to accrete their frosty interstellar grains into larger bodies. When one of these molecular clouds collapsed to form the future solar system, however, the interstellar grains did likewise and eventually formed a dusty disk around the central star—the proto-Sun. Accretion into objects of 10-kilometre diameter is more likely in dusty disks of this type. The outer grains of the disk had not lost their frost, and some of them were ejected into the Oort cloud during the accretion of planetesimals into giant planets after some very moderate processing by heat. It is hoped that one day, space probes will secure data that will make it possible to determine whether frosty interstellar grains have lost their identity or can still be recognized as pristine and unaltered objects in cometary dust.

Comets seem to be the most pristine objects of the solar system, containing intact the material from which it was formed. Included are the hydrogen, carbon, oxygen, nitrogen, and sulfur atoms needed to build the volatile molecules present in the terrestrial biosphere (including the oceans and the atmosphere). Comets also seem to be the link between interstellar molecules and the most primitive meteorites known—the carbonaceous chondrites (see below *Types of meteorites*). The molecules required to initiate prebiotic chemistry (*e.g.,* hydrogen cyanide, methyl cyanide, water, and formaldehyde) are present in interstellar space just as they are in comets; larger prebiotic chemistry molecules (*e.g.,* amino acids, purines, and pyrimidines) occur in some chondrites and possibly in comets. An early cometary bombardment of the Earth, predicted in some accretion models of the solar system, may have brought the oceans and the atmosphere, as well as a veneer of the molecules needed for life to develop on the Earth. Comets could well be the link between interstellar chemistry and life. (A.H.De.)

Meteoroids, meteors, and meteorites

METEOROIDS

General considerations. The solar system contains many small bodies that move in orbits sufficiently eccentric to cross over and intersect the orbit of the Earth. When their orbits do intersect that of the Earth, the probability of the objects colliding with the planet becomes quite high. A body of this kind entering the Earth's atmosphere is called a meteoroid. Such bodies range from small particles less than one micrometre in size (about 10^{-12} gram in mass) up through objects several centimetres or metres in diameter and grade into kilometre-sized bodies large enough to be observed as astronomical objects through telescopes. They enter the upper atmosphere with velocities of 11 to 72 kilometres per second. Interaction with the atmosphere heats incoming objects that are larger than about 10^{-2} centimetre to temperatures high enough to cause them to become incandescent, vaporize, and heat the surrounding air. As a result of this sequence, such objects are observable from the ground as meteors or, in more popular language, "shooting stars" or "falling stars." Strictly speaking, the term meteor refers only to the phenomena associated with the collision of a meteoroid with the Earth's atmosphere. Scientific usage is not all that strict, however, and the body itself is often called a meteor. Unusually luminous meteors are termed fireballs or bolides.

'ariations
ı size

Due to its fairly low entry velocity, large mass, and physical strength, a meteoroid sometimes survives its passage through the atmosphere, falls to the ground, and may be recovered as a meteorite. Recovered meteorites, ranging in mass from a few grams to several tons, are often exhibited in museums. Most meteorites either consist of rocky—chiefly silicate—material (stony meteorites) or are composed primarily of nickel-iron alloy (iron meteorites; Figures 51 and 52). In stony-iron meteorites, massive nickel-iron alloy is intermixed with silicate material.

In addition to these relatively large meteorites, it is possible to recover much smaller objects about 10^{-3} centimetre in diameter called micrometeorites on filters attached to aircraft flying in the stratosphere. These micrometeorites (often referred to as interplanetary or cosmic dust) also accumulate on the bottom of the deep ocean. The larger ones can be identified and separated from cores drilled from the muddy deposits on the seafloor.

Micrometeorites

The largest meteoroidal bodies are observable through telescopes as astronomical objects. These include the Apollo objects, bodies of asteroidal appearance with diameters ranging from a few hundred metres to several kilometres that come closer to the Sun than to the Earth's orbit. There are about 700 of these larger than one kilometre in diameter. Because comets can strike the Earth, they, too, can be thought of as large meteoroidal bodies.

When meteoroids are sufficiently large (*i.e.,* 100 metres

(Top left) J.A. Wood; (others) Smithsonian Institution

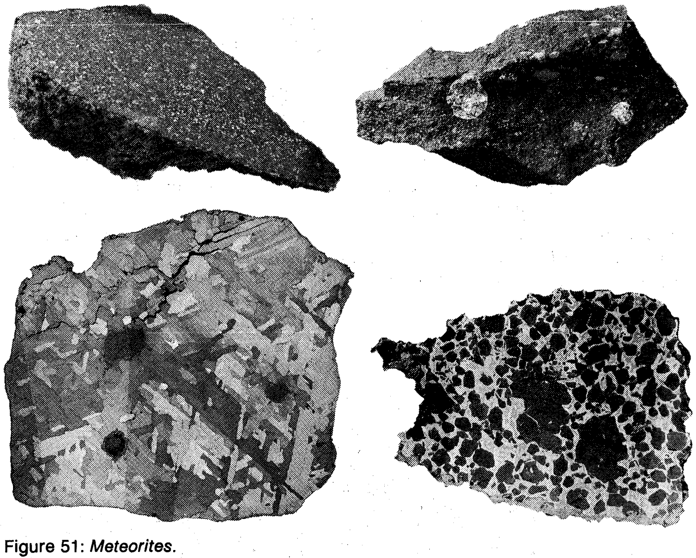


Figure 51: *Meteorites.*
(Top left) Ankober (Ethiopia) ordinary (H4 olivine-bronzite) chondrite (reduced). One surface has been sawed and polished, revealing the internal structure. The white spots are reflections from nickel-iron metal; the surrounding gray material is composed of silicate minerals. (Top right) Allende carbonaceous (CV3) chondrite (reduced 3.8 ×). The large white inclusions consist primarily of refractory aluminum- and calcium-bearing minerals, which are embedded in a dark gray matrix containing fine-grained minerals formed at much lower temperatures. (Bottom left) Sawed, polished, and acid-etched interior surface of the Osseo (Ontario) iron meteorite (coarsest octahedrite; reduced 4.2 ×). This treatment reveals the "Widmanstätten pattern" resulting from coarsely crystalline kamacite (α-iron-nickel alloy). (Bottom right) Polished and etched section of the Salta stony-iron (pallasite), composed of roughly equal amounts of olivine (magnesium-iron silicate in the form of dark grains) and nickel-iron alloy (the shiny crystals; reduced 4.1 ×).

to several kilometres in diameter), they can pass through the atmosphere without slowing down appreciably. When they strike the Earth's surface at velocities of many kilometres per second, the kinetic energy released is sufficient to produce an impact crater. In many ways such craters resemble those produced by nuclear explosions. They are often called meteorite craters, in spite of the fact that the impacting meteoroids themselves are almost entirely vaporized during the explosion. High-velocity impact by objects of this kind on the Moon, Mercury, and Mars are in large part responsible for the heavily cratered appearance of the surface of these bodies. The cratering record on the Earth and Moon also shows that there are many meteoroids in the intermediate mass range between the larger recovered meteorites (a few metres in diameter) and the Apollo objects.

G. Kurat

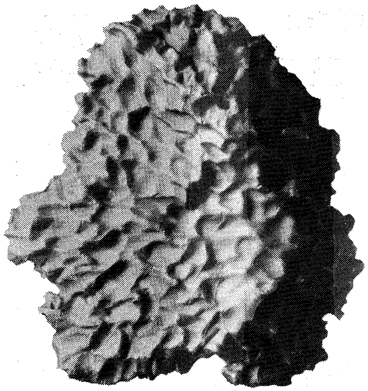


Figure 52: Cabin Creek (Arkansas) iron meteorite (medium octahedrite; reduced 8.1 ×). The dimpled appearance of the face of the meteorite was caused by melting of its surface upon entering the Earth's atmosphere.

Relationship of meteoroids to asteroids and comets. Most of the mass of the solar system resides in its larger bodies, the Sun and the planets. The planets move about the Sun in stable and well-separated orbits. It is almost certain that these orbits and thus the positions of the planets have undergone only minor changes since the formation of the solar system some 4,500,000,000 years ago. In addition, the planets are large enough to retain on their surfaces nearly all the debris produced by impact craters.

On the other hand, a smaller fraction of the mass of the solar system is found in bodies of such small size or in orbits so eccentric that their physical survival or orbital stability has been in jeopardy throughout the history of the solar system. Most of these bodies are found in either of two regions of the solar system: the asteroid belt, between the orbits of Mars and Jupiter (mostly between 2.2 and 3.4 a.u.), and the cometary, or Oort, cloud extending far beyond the orbits of Neptune and Pluto to distances of more than 10^4 a.u.

The members of the asteroid belt have high enough eccentricities and inclinations that they collide with one another at velocities averaging about five kilometres per second. Because of this, it is unlikely that most asteroids larger than about 75 kilometres in diameter have survived collisional destruction over the entire 4,500,000,000-year history of the solar system. The present-day smaller asteroids are debris formed by fragmentation of larger asteroids caused by this natural grinding process in the asteroid belt. The grinding extends down to bodies the size of meteorites exhibited in museums and even to much smaller particles in the form of fine dust.

Fragmentation of asteroids

Some of this collisional debris is orbitally unstable. The same collisions that produced the fragments propel a portion of them into chaotic orbits controlled by dynamical resonances with the motions of the planets. As a result of collisions and orbital instability, asteroidal material is injected into orbits that cross the Earth's orbit. Some of this matter collides with the atmosphere of the Earth and becomes visible meteors. A small number of the particles survive as meteorites.

The much more distant cometary bodies also are vulnerable to destruction. Their orbits extend to the threshold of interstellar space. Passing stars and interstellar molecular clouds gravitationally perturb some of these bodies into orbits with perihelia within several astronomical units of the Sun. A small fraction of the comets that are perturbed into orbits approaching the Sun experience close approaches to planets, particularly Jupiter, and this leads to further orbital perturbations. In some cases, this causes the orbital eccentricity and orbital period of a comet to be drastically reduced to periods from about three to several hundred years.

Comets are small bodies consisting of a mixture of ice and less volatile material, the latter mostly in the form of dust (see above *The nature of comets*). When a comet is perturbed into an orbit that comes within a few astronomical units of the Sun, the ice begins to evaporate, giving rise to the luminous object commonly thought of as a comet. Dust is swept away from the comet along with the evaporating ice. Most of the fine dust escapes the Sun's gravitational field and is lost to interstellar space. Some of the dust, however, is in larger particles typically millimetres to centimetres in size. As compared with the finer particles of dust, these larger bodies have a smaller ratio of surface area to mass, and this renders them less susceptible to the solar radiation forces, proportional to the cross-sectional area, that drive the finest dust out of the solar system. Thus, these larger cometary particles can remain in short-period orbits and can intersect the orbit of the Earth. When they collide with the atmosphere, meteoroids of this kind also appear as meteors.

The best evidence for the cometary origin of certain meteoroids is identification of the orbits of meteoroids with those of known periodic comets (see above *Modern cometary research*). It is possible to determine these orbits by photographing the meteors with special cameras as they pass through the Earth's atmosphere. This identification has been made for several of the meteors that produce meteor showers. In this way inferences can be made about the physical properties of particulate cometary material.

Associatior of meteor showers with meteoroids of cometary origin

The asteroid belt and the outermost part of the solar system should be thought of as long-lived though somewhat leaky reservoirs of meteoroidal bodies. They are long-lived enough to retain a significant quantity of primordial solar system material but leaky enough to permit the escape of the observed quantity of Earth-crossing material. This quantity of Earth-crossing material represents an approximate steady-state balance between the input from the storage regions and the loss by ejection from the solar system, collision with the Earth, the Moon, and other planets, or vaporization by impact.

Evidence for meteoroids of asteroidal origin: radiometric ages of meteorites. The most meaningful way of describing and classifying meteoroids would be in terms of their various sources in the solar system. To do so, it is necessary to find criteria that discriminate among these sources.

Unlike the association of some meteoroids with the actual observed orbits of comets, direct orbital identification is not possible for meteoroids of asteroidal origin. The asteroids in the asteroid belt that are the ultimate parents of these meteoroids are not in Earth-crossing orbits. A meteoroid entering the atmosphere must be in an Earth-crossing orbit. Therefore, the observed meteoroid orbit cannot be the same as that of its source. The evolution of the asteroidal collision debris from the asteroid belt is complex, involving collisional fragmentation and major orbital changes caused by planetary gravitational perturbations. Although this evolution is quantitatively understandable in terms of known physical processes, the complex history destroys the information required to directly link the meteoroid with its source. Unlike cometary meteors, there is, for example, no good evidence that asteroidal meteors occur in streams, nor would such clustering of orbits be expected on theoretical grounds. The link between asteroidal meteoroids and their sources must be established in ways that are less direct.

This link is supplied by laboratory investigations of meteoroids that have survived passage through the atmosphere—namely, meteorites. As a result of these investigations, it is generally believed that nearly all recovered meteorites are fragments of asteroids. A small fraction, about 0.5 percent, may be of lunar or Martian origin. It is possible, but not at all firmly established, that some meteorites may be derived from comets.

The reasoning that leads to identification of most meteorites as asteroidal meteoroids is fairly complex but quite compelling. Perhaps the best starting point would be that based on studies of the daughter isotopes produced by the radioactive decay of radioactive parent isotopes of such elements as uranium, thorium, rubidium, potassium, and samarium. As is commonly done for terrestrial rocks, the ages of meteorites can be determined by isotopic measurements of this kind.

Isotopic measurements

When this is done, all but a very few meteorites show good evidence of having been formed, from the mineralogical point of view, some 4,500,000,000 years ago. Similar techniques have been used to infer, somewhat less directly, the age of the Earth and the Moon, with the same result. Within an uncertainty of less than about 100,000,000 years, this age represents the time in the past when the Sun, the planets, and their natural satellites, along with the asteroids and comets, formed from interstellar dust and gas. Because of geologic processes that produced younger rocks on the Earth and Moon, their evidence for this primordial age of formation has been obscured but not erased. By contrast, the evidence for the age of the solar system is clearly preserved in meteorites.

Although present-day meteorites are the products of a long series of events that caused physical fragmentation, in most cases the mineral grains of which they are composed have undergone little chemical alteration since the early years of the solar system. As a consequence, the isotopic ratios from which the ages are calculated preserve the evidence for this very ancient time of formation.

If meteorites are asteroidal fragments, this is what would be expected. Theoretical calculations of the thermal evolution of bodies as small as asteroids predict that they would not exhibit the long-continuing history of igneous and metamorphic rock formation characteristic of large bodies such as the Earth and Moon. As in the case of the latter, however, the decay of radioactive isotopes produces heat in the interiors of asteroids. Yet, because of their small size, this heat is readily conducted to the surface and radiated to space. The internal temperature thus would never rise high enough to produce the metamorphism or melting that would chemically alter the ancient mineralogy. In the much larger Earth, on the other hand, the heat is not efficiently transported to the surface by convection until the temperature increases to near its melting point.

By itself, this preservation of ancient mineralogy does not, however, clearly indicate that the meteorites are of asteroidal rather than of cometary origin. Comets are even smaller bodies than asteroids and would be heated even less. If, in addition to the volatile material and fine dust emitted by comets as they approach the Sun, there exists ancient rocky material in their nuclei, one would expect the primordial isotopic age record to be preserved in cometary meteoroids as well.

Distinuishing etween steroidal nd ometary ources

The clue to the distinction between asteroidal and cometary sources for at least the most common meteorites is provided by minor mineralogical disturbances that slightly alter the otherwise clear preservation of the primordial isotopic age record in meteorites. Such disturbances are attributable to metamorphism associated with evidence for collisional shocks preserved in the detailed mineralogy of meteorites. For example, many representatives of one very common type of chondritic meteorite, the hypersthene chondrite, display glass veins and other evidence of shock-induced metamorphism of the sort that would be expected to result from collisions between asteroids. (Chondrites, the most common type of stony meteorite, are so called because they contain small primordial silicate spherules known as chondrules.) Ages measured by the decay of potassium to argon show that this shock metamorphism took place quite late in the history of the solar system—about 500,000,000 years ago. Ages measured by other isotopic decay systems on these same meteorites still display the more common value of 4,500,000,000 years (Figure 53). Furthermore, these shocked meteorites exhibit a very strong chemical and textural kinship to unshocked meteorites that record completely the isotopic age of the solar system. It is not plausible that such similar rocks should be provided by sources from regions as far removed from one another as the asteroid belt and the cometary cloud. Other evidence for impact events late in solar system history is provided by dating shocked fragments of meteoritic material embedded in meteorites of different types which are presumably a result of collisions in space. Quantitative comparison of the collisional history expected for the sparsely populated, low relative velocity source regions of comets with that expected for the more densely populated, higher velocity asteroid belt argues strongly for the asteroid belt being the actual site of these collisions and thus the source region for the meteorites.

(Left) Based on data from J.F. Minster (1979) and a series of papers by K. Gopalan, S.K. Kaushal, and G.W. Wetherill in *Journal of Geophysical Research* (1968–71), copyright by the American Geophysical Union. (Right) Data compiled by J.T. Wasson (1985) based on data from D. Bogard, *Journal of Geophysical Research*, vol. 81, p. 5664, (1976), copyright by the American Geophysical Union; G. Turner, *Meteorite Research* (1969); and G. Turner *et al.*, *Proceedings of the Ninth Lunar Planetary Scientific Conference* (1979)

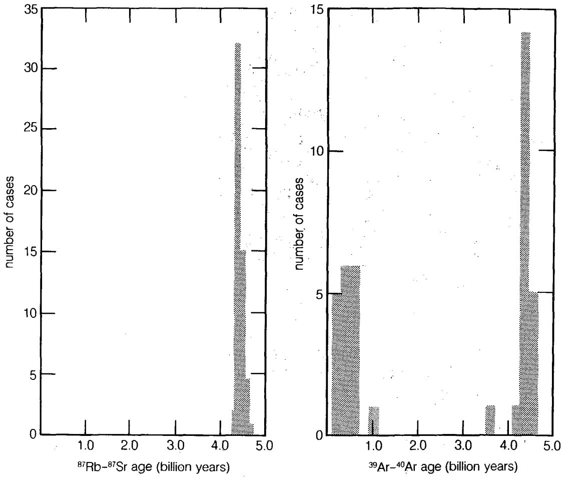


Figure 53: *Distribution of formation and metamorphism ages of ordinary chondrite meteorites.*
(Left) Ages measured by the decay of ^{87}Rb to ^{87}Sr. The "isotopic clocks" of these meteorites were set by heating events essentially at the beginning of solar system history.
(Right) Ages based on the decay of ^{40}K to ^{40}Ar. Many of these ages also cluster around 4,500,000,000 years, but a significant number have been "reset" by collisionally induced shock metamorphism on the asteroid from which the meteorite was fragmented.

Such conclusions are supplemented by observational and theoretical dynamical studies of meteorite orbits. The results of these studies are in agreement with an asteroidal origin for at least the most abundant types of meteorites but are not in agreement with a cometary origin.

Meteoroids of less certain origin. Identification of many meteoroids with either asteroids or comets can be securely established. It is then possible to extend these identifications by extrapolation to cases where the distinction is less obvious. There are, for example, many shower meteors in meteor streams for which no comet is presently identifiable; however, the orbits and characteristics of their passage through the atmosphere are very similar to those of stream meteors that are associated with known comets. Even when such meteors are not in streams, their orbits with aphelia and physical characteristics similar to those of cometary stream meteors leave little room to question their cometary parentage. It also has been possible to identify a number of very bright meteors (fireballs) with asteroidal meteorites by comparing the details of their atmospheric flight with those of recovered meteorites (see below *Fireball networks*).

There are many meteoroids whose asteroidal or cometary nature remains highly uncertain. These include objects that are presently in orbits very similar to identifiably asteroidal meteoroids but that exhibit in their atmospheric flight the greater fragility associated with meteoroids known to be derived from comets. Are these fragments of comets that have been perturbed into orbits traditionally associated with asteroids, or do they represent a type of asteroidal material that is too weak to survive as a meteorite? Answers to such questions are needed if scientists are to use the information preserved in primitive bodies—comets and asteroids—in the most effective way. There is much more that can be learned from further observational studies of meteors, laboratory investigations of meteorites, Earth-based observations of comets and asteroids, and spacecraft missions to these small primitive bodies. Continuation of this work holds much promise for clarifying the processes by which both the Earth and the solar system came into existence.

In addition to cometary and asteroidal sources, a very minor but identifiable contribution to the meteoroidal population is derived from the Moon. Such particles are ejected from primary cratering of the lunar surface by large meteorites. By comparison with lunar rocks returned by the U.S. Apollo missions, several small meteorites collected on the Antarctic ice sheet have been positively identified as lunar meteorites. It also appears probable that certain relatively rare types of meteorites (shergottites, nakhlites, and chassignites) are fragments of igneous rocks from Mars. Some day evidence for meteoroids from Mercury, Venus, Earth, and the interstellar medium may be identified.

METEORS

Characteristic observational effects. On any clear night in the countryside beyond the bright lights of cities, one can observe with the naked eye several meteors (or shooting stars) per hour as they streak through the sky, with durations ranging from a small fraction of a second up to several seconds. Quite often they vary in brightness along the path of their flight, appear to emit "sparks" or flares, and sometimes leave a luminous train that lingers after their flight has ended. These meteors are the result of the high-velocity collision of meteoroids with the Earth's atmosphere. Nearly all such interplanetary bodies are small fragments derived from comets or asteroids.

The observed apparent brightness of these easily observable meteors covers the same range of brightness as the stars visible to the unaided eye (*i.e.,* from about zero to fifth astronomical magnitude). They constitute a portion of an Earth-impacting interplanetary flux of similar bodies ranging in mass from less than one nanogram up to millions of tons. The smaller bodies are too faint to be seen with the naked eye but are observable with the aid of binoculars and telescopes or by radar reflection. Brighter meteors—of magnitudes ranging in brightness from that of Venus (−4 magnitude) to greater than that of the Full Moon—are less common but are not really unusual. These are produced by meteoroids with masses ranging from several grams up to about one ton.

The brightest meteor (possibly of cometary origin) for which historical documentation exists struck on June 30, 1908, in the Tunguska region of central Siberia and rivaled the Sun in brightness. The energy delivered to the atmosphere by this impact was roughly equivalent to that of a 10-megaton thermonuclear explosion and caused the destruction of forest over an area of about 2,000 square kilometres. The geologic record of cratering attests to the impact of much more massive meteoroids, including objects with kinetic energies equivalent to 100,000,000 megatons. Fortunately, impacts of this magnitude occur only once or twice every 100,000,000 years. It is hypothesized that large impacts of this kind may have played a major role in determining the course of biologic evolution by causing simultaneous mass extinctions of many species of organisms, possibly including the dinosaurs some 65,000,000 years ago. If so, the replacement of reptiles by mammals as the dominant land animals, the eventual consequence of which was the rise of the human species, would be the result of a grand example of a phenomenon observable every clear night.

Heights and velocities

The visibility of meteors is a consequence of the high velocity of meteoroids in interplanetary space. Before entering the region of the Earth's gravitational influence, their velocities range from a few kilometres per second up to as high as 72 kilometres per second. As they approach the Earth, within a few Earth radii, they are accelerated to even higher velocities by the planet's gravitational field. As a consequence, the minimum velocity with which a meteoroid can enter the atmosphere is equal to the Earth's escape velocity of 11 kilometres per second. Even at this minimum velocity, the kinetic energy of a meteoroid would be 6×10^4 joules per gram of its mass. This can be compared with the energy of about 4×10^3 joules per gram produced by chemical explosives, such as TNT. As the meteoroid is slowed down by friction with atmospheric gas molecules, this kinetic energy is converted into heat. Even at the low atmospheric density at altitudes of 100 kilometres (6×10^{-10} gram per cubic centimetre compared with 10^{-3} gram per cubic centimetre at sea level), this heat is sufficient to vaporize and dissociate and ionize the surface material of the meteoroid and dissociate and ionize the surrounding atmospheric gas as well. Electronic transitions effected by this excitation of atmospheric and meteoroidal atoms produce a luminous region, which travels with the meteoroid and greatly exceeds its dimensions. At deeper levels in the atmosphere, a shock wave may be produced in the air ahead of the meteoroid. This shock wave interacts with the solid meteoroid and its vapour in a complex way. About 0.1 to 1 percent of the original kinetic energy of the meteoroid is transformed into visible light.

Ablation and fragmentation

This great release of energy destroys meteoroids of small mass—particularly those with relatively high velocities—very quickly. This destruction is the result both of ablation (the loss of mass from the surface of the meteoroid by vaporization or as molten droplets) and of fragmentation caused by aerodynamic pressure that exceeds the crushing strength of the meteoroid. For these reasons, numerous meteors end their observed flight at altitudes above 80 kilometres, and penetration to as low as 50 kilometres is unusual. Nevertheless, some meteoroids survive to much lower altitudes due to a combination of relatively low entry velocity (< 25 kilometres per second), large mass (> 100 grams), and fairly high crushing strength (>10^7 dynes per square centimetre). Those that are recoverable as meteorites lose their kinetic energy before the meteoroid is completely destroyed. They are effectively stopped by the atmosphere at altitudes of five to 25 kilometres. Following this atmospheric braking, they begin to cool, their luminosity fades, and they fall to the Earth at low terminal velocities of 100 to 200 metres per second. This "dark flight" of the meteoroid may be several minutes in duration, in contrast to the few seconds of visible flight.

The passage of meteoroids through the atmosphere produces atmospheric shock waves that penetrate to the ground. The penetration of a meteoroid in the kilogram range to altitudes of about 40 kilometres can thereby produce sounds on the ground similar to sonic booms or thunder. Sometimes these sound waves are intense enough to become coupled to the ground and be recorded by seismometers.

The effect of the final impact with the ground of meteorites in the kilogram mass range could be considered an anticlimax. The fall can go unnoticed by those near the impact site, the impact being signaled only by a whistling sound and a thud. For this reason, many meteorites are recovered only because at least one fragment of the meteoroid strikes a house, drawing the attention of the residents to an unusual event.

Orbits of meteoroids and meteor showers. Prior to entering the gravitational field of the Earth, a meteoroidal body, like all bodies of the solar system, moves around the Sun in an elliptic Keplerian orbit. If this orbit can be determined, valuable information relevant to identifying the source—the parent body—of the meteoroid can be obtained.

When the coordinates in the sky of the trajectory of a meteor are observed from two or more well-separated stations, the direction in which the meteoroid was moving in space before it encountered the Earth can be estimated reasonably well by triangulation. This direction is called the radiant of the meteor. If the motion of the meteoroid is thought of as a velocity vector, such observations determine approximately the direction of this vector. To determine the meteoroid's orbit, however, requires ascertaining not only the direction but also the magnitude of the velocity vector. Although well-trained visual observers are able to estimate the coordinates of the meteor trajectories and thereby determine radiants fairly well, their velocity estimates have proved to be too uncertain to be useful for orbit determination.

This problem was overcome during the 1940s by the introduction of astronomical cameras specially designed for studying meteors. These wide-field cameras were equipped with a rotating shutter that periodically interrupted the light to the photographic plate. The shutter breaks permitted calculation of the speed of a meteor along its path. The position of the meteor's trajectory with respect to the stars photographed on the same plate also was measured accurately. Such observations made at two or more stations could then be used to calculate precisely the orbit of the meteoroid before it encountered the Earth. During the 1940s, special radar instruments also were applied to the study of meteors generally fainter than those observed photographically.

adiants
ˆ shower
ıeteors

"Showers" of meteors have been known since ancient times. On rare occasions, these showers are very dramatic, with thousands of meteors falling per hour. More often, the background hourly rate of roughly five observed meteors increases up to about 10–50. Shower meteors characteristically have nearly the same radiant. This means they are all moving in the same direction in space. As a consequence, plots of meteoroid trajectories on a star map converge at a single point, the radiant, for the same reason that parallel railroad tracks appear to converge at a distance. The new photographic data fully confirmed the belief that meteors belonging to a particular shower had not only the same radiant but similar orbits as well. In other words, the meteoroids producing the meteor showers move in confined streams around the Sun. The introduction of radar observation led to the discovery of several new meteor streams that were totally invisible to cameras because they came from radiants in the daytime sky.

A fact of great importance, fully confirmed by the photographic data, is the association of many meteor streams with the orbits of observed comets. A list of the more important meteor-stream orbits and associated comets is given in Table 27. The streams with cometary associations represent debris ejected from a comet along its orbit through space. A recently formed stream, the Leonids, tends to appear in great strength every 33 or 34 years, the same as the period of the parent comet, Temple-Tuttle. These meteoroids are clustered in a compact swarm moving in the orbit of the comet. With the passage of 1,000 years or so, the slightly different orbits of the meteoroids will cause them to disperse more uniformly along the orbit of the comet. In such cases, a shower, usually weaker, occurs annually when the Earth's orbit intersects the orbital plane of the meteor stream. After a still longer period, about 10,000 years, planetary perturbations will cause the orbits of the stream meteoroids to disperse into different orbits, and their identity as members of a stream will gradually disappear.

Sporadic meteors

Meteors that do not appear to belong to streams are called sporadic. It is likely that in some sense all meteoroids are, or have been, stream members because the physical processes that release meteoroids from either comets or asteroids do so in great numbers. Sporadic meteors are therefore the result of streams too weak to be distinguished from one another or old streams so dispersed as to be no longer recognizable.

There is one rather strange example of a major meteor shower that is clearly identifiable with an astronomical object that, at least at first glance, does not appear to be a comet: the Geminid shower and 3200 Phaeton, respectively. The latter (formerly designated 1983TB) exhibits none of the usual cometary features, a nebulous halo and tail; it simply looks like a small Earth-crossing asteroid. If the stream is cometary, it means that a comet that produced meteoroids prolifically only a few thousand years ago has now completely ceased its cometary activity and looks more like an asteroid. The orbits of 3200 Phaeton and the Geminids also are unlike those of comets in that their aphelia are at 2.4 a.u., well within the orbit of Jupiter. If 3200 Phaeton came from the Oort cloud of comets (see above), it must at one time have crossed the orbit of Jupiter. Over a span of million of years, it is not out of the question that close encounters with the Earth and Venus could gravitationally perturb a Jupiter-crossing orbit into an orbit of this kind. The observed rate at which matter is lost from comets, however, seems to indicate that their inventory is exhausted in only a few thousand years. Thus, the millions of years required for this orbital evolution does not appear to have been available.

It could be hypothesized that 3200 Phaeton was never a comet at all but simply an Apollo object that strayed from the asteroid belt by the reasonably well-understood resonant perturbations that can cause this to occur. The Geminid meteors would then be explained as fragments produced by an asteroidal collision while 3200 Phaeton is traversing the portion of its orbit that is in the asteroid belt. There are serious problems with this explanation. Quantitative calculations show that an asteroidal collision of the required magnitude during an interval of only a few thousand years is very unlikely. Studies of the historical orbital evolution of the Geminid stream suggest that its orbit is incompatible with a single outburst of meteoroids; it is more like one expected from a body that produced a series of outbursts. Finally, the orbit is not the kind one would expect for an Apollo object perturbed from the asteroid belt. Its perihelion is too close to the Sun. An understanding of the mystery of 3200 Phaeton and the Geminids is likely to contribute much to the understanding of comets, the origin of meteor streams, and the relationship between Apollo objects, comets, and asteroids.

Fireball networks. A very significant development in meteor science occurred during the 1960s. This was the establishment of large-scale networks for photographing very bright meteors, or fireballs. These networks were designed to provide all-sky coverage of meteors over areas of about 10^6 square kilometres. Three such networks were developed: the Prairie Network in the central United States, the MORP (Meteorite Observation and Recovery

Table 27: Principal Visually Observable Meteor Showers

shower	average date of maximum	normal duration (days)	visual strength (Northern Hemisphere)	entry velocity (km/sec)	associated comet
Quadrantid	January 3	1	medium	41	not known
Lyrid	April 22	1	irregular	48	1861 I (Thatcher)
Eta Aquarid	May 3	5	weak	66	Halley
S. Delta Aquarid	July 29	8	medium	41	not known
Capricornid	July 30	3	medium	23	not known
Perseid	August 12	5	strong	59	Swift-Tuttle
Andromedid	October 3	11	weak	21	Biela
Draconid	October 9	1	irregular	20	Giacobini-Zinner
Orionid	October 21	2	medium	66	Halley
Taurid	November 8	30	weak	28	Encke
Leonid	November 17	<1	irregular	71	Temple-Tuttle
Geminid	December 14	4	strong	34	3200 Phaeton*

*This body exhibits no cometary activity and is possibly of asteroidal rather than of cometary origin.
Source: Data derived primarily from A.F. Cook in *NASA SP-319* (1973).

Project) network in the prairie provinces of Canada, and the European Network with stations in Germany and Czechoslovakia. The most complete set of published data is that of the Prairie Network, which was operated by the Smithsonian Astrophysical Observatory from 1964 to 1974.

Principal objectives of the networks

An original goal of these networks was to recover a larger number of meteorites for laboratory studies. Other objectives were to determine the orbits of the recovered meteorites and to compare the inferences of meteor theory regarding the density and strength of meteoroids with "ground truth" provided by the study of the same meteoroids in the laboratory. The goal of recovering meteorites had only limited success. Three meteorites were recovered, one by each of the networks. All three meteorites were ordinary chondrites, the most abundant type of stony meteorite. During the operation of the networks, many more meteorites were actually recovered by chance collisions with the roofs of houses.

In spite of this meagre record for meteorite recovery, the networks compiled data that became the basis for a new outlook on meteor science and meteoroid sources. Prior to this effort, there was a tendency to regard the study of meteors and meteorites as independent scientific fields that had little to contribute to each other. Meteors were studied by astronomers and were thought to be associated almost entirely with fragile and low-density "dust balls." On the other hand, meteorites were dense rocks studied by geochemists in the laboratory as samples of the primordial solar system. Little thought was given to why meteor astronomers did not concern themselves with meteorites.

Straightforward application of conventional meteor physics to determine the density of the three recovered meteorites led to the incorrect conclusion that these dense rocks were also low-density objects. This clearly showed that there was something wrong with meteor physics as traditionally applied. A likely, but still not proven, explanation was that the value of luminous efficiency, conventionally used to relate the mass of a meteoroid to the brightness of a meteor, was too low. As a result, the mass of the meteoroid calculated from the luminosity of the meteor was too high. When this large "photometric" mass was combined to measure the cross-sectional area of the meteoroid (using the rate at which it was observed to slow down by atmospheric gas drag) and thereby its radius and volume, a spuriously low density was obtained.

Impact of photographic data on meteor studies

The photographic data from the three fireballs recovered by the networks permitted a more direct empirical approach to the analysis of meteor data. It was found that the atmospheric trajectories of the recovered meteorites, including the end height at which they ceased to be luminous, could be accurately reproduced if the "dynamic mass," determined by the deceleration of the meteor, were used in the theory instead of the photometric mass. It also was found that the ratio of the photometric mass to the dynamic mass was a constant. Laboratory measurements of cosmic-ray effects on the recovered meteorites led to a calculation of the "true mass," which was intermediate between the photometric mass and the dynamic mass. Finally, the light curve (the plot of brightness versus altitude) was similar for the three meteorites.

These results, obtained from the recovered meteorites, could then be used to identify similar objects in the other fireball photographs. Their presence certainly could be expected, because meteorites are produced by fragmentation processes in space similar to those studied in the laboratory. Both experimental and theoretical studies of these processes demonstrate that for every large fragment there must be many small ones.

The recovered fireballs were among the very brightest observed by the photographic networks. Accordingly, there were among the fireball data many objects physically identical to the recovered meteorites. In short, the problem of determining which fireballs were meteorites no longer was dependent on uncertain first principle measurements of density. The empirical data obtained from the recovered meteorites could be used to check the record of each individual fireball, testing quantitatively whether or not the object "looked like a meteorite."

To date, about 30 fireballs have been identified as stony meteorites in this way. The adoption of this approach has increased scientific knowledge of the distribution of meteorite orbits by an order of magnitude.

Application of the same method of analysis shows that fireballs from the Taurid shower, associated with Comet Encke, do not look like meteorites, or at least not like ordinary chondrites. On the other hand, they do not resemble dust balls either but appear to have significant physical strength. The stronger objects of this group have a strength comparable to that of ordinary dirt clods. These physical properties overlap with those of some carbonaceous meteorites. Further analysis of existing data can be expected to shed new light on important questions regarding the relationships between meteoroids of various kinds and their sources. If some way could be found to increase the rate at which fireball networks recover meteorites by about an order of magnitude, the empirical approach, proved valuable for identifying ordinary chondrite sources, could be extended to include less abundant types of meteorites.

METEORITES

As noted above, meteorites are meteoroids that survive passage through the Earth's atmosphere. Any source that can eject such material into interplanetary space should therefore, at least in principle, be thought of as a candidate source of meteorites. There is no fundamental reason why all meteorites must come from similar sources.

Asteroidal origin of most meteorites

It turns out, however, that in practice there are some regions in the solar system that are much more effective in introducing material of substantial strength into Earth-crossing orbits than others. Recent laboratory and theoretical studies fully confirm the older belief that most meteorites are fragments of asteroids. These same studies show that a small fraction, less than 1 percent of the meteorites, come from nonasteroidal sources. The lunar origin of several meteorites is well-established, and it is probable that at least eight others come from Mars. There is evidence from fireball data that a small part of the material in cometary orbits (*i.e.,* with aphelia beyond Jupiter) may possess sufficient strength to successfully penetrate the atmosphere. It is not known if any of this material is present in existing meteorite collections. If it is, the best candidate material would be carbonaceous stony meteorites, probably those of type CI (see below), of which five separate falls have been recovered.

With these few exceptions, it is safe to regard all meteorites as samples broken from outcrops of rock or metal, which until fairly recently in solar-system history were part of asteroidal bodies, mostly in the inner region of the asteroid belt (between about 2.2 and 2.6 a.u.). Like rocks from the Moon, the Earth, or any other similar planetary body, their present state is determined by the total effect of events that occurred on the body throughout the entire history of the solar system. There is no a priori reason why such samples must be pristine samples of a primordial solar nebula from which the present solar system evolved. On the other hand, the principal driving force behind asteroid studies has been the plausible belief that small "primitive" bodies such as asteroids and comets are those most likely to preserve evidence of events that took place in the early solar system. Insofar as this belief is correct, meteorites, samples of these bodies, share this property. Evidence derived from the study of meteorites themselves supports this conclusion.

Types of meteorites. The most fundamental distinction between the various meteorites—no two of which are exactly alike—is the division between chemically undifferentiated and differentiated meteorites. This concept arises from the fact that there is an average chemical composition of the solar system. This average composition must be very close to the composition of the Sun, because the Sun contains most of the mass of the solar system. Spectroscopic comparison of the Sun's chemical makeup with those of other stars shows that its composition in turn is closely related to a cosmic average of the relative abundances of the elements. Important deviations from such average abundances are observed, but these do not invalidate the view that they are deviations from normal abundance ratios determined by the processes by which the chemical elements are formed in stars at various stages of their evolution, returned to the interstellar medium where they are mixed, and then incorporated into new stars and their planetary systems when they are formed.

Since the late 1940s, important advances have been made in the chemical analysis of both meteorites and the Sun. A remarkable result has emerged from this work. Although at one time there appeared to be major differences between the Sun and typical meteorites in the ratios of elements to one another (*e.g.*, iron to silicon), these differences tended to disappear as the accuracy of the measurements improved. It turned out that, for most meteorites and most elements, the solar and meteoritic values of the element ratios relative to silicon (taken as a standard) agreed to within better than a factor of two.

3asic ıgreement ›etween he mete›ritic and olar abunlances ›f the lements

Two kinds of exceptions to this rule were found. Relative to the Sun, the meteorites were deficient in the more volatile elements. For the most volatile elements, hydrogen and the noble gases, the deficiencies were gross—more than a factor of 10^4. For less volatile elements, the deficiencies were smaller; and for the nonvolatile elements, or "refractory" elements, such as iron, magnesium, aluminum, and calcium, the meteoritic and solar abundance ratios were identical within the accuracy of the data.

The other kind of exception to the rule relates to differences between meteorites. For some meteorites, as, for example, those consisting primarily of metallic iron, the similarity between meteoritic and solar abundance ratios fails completely. This also is true for basaltic meteorites, those that appear to have been at one time volcanic magmas and have undergone chemical fractionation of the sort observed in terrestrial igneous rocks.

Undifferentiated meteorites. The meteorites that do obey the rule prove to be of a kind that had already been grouped together on textural grounds—namely, the chondrites. From observed fall rates, this is the most abundant type of meteorite (Table 28). The designation chondrite is based on the occurrence in these meteorites of small (about one millimetre in diameter) spherules called chondrules (Figure 54). In many chondrites, the composition of the chondrules is quite heterogeneous, and the space between the chondrules is filled with a fine-grained matrix material that is richer in volatile elements. In terms of terrestrial rocks, these meteorites seem more akin to sedimentary conglomerates composed of a mechanical mixture of a jumble of components rather than to rocks formed by a process of igneous differentiation.

Not all chondrites contain chondrules of heterogeneous composition. More often, the mixture of heterogeneous chondrules and matrix appears to have undergone a thermal metamorphism that caused the chemical components of the chondrules to come into equilibrium with one another and with the matrix material. This metamorphism was accompanied by the loss of relatively volatile elements. All of these chondrites share one common property: they do not appear to have ever experienced chemical differentiation associated with igneous melting. This shared property is the basis for grouping them all as undifferentiated meteorites even though they clearly differ from one another in the degree to which they have retained volatiles and in various other ways.

Undifferentiated meteorites are classified in two complementary ways. Based on their major element concentrations (Fe, Mg, O), carbon content, and abundance of chondrules, these meteorites naturally cluster into the distinct classes shown in Table 28. In addition, within each of these classes, the meteorites differ according to the degree that they have been thermally metamorphosed or experienced loss of volatile elements. This difference is referred to as the petrologic type (Table 29). For example, the Allende carbonaceous chondrite (Figure 51, top right) is classified CV3, indicating that it belongs to group CV (Table 28) and petrologic type 3 (Table 29).

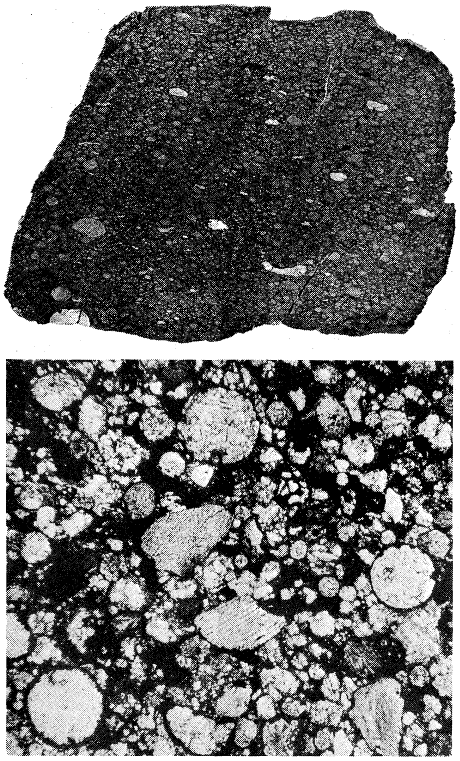


Figure 54: (Top) Sawed and polished section of the Leoville carbonaceous chondrite (CV3; reduced 2.3 X). The small, rounded gray objects are chondrules. The larger, whitish objects are refractory inclusions, similar to those seen in Allende (Figure 51, top right). (Bottom) Microscopic view of a thin section of the Tieschitz (Czechoslovakia) ordinary chondrite (olivine-bronzite H3). The round objects are chondrules, some of which have been fractured by collisions after their formation.

(Top) F. Wlotzka, Max-Planck-Institut fur Chemie, Mainz, W. Ger., (bottom) J.A. Wood

ifferences :tween .ondrite pes

Chondrites obviously differ from one another in several important respects. As has been pointed out, they vary in the extent to which they have undergone thermal metamorphism. Another important distinction is between the more abundant ordinary chondrites (of which there are three principal kinds) and the rarer chondritic meteorites that exhibit significant chemical differences. One of these types is the enstatite chondrite, which is, among other things, chemically more reduced than the ordinary chondrites. Almost all of the iron in these meteorites, for example, is in metallic form. As a result, most of the abundant silicate mineral, pyroxene, is present as nearly pure enstatite ($MgSiO_3$) rather than in magnesium-iron solid solution minerals, such as bronzite and hypersthene, found in the ordinary chondrites. In enstatite chondrites, the readily oxidized element silicon is even found in the reduced state, and calcium occurs as the sulfide mineral oldhamite (CaS) rather than in its more usual silicate forms.

Other very important varieties of chondrites are grouped together as the carbonaceous chondrites. As their name implies, they characteristically contain more carbon (0.5 to 5 percent) than the ordinary chondrites (only about 0.1%). The mineral constituents of the carbonaceous chondrites are less chemically equilibrated with one another than even the unequilibrated ordinary chondrites. In many cases, one finds in the same meteorite carbonaceous material that formed at low temperatures and inclusions of the most refractory minerals—perovskite ($CaTiO_3$), hibonite ($CaAl_{12}O_{19}$), and melilite (solid solutions of $Ca_2Al_2SiO_7$ and $Ca_2MgSi_2O_7$).

CI carbonaceous chondrites

Perhaps the most interesting type of meteorite is the CI carbonaceous chondrite. Strictly speaking, one could legitimately question why such meteorites are called chondrites at all inasmuch as they do not contain chondrules. When compared with solar abundances (Figure 55), however, it turns out that they are the least differentiated meteorites of all, and in making a classification scheme it certainly makes sense to group them with the other undifferentiated meteorites. In accordance with the correlation already observed between chemical undifferentiation and

Table 28: Classification of Undifferentiated Meteorites (Chondrites)

class	group	percentage of observed chondrite falls	total iron (weight %)	Fe—metal / Fe—total (%)	FeO / (FeO + Mg) (mole %)	chondrules (%)	carbon (weight %)
Ordinary	H	38.1	25–31	58–65	17	80	0.1
	L	46.3	21–23	30–39	22	80	0.1
	LL	8.5	20–23	6–25	27	80	0.1
Enstatite	E	1.8	22–35	70–88	0.05	20	0.4
Carbonaceous	CI	0.7	18–19	0	45	<1	3.1
	CM	2.3	21–24	0.1–0.6	43	2	2.5
	CO	1.0	24–26	3–19	35	70	0.5
	CV	1.3	22–25	0.8–25	35	30	0.5

Sources: Data primarily from B. Mason, *Handbook of Elemental Abundances in Meteorites* (1971); A.L. Graham, A.W.R. Bevan, and R. Hutchison, *Catalogue of Meteorites* (1985); and J.T. Wasson, *Meteorites, Their Record of Early Solar-System History* (1985).

chemical disequilibrium, the constituents of CI carbonaceous chondrites are far from equilibrium. The iron in these meteorites, for example, is highly oxidized, most of it occurring in the ferric iron-bearing mineral magnetite (Fe_3O_4). At the same time, carbon is present as highly reduced complex hydrocarbons. In equilibrium, the carbon would be oxidized to carbon monoxide and carbon dioxide, and the iron would be reduced to metallic iron.

Because CI chondrites are the most undifferentiated meteorites known, it has been speculated that, unlike most meteorites, they are of cometary rather than of asteroidal origin, since comets are believed to represent the most unaltered material in the solar system. There are difficulties in accepting this speculation as being correct. For example, detailed study of these meteorites shows that, in spite of their chemically undifferentiated and unequilibrated nature, they have had a complex chemical and physical history and are not simply a collection of interstellar dust. On the other hand, scientific knowledge about the nature and origin of comets is still limited, so that it would be unwise to dismiss this intriguing hypothesis prematurely.

Differentiated meteorites. Differentiated meteorites exhibit the type of chemical fractionation one would expect to occur on a planetary body that underwent core formation and magmatic differentiation similar to that observed in terrestrial and lunar volcanic rocks. Indeed, at one time it was thought likely that the most abundant type of differentiated stony meteorite, the basaltic achondrites, were actually lunar mare basalts. Their similarities to the lunar basalts subsequently returned by the Apollo missions showed that this was by no means a farfetched idea, but detailed considerations such as oxygen isotope ratios showed that it was not correct.

Iron mete- orites and achondrite

In classifying differentiated meteorites, the major division is made between the iron meteorites and the stony differentiated meteorites, the achondrites ("not chondrites"). In addition, there are a number of stony-iron meteorites that contain mixtures of large masses of nickel-iron metal and

Table 29: Classification of Undifferentiated Meteorites in Terms of Petrologic Type and Metamorphic Grade*

	petrologic type					
	1	2	3	4	5	6
Homogeneity of olivine and pyroxene compositions	—	mean deviations of pyroxene ≥ 5%, of olivine ≥ 3%		5% > mean pyroxene deviation > 0%	uniform ferromagnesian minerals	
Structural state of low-Ca pyroxene	—	predominantly monoclinic		abundant monoclinic crystals	orthorhombic	
Degree of development of secondary feldspar	—	absent		predominantly as microcrystalline aggregates		clear, interstitial grains
Igneous glass	—	clear and isotropic primary glass, variable abundance		turbid if present	absent	
Metallic minerals	—	taenite absent or very minor (Ni < 200 mg/g)	kamacite and taenite (Ni > 200 mg/g) present			
Mean Ni content of sulfide minerals	—	> 5 mg/g	< 5 mg/g			
Overall texture	no chondrules	very sharply defined chondrules		well-defined chondrules	chondrules readily delineated	poorly defined chondrules
Texture of matrix	all fine-grained, opaque	much opaque matrix	opaque matrix	transparent microcrystalline matrix	recrystallized matrix	
Bulk carbon content	30–50 mg/g	8–26 mg/g	2–10 mg/g	< 2 mg/g		
Bulk water content	180–220 mg/g	20–160 mg/g	3–30 mg/g	< 15 mg/g		

*Following R. Van Schmus and J. Wood in *Geochim. Cosmochim. Acta* 31:747 (1967).
Source: J.T. Wasson, *Meteorites, Their Record of Early Solar-System History* (1985).

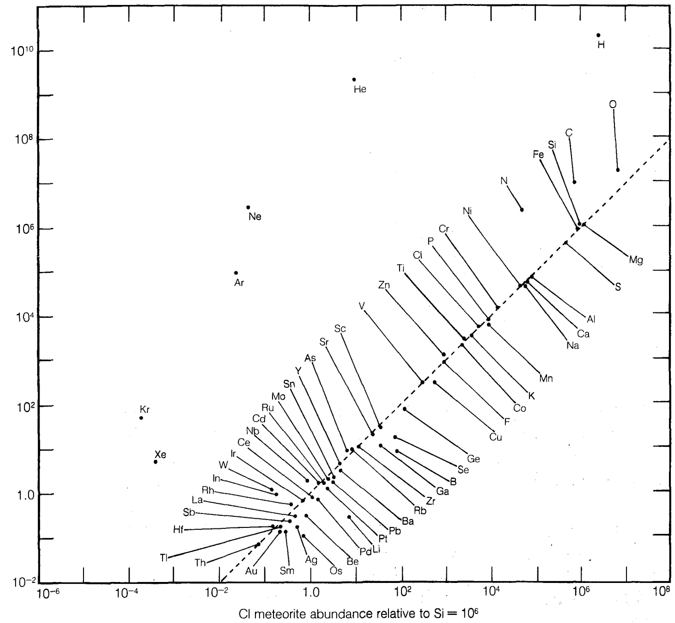


Figure 55: *Comparison of solar and meteoritic carbonaceous chondrite (CI) abundances of elements (relative to silicon = 10^6).* For most elements, the solar and meteoritic abundances are very similar and therefore fall near a line with a slope of 45°. For the most volatile elements (hydrogen, noble gases, carbon, nitrogen, oxygen), the meteorites are highly depleted relative to the solar composition.

Based on data from *Meteorites: Their Record of Early Solar-System History* by John T. Wasson, W.H. Freeman and Company, 1985, and B. Mason, *Handbook of Elemental Abundances in Meteorites* (1971), Gordon and Breach, Science Publishers, Inc.

differentiated silicate rock. The usual classification scheme for differentiated meteorites is given in Tables 30 and 31.

There are compelling reasons for believing that, like the chondrites, all differentiated meteorites, with the exception of those few from the Moon and possibly Mars, are asteroidal fragments. Because asteroids are too small to have experienced the accretional and long-lived radioactive heating that powers igneous processes on the inner planets, this may seem surprising. A partial solution to this enigma comes from radiometric dating, which shows that this differentiation took place very early in the history of the solar system, about 4,500,000,000 years ago. At that time other heating mechanisms may have been available—*e.g.,* heating by short-lived radioactive isotopes such as aluminum-26 (which has a half-life of about 700,000 years) or inductive heating by intense solar activity. It is even possible that asteroids from which differentiated meteorites seem to be derived are fragments of much larger differentiated planetesimals originally present in the inner planet region during the development of the Earth and Venus and which were stored in quasi-stable orbits in the innermost asteroid belt (2.2 a.u.) of the early solar system.

Specific asteroidal source regions for recovered meteorites. There is compelling, even if circumstantial, evidence that nearly all meteorites are derived from the asteroid belt. For scientific understanding of this matter to be complete, it is necessary to know which asteroids are the sources of particular types of meteorites and the mechanisms by which meteorites are transported from the asteroid belt to Earth. It is possible that the ultimate answer will not be found until asteroids are explored by spacecraft. Nevertheless, considerable information relevant to this question is already available.

Orbits of Earth-impacting ordinary chondrites

For the most abundant meteorite type, the ordinary chondrites, there is a fairly large body of evidence that indicates that most of these meteorites strike the Earth while traveling in a rather special type of orbit. Such an orbit has its perihelion just inside the Earth's orbit, as well as low inclination (less than about 10°) and fairly high eccentricity (about 0.6). This has been determined on the basis of the time of day during which meteorites are observed to fall. It has been noted that about twice as many ordinary chondrites fall during the daylight PM hours as during daylight hours before noon. This fact requires that the meteorite must be moving faster than the Earth at the time of impact and, therefore, must be near its perihelion according to Keplerian dynamics. Quantitative consideration of the implications of time of fall data is entirely consistent with evidence obtained by visual observations of meteorite radiants and with about 30 fireballs identified by the Prairie Network as being, at least physically, equivalent to stony meteorites (see above).

This distribution of orbits places strong constraints on the source region in the asteroid belt from which ordinary chondrites can be derived. If meteorites in the kilogram mass range are to be derived directly from the asteroid belt, the mechanism by which they are extracted from the belt and transferred to an Earth-crossing orbit must operate quite rapidly on a time scale of a few million years. Only then can a meteoroid escape destruction by collision while near its aphelion in the asteroid belt. This collisional lifetime is not simply a theoretical result but is measured directly by the cosmic-ray exposure ages of meteorites, as discussed below.

High-energy galactic cosmic rays—primarily protons—have a range of penetration on the order of a few metres in meteoritic material. Any meteoroid of smaller dimensions will be radiated throughout by this proton bombardment. The high-energy protons cause spallation reactions (nuclear interactions that result in the release of many nucleons) on the abundant elements in the meteoroidal target. As a consequence, a large number of otherwise rare isotopic

Table 30: Classification of Iron Meteorites*

class	subclass	symbol	nickel (%)	Widmanstätten structure kamacite (α-iron) bandwidth (mm)	percentage of total irons
Octahedrites		O	9–13		76
	coarsest	Ogg		>3.3	4
	coarse	Og		1.3–3.3	17
	medium	Om		0.5–1.3	40
	fine	Of		0.2–0.55	10
	finest	Off		<0.2	1
	plessitic	Opl		<0.2 (kamacite spindles)	4
Hexahedrites		H	5.5	pure kamacite	10
Nickel-rich ataxites		D	9–69	fine intergrowth of kamacite and taenite (γ iron; called plessite)	6

*Differentiated meteorites of this type consist principally of nickel-iron metal and iron sulfide; they make up 4 percent of all meteorite falls. The classification given here is structural. Another classification is based on the trace element concentrations of iron meteorites, particularly gallium, germanium, and iridium. This classification tends to group the same meteorites that are grouped in the structural classification, but the grouping is not exactly the same. For either classification, a significant number (10–15 percent) of meteorites do not fit well and must be classified "anomalous."

Sources: B. Mason, *Handbook of Elemental Abundances in Meteorites* (1971); A.L. Graham, A.W.R. Bevan, and R. Hutchison, *Catalogue of Meteorites* (1985); J.T. Wasson, *Meteorites, Their Record of Early Solar-System History* (1985); and V.F. Buchwald, *Handbook of Iron Meteorites* (1975).

Table 31: Silicate-Rich Differentiated Meteorites

class	principal silicate minerals	iron metal (weight %)	Fe-Mg silicate $\frac{FeO}{(FeO + MgO)}$	percentage of total observed falls
Pallasite	olivine	28–88	11–14	0.3
Mesosiderite	orthopyroxene plagioclase	30–55	23–27	0.8
Achondrites				
Basaltic-eucrite	pigeonite, plagioclase	<0.1	50–67	2.7
Basaltic-howardite	orthopyroxene, plagioclase	<1	25–40	2.4
Enstatite (aubrite)	enstatite	1	0	1
Ureilite	olivine, clinopyroxene	0.3–6	10–25	0.4
Diogenite	hypersthene	<1	25–27	1
Shergottite*	pigeonite, augite, plagioclase	0	59–76	0.3
Nakhlite*	augite, olivine	0	69	
Chassignite*	olivine	0	53	

*Isotopic and geochemical data suggest that these meteorites are from the planet Mars. Several achondrites found in Antarctica are of lunar origin. A small number (0.3 percent of total observed falls) do not fit into any of these categories and are classified as "anomalous."

Sources: B. Mason, *Handbook of Elemental Abundances in Meteorites* (1971); A.L. Graham, A.W.R. Bevan, and R. Hutchison, *Catalogue of Meteorites* (1985); and J.T. Wasson, *Meteorites, Their Record of Early Solar-System History* (1985).

Cosmic-ray exposure age

species, both stable and radioactive, are produced. These include the stable noble gas isotopes helium-3, neon-21, and argon-36, as well as the rare isotope potassium-40 and various short- and moderately long-lived radioactive isotopes, including hydrogen-3 (with a half-life of 12.26 years), beryllium-7 (53.29 days), beryllium-10 (1.6×10^6 years), aluminum-26 (7.2×10^5 years), manganese-53 (3.7×10^6 years), and cobalt-60 (5.272 years). The concentration of the shorter-lived radioactive isotopes can be used to monitor the cosmic-ray bombardment rate, and the accumulation of the stable species (*e.g.,* neon-21) measures the time in the past that this bombardment began—*i.e.,* the time that the meteoroidal fragment was separated from a larger object, that was large enough to have shielded it from cosmic-ray bombardment. For chondritic meteorites, the distribution of cosmic-ray exposure age falls off quite quickly with age. This is to some extent a consequence of the dynamic evolution of the meteoroid orbits but for the most part should be attributed to a "collision half-life" of 5,000,000 to 10,000,000 years.

Mechanisms by which meteorites are transported into Earth-crossing orbits

There are only two processes known that can accelerate meteoroidal fragments into Earth-crossing orbits on this short time scale given by cosmic-ray exposure ages. These processes are direct collisional ejection at velocities of about five kilometres per second and gravitational acceleration by dynamic resonances in the asteroid belt, of which the 3:1 resonance at 2.5 a.u. is of dominant importance. In a hypervelocity collision, some material is ejected at the required high velocity, but the quantity of this material is small and most of it is pulverized by the associated shock pressures. High-velocity ejection is likely to be responsible for the occurrence of meteorites from Mars or the Moon, but it completely fails to provide the observed quantity of meteoroids from the asteroid belt. The resonant mechanisms are therefore of much greater importance. Bodies orbiting the Sun with a semimajor axis near 2.5 a.u. will complete three revolutions about the Sun in the time that Jupiter, a strong source of gravitational perturbations, executes one revolution. The resulting resonant acceleration will cause the orbit of the body to become "chaotic," and its perihelion will become Earth-crossing in about 1,000,000 years.

The calculated quantity of asteroidal material in the meteorite-size range delivered to the Earth from the 3:1 resonance agrees well with the 10^8 gram per year terrestrial impact rate of ordinary chondritic meteoroids. Moreover, this resonant acceleration presents a natural mechanism for concentrating meteoroid perihelia near the Earth's orbit, thereby explaining the special distribution of orbits observed for ordinary chondrites. In fact, as it turns out, meteoroids derived by this resonance acceleration mechanism should be overly concentrated near 1 a.u.; the ratio of afternoon to morning falls should be about 3:1 instead of the observed 2:1. This discrepancy is removed when one takes into account the fact that larger asteroidal fragments (those reaching the size of Apollo objects) will also be accelerated into an Earth-crossing orbit by the same resonant mechanism. Meteorite-sized fragments will be produced as collision debris from these larger bodies, but they will not have the special distribution of orbits exhibited by the smaller asteroidal fragments introduced more directly into Earth-crossing orbits. When the contribution from resonant asteroidal ejecta over the entire size range is averaged, the predicted and observed orbital distribution match rather well.

Orbital statistics are not very well known for other meteorite types. It does appear that the differentiated basaltic achondrites fail to share the special orbital distribution observed for the ordinary chondrites. Basaltic achondrites are most likely the collision debris of Apollo objects that were extracted by another resonance mechanism known to operate in the innermost belt; this mechanism can be shown to be effective only in providing larger Earth-crossing bodies. Although more work remains to be done on this problem, it seems likely that known resonant mechanisms are adequate to explain the dynamical processes by which all classes of meteorites are delivered from their asteroidal source regions into Earth-crossing orbits.

Derivation of meteoritic material from these designated regions of the inner asteroid belt implies that asteroids in such regions have the chemical and mineralogical composition observed in the meteorites. The surface mineralogical composition of asteroids can be determined directly by Earth-based reflectance spectrometry. These measurements have been made for most of the larger asteroids. Although no two reflectance spectra are exactly alike in detail, most asteroids fall into one of two general groups, the S class and the C class. The S-class asteroids have moderate albedos (though comparatively high overall) and contain mixtures of olivine (magnesium, iron silicates), pyroxene (silicates containing magnesium, iron, calcium, and aluminum), and metallic iron. These are the same minerals found in ordinary chondrites and in basaltic achondrites. The C-class asteroids have low reflectance, and their more featureless spectra indicate the presence of light-absorbing opaque minerals. It is plausible to consider these asteroids as candidate sources for carbonaceous meteorites.

Chemical and mineralogical similaritie between the major classes of asteroids and meteorite

When considered in more detail, there are certain difficulties in identifying the required number of S-class asteroids as ordinary chondrite sources. Although qualitatively the mineralogy of these asteroids agrees with that of ordinary chondrites, the proportions of the constituent minerals to one another does not match as well. In particular, the S-class asteroids appear to have, at least on their surfaces, about twice as much metallic iron as the ordinary chondrites. The solution to this discrepancy is not known at present. It may be that the surfaces of the asteroids are not truly representative of their interiors, having been altered by collisional bombardment or exposure to solar radiation and particles. Another possibility is that a systematic change in mineralogical composition occurs during the process of collisional grinding of asteroids into ever-smaller bodies, ultimately to those of meteorite size. There is some indication that the discrepancy is less serious for small asteroidal fragments observed as Apollo objects, which is consistent with this hypothesis.

Meteorites, asteroids, and the early solar system. A prime long-range objective of exploring the Moon and planets by spacecraft is to collect samples of these bodies for detailed laboratory study. With the few exceptions noted earlier, meteorites are samples of asteroids delivered to Earth by fairly well-understood natural mechanisms. It also is believed that samples collected from asteroids will prove especially valuable, because such small bodies are most likely to be "primitive" and thus retain the record of events that occurred during their own formation and that of the solar system in general. The study of meteorites is already providing scientists with much valuable information directly related to this matter.

emnants
' presolar
terstellar
ains

Isotopic records. Meteorites indicate to investigators that the asteroid belt must always have been a relatively tranquil region of the solar system. Some unequilibrated chondrites have inherited and preserved without complete mixing the remnants of presolar interstellar grains. This is demonstrated by variations of the ratios of the oxygen isotopes oxygen-16, oxygen-17, and oxygen-18 within a single meteorite and between different meteorites. It is well known that the oxygen isotopes are fractionated by natural chemical processes. Variations in the ratio of oxygen-18 to oxygen-16, for example, form the basis for paleotemperature studies of ancient terrestrial sedimentary rocks by making use of the temperature dependence of the isotopic chemical equilibrium between seawater and the calcium carbonate that forms the shells of marine organisms. It is characteristic of this natural fractionation dependent on isotopic mass that the degree of fractionation is proportional to the difference of the masses of the isotopes. Thus, in these marine sediments, the variation in the ratio oxygen-18 : oxygen-16 is twice that of oxygen-17 to oxygen-16. In some unequilibrated chondrites, however, the variations in the oxygen-18 to oxygen-16 ratios are equal to those in oxygen-17 : oxygen-16. It is possible to devise special chemical processes that could produce fractionation of this kind, but it is much more likely that the variations found in the meteorites are caused by nuclear processes that predated the formation of the Sun and solar system. The interstellar grains that were the carriers of these isotope anomalies were probably formed in stellar atmospheres and preserved the signatures of the isotope-formation processes of stars of particular types. These isotopic variations of nuclear origin are found not only in oxygen but also in other less abundant elements, including neon, xenon, titanium, and chromium.

Probable early evolution of the solar system based on meteoritic evidence. The need to provide an environment sufficiently tranquil to preserve this isotopic record, as well as other fragile relics of early solar system events observed in meteorites, places important constraints on the formation of the solar system. If one examines the distribution of matter in the present solar system, it is seen that the density is high both in the region of the inner planets and in the region of the giant planets in the outer solar system but very low in the wide space between Mars and Jupiter. This fact itself is surprising: Why should the solar nebula from which the solar system formed have had a great hole in it? The answer is that it probably did not. In order for asteroids to have formed and developed at all on the time scale of a few million years indicated by the radiometric dating of meteorites, densities more like those in the regions occupied by the giant planets would have been required, as shown by theoretical calculations. It is difficult to escape the conclusion that the quantity of matter in what is now the asteroid belt must have been much greater, perhaps by as much as 10^4 times as is observed there today. Therefore some natural process has to have removed almost all of the material in this region of the solar system after the formation of asteroidal bodies, but in a sufficiently gentle way to have preserved the relics of pre-solar and early solar events found in meteorites.

Although the details are not yet understood, it seems most likely that the formation of the giant planets, particularly Jupiter, quickly resulted in the evacuation of most of the matter from this region of the solar system. This means that Jupiter formed rapidly, before bodies in the asteroid belt had grown to become full-fledged planets. The mineralogical and chemical record of the undifferentiated meteorites is not compatible with their once having been part of a planet even as large as the Moon. Also, the very energetic collisional events that would be associated with the dispersal of large planets in the asteroid belt would preclude preservation of the observed relics.

Planetary
growth and
changes
in mass
distribution
in the
outer solar
system

These constraints, for the most part based on meteoritic evidence, define a conceivable chain of events for the early evolution of the solar system beyond the region of the inner planets. Even though the density of matter was at least as great in the asteroid belt as in the region of Jupiter, planetary growth must have been more rapid in Jupiter's vicinity than in the asteroid belt. (Some suggestions have been made as to how this may be possible, but it remains to be seen if they are satisfactory.) Within about 1,000,000 years, proto-Jupiter(s) began to capture the massive quantities of hydrogen and helium from the solar nebula that constitute most of the giant planet today. At the same time, thousands of large asteroids greater than 100 kilometres in diameter, including some as big as the largest present-day asteroids but not much bigger, had formed. Shortly thereafter, as Jupiter approached its present mass, most of the residual nebular gas was removed by the intense solar ultraviolet radiation characteristic of young stars.

Up until the time Jupiter approached its present mass, the asteroids moved in nearly circular orbits in accordance with the weak mutual gravitational perturbations expected for small bodies. During the final formation of Jupiter and Saturn and the removal of the nebular gas, the changing mass distribution in the outer solar system caused waves of resonant perturbations to sweep through the asteroid belt, increasing the eccentricities and inclinations of the asteroids to the moderate values observed today. Because the asteroids had not grown larger than the asteroids of today, collisions at the relative velocities of about five kilometres per second, associated with their present-day eccentricities and inclinations, would begin to grind down the smaller bodies. This would occur without the shock effects associated with disruption of larger planetary bodies at higher impact velocities.

The foregoing scenario of planetary growth is certain to be wrong in detail and may well be wrong altogether. Nevertheless, without the samples of asteroids provided by recovered meteorites, there would be little observational basis at all for formulating models of this kind. All of the qualitative statements made above should be considered not as established facts but rather as statements of theoretical problems that need to be thoroughly worked out. Their results must be compared with old and new observational data and reformulated and reworked until a satisfactory understanding of planetary formation is achieved.

Thermal evolution of the solar nebula and planetesimals. The available meteoritic evidence is relevant to many other unsolved central questions concerning the early solar system, including some that bear on the way the Sun was formed and its early history. One of these is the question of the thermal evolution of the solar nebula and the planetesimals that formed and grew within it.

As discussed above, the environment of the early asteroid belt must have been a rather "tranquil" one. By tranquil, one actually means thermally, rather than physically, uneventful. The collisions that led to the removal of most of the material from the asteroid belt were highly disruptive, and there is ample evidence of extensive physical disruption in the textures of even the most primitive meteorites. Yet, the preservation of primordial relics, such as the isotope anomalies, argues against widespread heating of the asteroidal region to temperatures as high as 1,000 K.

Low-tempera-
ture origin
of the
asteroidal
region

A relatively cool solar nebula at distances this far from the Sun is in agreement with most current theoretical calculations of the formation of the Sun. In addition to the evidence for an overall low-temperature origin for this region of the solar system, however, the meteoritic record clearly shows the imprint of high-temperature events of major importance, which are not well understood at present. The most apparent of these is the very abundant presence of chondrules in all undifferentiated meteorites except the CI chondrites. Chondrules appear to have once been melted droplets primarily of silicate composition and would have required temperatures of about 1,500 K to have formed.

Formation of chondrules

If chondrules were a relatively rare meteoritic curiosity, one could legitimately consider them an interesting detail to be explained someday but not a matter of central importance. Yet, the fact that chondrules (or their broken fragments) make up most of the mass of the most abundant class of meteorites, the ordinary chondrites, and a major portion of other chondrites, indicates that their formation must have been of major importance in the early solar system. Even if ordinary chondrites formed only within a restricted region of the asteroid belt adjacent to the 3:1 Kirkwood gap (*i.e.,* between 2.44 and 2.56 a.u.), this is still about 10 percent of the entire asteroid belt. It also is likely that other chondrule-bearing meteorites formed outside this region, even though their asteroidal sources may be in that region today.

It seems impossible that chondrules are the product of igneous differentiation because they are of nearly undifferentiated solar composition except for the most volatile elements. Nor does it seem likely that they are impact droplets, such as those found in the lunar soil. The expected low impact velocities of asteroidal planetesimals argue against a ubiquitous high-velocity impact environment. Thus, there seems to be no way by which the chondrules could have formed in or on planetesimals. In all likelihood, they were formed in the solar nebula. At the same time, the evidence for rapid cooling of chondrules argues against their formation by large-scale condensation from a very hot solar nebula. Local, transient heating events appear to have been important on a wide scale in the solar nebula, but the nature and cause of these events remain unknown.

A problem of similar difficulty is that of the origin of the large (up to more than one centimetre in diameter), highly refractory inclusions found in the CV meteorites (see Table 28), especially prominent in the Allende carbonaceous chondrites. Though not as ubiquitous as chondrules, these inclusions are by no means of negligible abundance. Unlike chondrules, they are highly fractionated chemically, apparently as a result of more prolonged heating to about 1,500 K. Because many of the isotopic anomalies are associated with these refractory inclusions, interpretation of this important evidence is limited by a poor understanding of their origin.

Evidence of internal heating in differentiated meteorites

The foregoing thermal events most likely occurred in the solar nebula rather than on growing asteroidal bodies. In addition, thermal effects are observed in meteorites associated with internal heating. The most apparent of these are the differentiated meteorites, which probably represent about 10 percent of the asteroidal region sampled. The asteroids from which these meteorites were fragmented, though probably formed from a relatively cool solar nebula, experienced internal heating, core formation, and igneous differentiation within a few million years of the formation of the solar nebula itself. Clear and detailed evidence for this is based on (1) radiometric dating of the minerals formed by this igneous differentiation, (2) the mineral assemblages of the resulting igneous rocks, and (3) the slow cooling that produced large crystals of differentiated minerals (*e.g.,* the so-called Widmanstätten structure observed in many nickel-iron meteorites). The actual process responsible for this heating is yet unknown, but several good possibilities are being evaluated. These include heating by the short-lived radioactive isotope aluminum-26, heating by electric currents induced by early solar activity, and accretional heating of planetesimals in the terrestrial planetary region, followed by fragmentation and transfer to the innermost asteroid belt.

The ordinary chondrites also experienced heating after the formation of chondritic planetesimals, but not enough to produce melting. As a result, chondritic material, presumably once resembling the unequilibrated ordinary chondrites, was metamorphosed to produce the more abundant equilibrated ordinary chondrites. The time scale for this metamorphism is, within uncertainties, the same as that which produced the parent asteroids of the differentiated meteorites. It is plausible that some of the same heat sources (*e.g.,* aluminum-26) may have been responsible. (G.W.We.)

BIBLIOGRAPHY

General works: GEORGE ABELL, *Exploration of the Universe,* 4th ed. (1982), contains chapters on the solar system written at an elementary level. Popular accounts include: F.L. WHIPPLE, *Orbiting the Sun: Planets and Satellites of the Solar System* (1981); J. KELLY BEATTY, BRIAN O'LEARY, and ANDREW CHAIKIN (eds.), *The New Solar System,* 2nd ed. (1982); and WERNHER VON BRAUN and FREDERICK I. ORDWAY III, *New Worlds: Discoveries from Our Solar System* (1979). See also issues in the University of Arizona Space Science series, including: TOM GEHRELS (ed.), *Jupiter* (1976), *Protostars and Planets* (1978), and *Asteroids* (1979); and JOSEPH A. BURNS (ed.), *Planetary Satellites* (1977). Authoritative reviews are found in *Annual Review of Earth and Planetary Sciences* and in *Annual Review of Astronomy and Astrophysics.* Most original work on the solar system is published in technical journals, including: *Icarus* (monthly); *Astronomical Journal* (monthly); and *Science* (weekly).

(R.L.Ne./G.S.O./D.C.P./Ed.)

The Sun: A large part of modern solar research is published in *Solar Physics,* a monthly journal for solar research and the study of solar–terrestrial relations. Popular works on the Sun's atmosphere include: ROBERT W. NOYES, *The Sun, Our Star* (1982); G. GAMOW, *The Birth and Death of the Sun* (1945, reissued 1976); and E.N. PARKER, "The Sun," *Scientific American,* 233:42–50 (September 1975). Works of a more technical nature include: H. ZIRIN, *The Solar Atmosphere* (1966); G.P. KUIPER (ed.), *The Sun* (1953); and E. TANDBERGHANSSEN, *Solar Activity* (1967). Specific subjects can be found treated in the following technical books: R.J. BRAY and R.E. LOUGHHEAD, *Sunspots* (1964, reprinted 1979), and *The Solar Granulation* (1967); H.J. and E.V.P. SMITH, *Solar Flares* (1963); Y. OHMAN (ed.), *Mass Motions in Solar Flares and Related Phenomena* (1968); N. ROBINSON (ed.), *Solar Radiation* (1966); D.E. BILLINGS, *A Guide to the Solar Corona* (1966); J.W. EVANS (ed.), *The Solar Corona* (1963); A.J. HUNDHAUSEN, *Coronal Expansion and Solar Wind* (1972); R.N. THOMAS and R.G. ATHAY, *Physics of the Solar Chromosphere* (1961); J. MEEUS, C.C. GROSJEAN, and W. VANDERLEEN, *Canon of Solar Eclipses* (1966); K.O. KIEPENHEUER (ed.), *Structure and Development of Solar Active Regions* (1968); M.R. KUNDU, *Solar Radio Astronomy* (1965); A.G. SMITH, *Radio Exploration of the Sun* (1967); S. CHAPMAN, *Solar Terrestrial Physics* (1972), and *Solar and Magnetospheric Science* (1976); V.V. SOBOLEV, *Treatise on Radiative Transfer,* trans. by S.I. GAPOSCHKIN (1963, reissued 1978); I.S. SHKLOVSKII, *Stars: Their Birth, Life and Death* (1975; Eng. trans. from the Russian, 1978); and R.W. NOYES, "New Developments in Solar Research," in E.H. AVRETT (ed.), *Frontiers of Astrophysics* (1976). Atlases of Fraunhofer lines include G. BRUCKNER, *Photometric Atlas of the Near Ultraviolet Solar Spectrum 2988 Å–3629 Å* (1960); O.C. MOHLER *et al., Photometric Atlas of the Near Infra-Red Solar Spectrum λ 8465 to λ 25242* (1950, reissued 1978); L. DELBOUILLE and G. ROLAND, *Photometric Atlas of the Solar Spectrum from λ 7498 to λ 12016* (1963); and M. MIGEOTTE, L. NEVEN, and J. SWENSSON, *The Solar Spectrum from 2.8 to 23.7 Microns* (1956).

(A.K.P./E.A.M.)

Mercury: Nontechnical discussions of Mercury may be found in BRUCE C. MURRAY, "Mercury," *Scientific American,* 233:58–68 (September 1975); J. KELLY BEATTY, "Mariner 10's Second Look at Mercury," *Sky and Telescope,* 48:307–314 (November 1974); KENNETH F. WEAVER, "Mariner Unveils Venus and Mercury," *National Geographic,* 147:858–869 (June 1975); B.C. MURRAY and ERIC BURGESS, *Flight to Mercury* (1977); and B.C. MURRAY, MICHAEL C. MALIN, and RONALD GREELEY, *Earthlike Planets* (1981). An excellent collection of Mariner 10 photographs is available in MERTON E. DAVIES *et al., Atlas of Mercury* (1978). Technical discussions of Mercury may be found in *Science,* 185:141–180 (July 12, 1974); *Journal of Geophysical Research,* 80:2341–2514 (June 10, 1975); *Icarus,* 28:429–609 (August 1976); and *Physics of the Earth and Planetary Interiors,* 15:113–312 (November 1977). See also R.G. STROM, "Mercury: A Post-Mariner 10 Assessment," *Space Science Reviews,* 24:3–70 (1979).

(M.C.Ma.)

Venus: A reference work on Venus that presents the primary results of the Pioneer Venus mission and the current data on the planet is *Journal of Geophysical Research,* 85:7573–8337 (Dec. 30, 1980). A summary article on results of the Venera probe

missions is M.YA. MAROV, "Results of Venus Missions," in *Annual Review of Astronomy & Astrophysics,* 16:141–169 (1978). A general review article on Venus is ANDREW and LOUISE YOUNG, "Venus," *Scientific American,* 233:70–78 (September 1975). A classic paper on infrared temperatures of the clouds is W.M. SINTON and J. STRONG, "Radiometric Observations of Venus," *Astrophys. J.,* 131:470–490 (1960). The greenhouse model as initially proposed is described in CARL SAGAN, "The Radiation Balance of Venus," Jet Propulsion Laboratory Tech. Rep.:32–34 (Sept. 15, 1960). The initial report of the high radio flux from Venus is in C.H. MAYER, *et al.,* "Observations of Venus at 3.15 cm Wavelength," *Astrophys. J.,* 127:1–16 (1958). The discovery of trace gases in Venus' atmosphere by high-resolution spectroscopic techniques is described in P. and J. CONNES *et al.,* "Traces of HCl and HF in the Atmosphere of Venus," *Astrophys. J.,* 147:1230–37 (1967). The first investigation of the clouds by means of Earth-based optical measurements is in J.E. HANSEN and A. ARKING, "Clouds of Venus," *Science,* 171:669–672 (Feb. 19, 1971). An illustrated book with popular descriptions is C. SAGAN and J.N. LEONARD, *Planets* (1966 and subsequent ed.). G.E. HUNT and PATRICK MOORE, *The Planet Venus* (1983), a good introduction to the planet and the space vehicles that have studied it.

(A.S.)

Earth: FRANK PRESS and RAYMOND SIEVER, *Earth,* 4th ed. (1986), an excellent overview of the physical properties and dynamics of the Earth; I.G. GASS, P.J. SMITH, and R.C.L. WILSON (eds.), *Understanding the Earth* (1971), some 25 articles by as many authors who are leading experts (readable, nonmathematical); SCIENTIFIC AMERICAN, *Readings in the Earth Sciences,* 2 vol. (1969), an excellent collection containing 74 offprints of articles that appeared in *Scientific American* between 1948 and 1969—see also current issues of this periodical for further valuable articles; OWEN M. PHILLIPS, *The Heart of the Earth* (1968), introductory for a college science student; FRANK D. STACEY, *Physics of the Earth* (1969), an outstanding text for the advanced undergraduate or beginning graduate student of science with a corresponding mathematics background, including an extensive and carefully selected bibliography; GEORGE D. GARLAND, *Introduction to Geophysics* (1971), a thorough and reliable work that covers several subjects not adequately treated by Stacey; J.A. JACOBS, R.D. RUSSELL, and J.T. WILSON, *Physics and Geology* (1959), a classic text to introduce geology students to the underlying physics with emphasis on those aspects that have direct bearing on large-scale geology; S.K. RUNCORN *et al.* (eds.), *International Dictionary of Geophysics,* 1 vol. and 1 vol. of maps (1967), articles contributed by nearly a thousand authorities, with short bibliographies following the articles. JOHN W. HARRINGTON, *Dance of the Continents* (1983), an introduction to historical geology.

(W.M.E.)

The Moon: BEVAN M. FRENCH, *The Moon Book* (1977); C.W. ALLEN, *Astrophysical Quantities,* 3rd ed. (1973); HER MAJESTY'S NAUTICAL ALMANAC OFFICE, *Explanatory Supplement to the Astronomical Ephemeris and the American Ephemeris and Nautical Almanac* (1963). Works focusing on the motion and rotation of the Moon include D.W.G. ARTHUR, "Selenography," in G.P. KUIPER and B.M. MIDDLEHURST (eds.), *The Moon, Meteorites and Comets* (1963); JACQUES HENRARD, "A New Solution to the Main Problem of the Lunar Theory," in *Celest. Mech.,* 19:337–355 (1979); K. GARTHWAITE *et al.,* "A Preliminary Special Perturbation Theory for the Lunar Motion," in *Astron. Jour.,* 75:1133–39 (1970), the first published use of a numerical lunar orbit, showing the inadequacy of Brown's theory; MARTIN A. SLADE, *The Orbit of the Moon* (1971), including mathematical details of numerical integration of the lunar motion; J. DERRAL MULHOLLAND, "Scientific Achievements from Ten Years of Lunar Laser Ranging," in *Rev. Geophys. Space Phys.,* 18:549–564 (1980) and "How High the Moon: A Decade of Laser Ranging," in *Sky & Telescope,* 60:274–279 (October 1980); J. DERRAL MULHOLLAND, "The Rotation of the Moon," in *Bull. Cl. Sciences Acad. Roy. Belgique,* Ser. V, 55:1088–1113 (1974), a comprehensive overview; STANTON PEALE, "Generalized Cassini's Laws," in *Astron. Jour.,* 74:483–489 (1969), an analysis of the resonance; and ROGER J. CAPPALLO, *The Rotation of the Moon* (1980), which applies numerical integration to the Moon's libration. The use of different kinds of observations is described in R.W. KING, C.C. COUNSELMANN III, and IRWIN I. SHAPIRO, "Lunar Dynamics and Selenodesy: Results from Analysis of VLBI and Laser Data," in *Jour. Geophys. Res.,* 81:6251–56 (1976); J. DERRAL MULHOLLAND (1980, *op. cit.*); and LESLIE V. MORRISON, "An Analysis of Lunar Occultations in the Years 1943–1974 for Corrections to the Constants in Brown's Lunar Theory," in *Mon. Not. Roy. Astron. Soc.,* 187:41–82 (1979). THOMAS C. VAN FLANDERN, "Is Gravity Getting Weaker?" *Scientific American,* 234:44–52 (February 1976), an explanation of the effect of gravity on the Moon's orbit. The mass of the Moon is discussed in A.J. FERRARI *et al.,* "Geophysical Parameters of the Lunar Orbit," in *Jour. Geophys. Res.,* 85:3939–51 (1980). For the gravity field of the Moon, see WILLIAM M. KAULA, *An Introduction to Planetary Physics* (1968); P.M. MULLER and W.L. SJOGREN, "Mascons: Lunar Mass Concentrations," in *Science,* 161:680–684 (1968), which announced the discovery of mascons; and A.J. FERRARI *et al., op. cit.,* which gives an informative description of the harmonic field.

Types of lunar features are covered in GILBERT FIELDER, *Lunar Geology* (1965); ELBERT A. KING, *Space Geology* (1976); JOSIAH E. SPURR, *Geology Applied to Selenology:* vol. 1, *The Imbrium Plain Region of the Moon,* vol. 2, *Features of the Moon,* vol. 3, *Lunar Catastrophic History,* and vol. 4, *The Shrunken Moon* (1944–49); GILBERT FIELDER and LIONEL WILSON (eds.), *Volcanoes of the Earth, Moon, and Mars* (1975); NICHOLAS M. SHORT, *Planetary Geology* (1975); THOMAS A. MUTCH, *Geology of the Moon* (1970); PETER H. SCHULTZ, *Moon Morphology* (1976); ROYAL SOCIETY OF LONDON, *The Moon* (1977); G.P. KUIPER (ed.), *Photographic Lunar Atlas* (1960, with suppl., *Orthographic Atlas of the Moon*); E.A. WHITAKER *et al., Rectified Lunar Atlas* (1963); *The Times Atlas of the Moon* (1969); J.E. BALDWIN, "Thermal Radiation from the Moon and the Heat Flow Through the Lunar Surface," *Mon. Not. R. Astr. Soc.,* 122:513–522 (1961); ROBERT HOOKE, *Micrographia* (1st ed., 1665; 2nd issue, 1667, reprinted 1938); JAMES NASMYTH and JAMES CARPENTER, *The Moon,* 4th ed. (1903); WILLIAM H. PICKERING, "Lunar and Hawaiian Physical Features Compared," *Mem. Am. Acad. Arts Sci.,* 13:151–182 (1908); J.E. SPURR (*op. cit.*); G.K. GILBERT, "The Moon's Face," *Bull. Phil. Soc. Wash.,* 12:241–292 (1893); RALPH B. BALDWIN, *The Measure of the Moon* (1963), and *A Fundamental Study of the Moon* (1965); BRIAN MASON and WILLIAM G. MELSON, *The Lunar Rocks* (1970); G.M. BROWN, "Geochemistry of the Moon," *Endeavour,* 30:147–151 (1971); JOHN F. LINDSAY, *Lunar Stratigraphy and Sedimentology* (1976); ALFRED A. LEVINSON and S. ROSS TAYLOR, *Moon Rocks and Minerals* (1971); TYPHOON LEE, "New Isotopic Clues to Solar System Formation," in *Rev. Geophys. and Space Phys.,* 17:1591–1611 (1979); ALAN H. COOK, *Physics of the Earth and Planets* (1973); S. ROSS TAYLOR, *Lunar Science: A Post-Apollo View* (1975). The results of intensive studies of returned lunar rocks have been published in *Science,* vol. 167, no. 3918 (1970); in the reports of the Lunar Science Conferences held since 1970; and by BRIAN MASON and WILLIAM G. MELSON (*op. cit.*). Discussions on the evolution of the Moon's orbit are found in SIR GEORGE H. DARWIN, *The Tides and Kindred Phenomena in the Solar System* (1898, reprinted 1962), the classic first work on the subject; W.H. MUNK and G.J.F. MACDONALD, *The Rotation of the Earth* (1960), a landmark in the study of dissipation in the solar system; G.D. ROSENBERG and S.K. RUNCORN (eds.), *Growth Rhythms and the History of the Earth's Rotation* (1975), a good summary of the use of ancient fossil records, but with obsolete astronomical discussions; KURT LAMBECK, *The Earth's Variable Rotation* (1980). Theories of lunar origin are dealt with by BRIAN G. MARSDEN and A.G.W. CAMERON (eds.), *The Earth-Moon System* (1966); S. FRED SINGER, "The Origin of the Moon and Geophysical Consequences," in *Geophys. Jour. Roy. Astron. Soc.,* 15:205–226 (1968); WILLIAM M. KAULA and A.W. HARRIS, "Dynamics of Lunar Origin and Orbit Evolution," in *Rev. Geophys. Space Phys.,* 13:363–372 (1975).

(J.D.Mu./G.Fi.)

Mars: Early ideas on the possibility of life on Mars are summarized in H. STRUGHOLD, "Synopsis of Martian Life Theories," in *Advances in Space Science and Technology,* 9:105–122 (1967); post-Viking views are discussed in HAROLD P. KLEIN, "The Viking Mission and the Search for Life on Mars," in *Reviews of Geophysics and Space Physics,* 17:1655–62 (October 1979). The nature of Martian phenomena as understood before 1969, including several enlightening chapters on the possibilities of extraterrestrial life, are discussed in SAMUEL GLASSTONE, *The Book of Mars* NASA SP-179 (1968). The era of space exploration is covered by several books and collections of papers, each dedicated to an exposition of the subject at the conclusion of a project. Mariners 6 and 7 are discussed in N.W. CUNNINGHAM and H.M. SCHURMEIER, *Mariner-Mars 1969: Preliminary Report* NASA SP-225 (1969). An illustrated summary of the results of the Mariner 9 orbiter mission is to be found in *Mars as Viewed by Mariner 9* NASA SP-329 (1974). In addition, *Icarus,* 17:289–561 (October 1972) is dedicated to a collection of original papers by Mariner 9 experimenters. A synthesis of views on the nature of the planet between Mariner 9 and Viking is developed in THOMAS A. MUTCH *et al., The Geology of Mars* (1976). The Viking results are extensive; the original papers describing the results of the mission are to be found in the following entire issues of *Journal of Geophysical Research,* 84:2795–3007 (June 1979), 84:7909–8519 (December 1979), and 82:3959–4680 (September 1977). For an excellent review of the geologic evolution of Mars, see RAYMOND E. ARVIDSON, KENNETH A. GOETTEL, and CHARLES M. HOHENBERG, "A Post-Viking View of Martian Geologic Evolution," in *Reviews of Geophysics and Space Physics,* 18:565–603 (August 1980). An overview of the morphology of surface structures found by Viking and Mariner 9 is MICHAEL H. CARR, "The Morphology of the Martian Surface," in *Space Science Reviews,* 25:231–284 (1980). A detailed account of factual knowledge about the Martian satellites is JOSEPH VEVERKA and JOSEPH A. BURNS, "The Moons of Mars," in *Annual Review of Earth and Planetary Sciences,* 8:527–558 (1980). Knowledge of the meteorology of the Martian atmosphere is described in CONWAY B. LEOVY, "Martian Meteorology," in *Annual Review of Astronomy and Astrophysics,* 17:387–413 (1979). ARTHUR KOESTLER, *The Sleepwalkers: A History of Man's Changing Vision of the Universe* (1959), provides an excellent account of the role of Mars in the work of Brahe and Kepler. Maps of Mars can be obtained from the superintendent of the U.S. Geological Survey, Washington, D.C.

(M.J.S.B.)

Jupiter: A discussion of telescopic observations of Jupiter consisting of descriptions of motions, colours, and transformations observed in the Jovian cloud layers may be found in B.M. PEEK, *The Planet Jupiter* (1958). An extensive compilation of papers on all aspects of the Jovian system through 1975 is presented in T. GEHRELS (ed.), *Jupiter* (1976). The Pioneer missions are described in R.O. FIMMEL, W. SWINDELL, and E. BURGESS, *Pioneer Odyssey: Encounter with a Giant* (1974). An early summary of information about the satellites appears in J.A. BURNS (ed.), *Planetary Satellites* (1977). A later account is given in D. MORRISON (ed.), *Satellites of Jupiter* (1981), which includes a discussion of the Voyager results pertaining to the satellites. Summary articles of other Voyager discoveries may be found in the following journals: *Science,* 204:945–957, 960–1008 (June 1, 1979), and 206:925–996 (Nov. 23, 1979); *Nature,* 280:725–806 (Aug. 30, 1979); *J. Geophys. Res.,* 86 (Sept. 30, 1981), the entire issue being devoted to "Voyager Missions to Jupiter." The Voyager spacecraft and their experiments are described in *Space Science Reviews,* vol. 21, no. 2 and 3 (1977). An overview of this mission is given in D. MORRISON and J. SAMZ, *Voyage to Jupiter* (1980); and many of the results and their interpretation are presented in articles in J.K. BEATTY, B. O'LEARY, and A. CHAIKIN (eds.), *The New Solar System* (1981).

(T.C.O.)

Saturn: A.F.O'D. ALEXANDER, *The Planet Saturn: A History of Observation, Theory and Discovery* (1962, reissued 1980), contains a list of all important papers published up to 1962. For an overview of information gathered by the Voyager space probes, see DAVID MORRISON, *Voyages to Saturn* (1982). See also *Science,* 212:159–243 (April 10, 1981); *Science,* 215:499–594 (Jan. 29, 1982); and ANDREW P. INGERSOLL, "Jupiter and Saturn," *Scientific American,* 245:90–108 (December 1981).

(P.Mo./Ed.)

Uranus: A comprehensive and detailed summary of essentially all the serious work done on Uranus from its discovery through the early 1960s is ARTHUR F.O.'D. ALEXANDER, *The Planet Uranus* (1965). More technical treatments are found in the articles by Brouwer and Clemence, Wildt, and Harris in *Planets and Satellites,* ed. by G.P. KUIPER and B.M. MIDDLEHURST (1961). The mass of Uranus may also be determined by its perturbations on the orbit of Saturn; this technique is summarized by W.J. KLEPCZYNSKI, P.K. SEIDELMANN, and R.L. DUNCOMBE in "The Masses of Jupiter and Saturn," *Astr. J.,* 75:739–742 (1970). Details of the atmosphere of Uranus may be found in "The Atmospheres of the Outer Planets and Satellites," by L. TRAFTON in *Reviews of Geophysics and Space Physics,* 19:43–89 (1981). Findings by the Voyager 2 flyby are reported in several articles in *Science,* vol. 233, no. 4759 (July 4, 1986), including information on the planet's atmosphere, rings, and satellites; and in TORRENCE V. JOHNSON, ROBERT HAMILTON BROWN, and LAURENCE A. SODERBLOM, "The Moons of Uranus," *Scientific American,* 256(4):48–60 (April 1987).

(H.Sm.)

Neptune: B.A. GOULD, *Report on the History of the Discovery of Neptune* (1850), is a classic work that provides detailed accounts. R.A. LYTTLETON, *Mysteries of the Solar System* (1968), includes a chapter on Neptune's discovery, along with a discussion of its orbits. Two articles, with illustrations, for the general reader are D.M. HUNTEN, "The Outer Planets," *Scientific American,* 233:130–140 (September 1975); and M.J.S. BELTON, "Uranus and Neptune," *Astronomy,* 5:6–17 (February 1977). Findings and pictures from the Voyager 2 flyby are included in JUNE KINOSHITA, "Neptune," *Scientific American,* 261(5):82–91 (November 1989); a set of articles in *Science,* vol. 246, no. 4936 (Dec. 15, 1989); and two articles in *Sky & Telescope,* vol. 79, no. 2 (February 1990): "Neptune and Triton: World Apart," pp. 136–145; and KELLY BEATTY, "Getting to Know Neptune," pp. 146–155.

(G.E.T./Ed.)

Pluto: W.G. HOYT, *Planets X and Pluto* (1980), a good account of the history of the prediction, discovery, and subsequent study of Pluto; A.J. WHYTE and H.A. WISE, *The Planet Pluto* (1980), a comprehensive collection of references to, and summaries of, original research reports. Pluto's properties as well as its satellite are described in J.W. CHRISTY and R.S. HARRINGTON, "The Satellite of Pluto," *Astr. J.,* 83:1005–08 (1978); W. BAADE, "The Photographic Magnitude and Color Index of Pluto," *Publs. Astr. Soc. Pacif.,* 46:218–221 (1934); M.F. WALKER and R.H. HARDIE, "A Photometric Determination of the Rotational Period of Pluto," *Publs. Astr. Soc. Pacif.,* 67:224–231 (1955); L.E. ANDERSSON and J.D. FIX, "Pluto, New Photometry and a Determination of the Axis of Rotation," *Icarus,* 20:279–283 (1973); I. HALLIDAY *et al.,* "An Upper Limit for the Diameter of Pluto," *Publs. Astr. Soc. Pacif.,* 78:113–124 (1966); S.J. ARNOLD, A. BOKSENBERG, and W.L.W. SARGENT, "Measurement of the Diameter of Pluto by Speckle Interferometry," *Astrophysical Journal,* 234:L159–L163 (1979). CLYDE W. TOMBAUGH, *Out of the Darkness: The Planet Pluto* (1980), an account of its discovery by its discoverer.

(R.H.H.)

Asteroids: Summary articles can be found in CLARK R. CHAPMAN, "The Nature of Asteroids," *Scientific American,* 232(1):24–33 (January 1975), and "Asteroids," ch. 10 in J. KELLY BEATTY, BRIAN O'LEARY, and ANDREW CHAIKIN (eds.), *The New Solar System,* 2nd ed. (1982), pp. 97–104. See also GEORGE O. ABELL, DAVID MORRISON, and SIDNEY C. WOLFF, "The Asteroids," ch. 19 in *Exploration of the Universe,* 5th ed. (1987). Review and research papers are collected in TOM GEHRELS and MILDRED SHAPLEY MATTHEWS (eds.), *Asteroids* (1979), including three broad review chapters; and C.-I. LAGERKVIST and H. RICKMAN (eds.), *Asteroids, Comets, Meteors* (1983), and *Asteroids, Comets, Meteors II* (1986). International reports of research on asteroids, as well as comets and meteorites, can be found in *Astrophysical Journal* (semimonthly); and *Astronomy and Astrophysics* (34/year).

(E.F.T.)

Comets: General introductory works are FRED L. WHIPPLE, *The Mystery of Comets* (1985, reprinted 1986); and ROBERT D. CHAPMAN and JOHN C. BRANDT, *The Comet Book: A Guide for Return of Halley's Comet* (1984), a historical treatment. More advanced are JOHN C. BRANDT and ROBERT D. CHAPMAN, *Introduction to Comets* (1981); and K.S. KRISHNA-SWAMY, *Physics of Comets* (1986), including a section on spectroscopy. LAUREL L. WILKENING and MILDRED SHAPLEY MATTHEWS (eds.), *Comets* (1982), is a definitive collection of essays covering all aspects at a technical level. See also D.A. MENDIS, H.L.F. HOUPIS, and M.L. MARCONI, "The Physics of Comets," *Fundamentals of Cosmic Physics,* vol. 10, no. 1–4 (1985); and A.H. DELSEMME (ed.), *Comets, Asteroids, Meteorites: Interrelations, Evolution and Origins* (1977). BRIAN G. MARSDEN, *Catalog of Cometary Orbits,* 5th ed. (1986), covers 1,187 cometary orbits, with detailed references and notes; a complementary volume is GARY KRONK, *Comets: A Descriptive Catalog* (1984).

(A.H.De.)

Meteoroids, meteors, and meteorites: Introductory information can be found in ROBERT T. DODD, *Thunderstones and Shooting Stars: The Meaning of Meteorites* (1986); ROBERT HUTCHISON, *The Search for Our Beginning: An Enquiry, Based on Meteorite Research, into the Origin of Our Planet and of Life* (1983); and JOHN A. WOOD, *Meteorites and the Origin of Planets* (1968). More advanced treatments are JOHN T. WASSON, *Meteorites: Their Record of Early Solar-System History* (1985), and *Meteorites: Classification and Properties* (1974); V.A. BRONSHTEN, *Physics of Meteoric Phenomena* (1983; originally published in Russian, 1981); and ROBERT T. DODD, *Meteorites: A Petrologic-Chemical Synthesis* (1981). A descriptive and historical treatment of iron meteorites, including beautiful photographs, is VAGN F. BUCHWALD, *Handbook of Iron Meteorites, Their Distribution, Composition, and Structure,* 3 vol. (1975). H.H. NININGER, *Out of the Sky: An Introduction to Meteorites* (1952, reprinted 1959), provides firsthand experiences of fall phenomena on a nontechnical level. See also D.E. BROWNLEE, "Cosmic Dust: Collection and Research," *Annual Reviews of Earth and Planetary Sciences,* 13:147–173 (1985). A catalog of known meteorites, including data regarding their fall, is A.L. GRAHAM, A.W.R. BEVAN, and R. HUTCHISON (eds.), *Catalogue of Meteorites,* 4th ed. rev. and enl. (1985). There are two journals devoted to papers on meteorites and related bodies: *Meteoritika* (annual), published in the U.S.S.R.; and *Meteoritics* (quarterly). Many papers on meteorites are published in *Geochimica et Cosmochimica Acta* (monthly).

(G.W.We.)

Sound

A human being speaking is said to produce sound, and anyone nearby who has adequate hearing ability is said to hear the sound. The experience is a manifestation of a phenomenon of great significance to all persons, who are immersed in sound stimuli that influence all of their activities.

One suggestion of the importance of sound is the large number of words in every language that are descriptive of the different kinds of sound within human awareness. In nature, the storm thunders, the gale howls, surf roars and pounds, wind whistles and moans, trees whisper, rain patters, and running water ripples and gurgles. The vocabulary descriptive of the sounds of living things is even richer: the barking and snarling of dogs, the mewing of cats, crowing of fowl, roaring of lions, hissing of snakes, lowing and bellowing of cattle, blatting of goats, grunting of pigs, chirping of birds and insects, screaming of gulls, crying and yelling of infants, not to mention the sneezing, wheezing, and snoring of humans.

In addition to the racket all around them, over which they have little or no control, humans are subjected to an almost limitless variety of sounds that they themselves produce. There are the unpleasant blast of an explosion, roar of engines, rumble of traffic, whine of jet planes; the more agreeable if distracting whir of machinery and ticking of clocks; or the melody and harmony of music.

The purpose of this article is to set forth the physicist's explanation of the phenomena of sound, though reference will be made, when necessary, to other disciplines that play a considerable role in human production of and reaction to sound. The branch of physics most concerned with understanding sound is called acoustics, from the Greek word meaning hearing. The various applications of acoustics, including the recording and reproduction of sound and the control of noise, are reviewed.

Inevitably man's major association with sound is through the human ear. Though the anatomical and physiological details of this interesting and important sensory organ are dealt with in the article on SENSORY RECEPTION, a few remarks here will set the stage for the further discussion of sound. The normal human ear is a remarkably sensitive organ, and it has been estimated that as little as 10^{-11} erg or about 10^{-25} kilowatt-hour of sound energy is enough to excite it and produce recognition of the sound.

The ear is built ruggedly enough to tolerate a wide range of sound energy, but hearing becomes painful, and irreversible ear damage may ensue if the amount of sound energy reaches 10^{-11} kilowatt-hour, which is, however, 10^{-14} times that of the minimum audible energy level. These figures must be taken as order of magnitude values and not as specific results of precise measurement. In any case, human ears differ from individual to individual and vary even in the same individual with age and environmental conditions. The noise in large cities takes a heavy toll in physiological damage and psychological impact.

The fact that the human being has two ears enables a hearer to judge the direction from which sound in the open reaches him (the binaural effect) or to separate a signal from ambient noises. A fundamental characteristic of what may be called regular sounds, like musical notes, is their frequency—*i.e.,* the number of cycles per second expressed in hertz (one cycle per second), abbreviated Hz, of the causative vibration. Low-frequency sound is heard as low pitch, and higher frequencies as correspondingly higher pitch (see MECHANICS: *Vibrations*). The normal human ear is sensitive to sound within the range of about 20 Hz to 20,000 Hz, with some variation from individual to individual. Advancing age usually leads to a decrease in the upper limit. Sound of less than 20 Hz is called infrasonic; that of more than 20,000 Hz, ultrasonic. These inaudible sounds are actually of greater interest in modern physics and in their technical application than those in the sonic or audible range.

The article is divided into the following sections:

Historical survey

The fact that a solid object struck in air produces sound must have been observed from the earliest times. The additional observation that under certain conditions the sounds so produced are particularly agreeable to the ear furnished the basis for the creation of music, which must have taken place long before the beginning of recorded history. Thus the germ of the idea that the pitch of a musical sound depends somehow on the frequency of vibration of the sound-producing object is usually attributed to the 6th-century-BC Greek philosopher Pythagoras.

Further, that sound transmission through the air has something to do with the motion of the air was recognized at least by Aristotle, though it is not clear that he grasped the idea that the motion of air involved in sound is a wave motion rather than a streaming of air from the sound source to the listener. The Roman architectural engineer Marcus Vitruvius Pollio of the 1st century BC understood the wave concept and actually drew analogies between sound travelling in air and surface waves in water, but his ideas were more or less forgotten during the Middle Ages, when a corpuscular, or atomic, theory of the transmission of sound was widely held. In the 17th century, a French philosopher and mathematician, Pierre Gassendi, in reviving the atomic theory, attributed the propagation of sound to the emission and transfer of a stream of very small, invisible particles, which, after moving through the air, are able somehow to affect the ear.

Irrespective of the theory of sound transmission, the question of the speed of propagation began to engage attention in the early 17th century. Gassendi in 1635 made one of the first recorded measurements of the velocity of sound in air. His experiments involved firearms and were based on the assumption that the flash of light is seen by the observer instantaneously. Timing the delay between powder flash and the noise of the discharge, he determined a speed of sound the equivalent of 478.4 metres per second (1,569 feet per second, or 0.3 mile per second). This was shown to be too high by other observers during the 17th and 18th centuries. It was about 1750 that the first measurements made in windless, open air that can be considered at all precise were carried out under the direction of the Academy of Sciences of Paris, with a cannon as the source of sound. The result, calibrated for 0° C (32° F), was 332 metres or 1,089 feet per second, or 0.206 mile per second. Careful repetitions during the succeeding two centuries resulted in figures differing from this last result by less than 1 percent.

Velocity of sound in liquids

The first serious attempt to measure the velocity of sound in a liquid, such as water, was apparently made in the 19th century by Daniel Colladon, a Swiss physicist who, with the assistance of a French mathematician, Charles-François Sturm, conducted such studies in Lake Geneva. The average velocity as found by their measurements was 1,435 metres (4,707 feet) per second at 8° C (46° F). It had already been established around 1808 that the velocity of sound in an iron pipe greatly exceeds that in either air or water and in fact is of the order of 5,000 metres (16,000 feet) per second.

Though the wave theory of sound had been in existence in a qualitative way for many centuries, the first attempt to apply it to a theoretical derivation of the velocity of sound was probably made by Sir Isaac Newton, who, on the basis of arbitrary assumptions, reported that the velocity of sound in air (or any gaseous fluid) should be equal to the square root of the ratio of the atmospheric pressure to the density.

Newton's derivation was criticized for giving too low a speed when appropriate pressure and density values were substituted in his equation. This difficulty was resolved by a French mathematician, Pierre-Simon Laplace, who pointed out that the Newtonian formula was based on the assumption that the compressions and rarefactions associated with the propagation of sound in air take place at constant temperature. Laplace felt it to be more reasonable to assume that the compressions and rarefactions in sound propagation do not take place at constant temperature. His assumption led to the correction of Newton's formula by multiplying the ratio of the pressure to the density by that of the specific heats (heat required to raise the temperature of one gram of gas one degree Celsius) of the gas at constant pressure and constant volume, respectively. Thus corrected, the result agreed with that from experiments at ordinary frequencies.

Mathematical analysis of wave motion

In the latter half of the 18th century and the first quarter of the 19th century, many scientists tried to theorize about mechanical waves in continuous media, mainly on the basis of a mathematical expression for wave motion proposed by d'Alembert in 1747. Much of this work was stimulated by the work of two mathematicians, the Swiss Leonhard Euler and Joseph-Louis Lagrange in France. These mathematical considerations made it possible to understand the vibrations of sound sources like strings, rods, membranes, plates, organ pipes, and musical instruments in general. Though earlier work—based largely on empirical or relatively simple mathematical analysis—had been carried out, it required the methods of the calculus to understand in any precise way the behaviour of even the simplest types of sound vibrators.

The ear early attracted scientific attention as the receiver of sound for human beings. A French physicist, Félix Savart, in 1830 attempted to establish the frequency limits of audibility and obtained results not too different from those of modern measurements made with more sophisticated equipment. The corresponding problem of the threshold of audibility was attacked around 1870 by the Viennese physicists, August Toepler and Ludwig Boltzmann, who arrived at the figure 10^{-11} watt (one watt is a unit of power equal to one joule of work per second) per square centimetre, considerably in excess of the modern figure of 10^{-16} watt per square centimetre but indicating, nevertheless, the remarkable sensitivity of the ear as a sound receiver. Further important studies on the ear were made in Germany by Georg Simon Ohm, a physicist, who in 1843 theorized that the ear is able to analyze any complex sound into a combination of simple tones in terms of which the sound may be expressed mathematically by the use of a theorem of a French mathematician, Jean Fourier (see ANALYSIS IN MATHEMATICS). Ohm's work was further extended by Hermann von Helmholtz, a German physicist and physiologist, whose book, *Sensations of Tone,* marked the beginning of physiological acoustics.

Though early scientific work in the production and reception of sound in closed spaces (*i.e.,* room or architectural acoustics) was reported around 1850, the modern quantitative foundation of the subject dates from the work of a United States physicist, Wallace Clement Sabine, who in 1900 found a law to connect reverberation (echo) time in a closed room with the volume of the room and the amount of absorbing material. His finding made applied architectural acoustics possible.

Rayleigh's contributions

In another sense, however, the modern period in acoustics may be said to date from the publication in 1877 by the English physicist Lord Rayleigh (John William Strutt) of *The Theory of Sound,* in which the whole basic theory of all aspects of the production, propagation, and reception of sound was presented in rigorous fashion, together with much of the relevant experimental material. Most of 20th-century acoustics is an exploitation and development of the fundamental ideas in Rayleigh's treatise with the use of more sophisticated equipment.

The nature and properties of sound

THE GENERAL NATURE OF SOUND WAVES

Wave motion, velocity, and intensity. As has been noted, it was early recognized that sound is propagated as wave motion from a sound-producing object. It is important to distinguish this motion from that of such things as automobiles, projectiles, and flowing water. When a person opens his mouth and utters a sound, he moves the air in front of his mouth, but he does not push away all of the air in his vicinity. Instead, the motion is a squeezing one that compresses a small mass of air near his mouth into less than its normal volume. Being elastic, air when compressed tends to expand again when the compressing influence is removed. In expanding, the first mass of air moves some of the air adjacent to it, which is thereupon compressed in turn and in turn expands and pushes against another adjacent bit of air. So the energy of the original compressing motion is passed on in temporal succession through the space surrounding the speaker. When a portion of it is intercepted by a receiver of sound like the human ear, it finally yields up some of itself to the receiver as a stimulus. The response in the case of the human ear is audition or hearing.

A sound wave in air or any fluid medium is thus a compressional wave. The simplest example is the pressure pulse produced by the firing of a pistol or a short speech ejaculation: it takes time for such a pulse to travel through the medium, and hence such a wave has a finite velocity. The velocity of sound in air is about 344 metres (1,128 feet) per second at 20° C (68° F). It increases with temperature in a way that is treated below in the section on the properties of sound waves. It is independent of the pressure or density of the air over wide ranges. Because the velocity of sound in air is so much less than the velocity of light (3×10^8 metres per second), the sound of a distant cannon is heard sometime after the flash or the puff of smoke is seen.

The everyday distinction between loud and soft sounds rests on the differences in the magnitude of the pressure changes involved in sound-wave propagation. In normal conversation the pressure of the air in front of the mouth of the speaker is changed at most by only about one millionth of the normal atmospheric pressure (*i.e.,* by only about one dyne per square centimetre out of 10^6 dynes per square centimetre; the dyne, a unit of force, is that force required to change the velocity of a mass of one gram by one centimetre per second each second). At the same time the accompanying motion that produces the compression corresponds to a fluctuating air-flow velocity with a maximum value of only about 0.02 centimetre per second (about 0.008 inch per second).

A sound wave represents the transmission of mechanical energy. An average rate of transfer of only 10^{-16} watt per square centimetre is enough to produce the hearing of an identifiable sound in a normal young person. This means that mechanical energy of only about 10^{-11} erg or about 10^{-25} kilowatt-hour can produce in the ear an identifiable sound stimulus. The average power transmission in a sound wave per unit area of the medium perpendicular to the direction of the wave is called the intensity. The unit for this quantity is the bel, but in practice it is common to use a unit ten times smaller, the decibel (abbreviated dB). If the intensity of a given sound in absolute units is I and the minimum audible intensity is I_0, the sound is said to be a number of decibels above the minimum audible equal to ten times the logarithm to the base ten of I divided by I_0. The decibel scale is thus relative to some arbitrary level, but the latter is commonly taken as the minimum audible. On this basis, ordinary conversational speech has an intensity of 60 decibels at a distance of about three feet from the mouth. Traffic at a busy street intersection can provide an average of 75 decibels, and a boiler factory can reach 130 decibels.

finition
decibel

As a source of sound in the open, such as an airplane, moves away from the listener, it is heard less distinctly. This is basically a matter of geometry. The intensity of a sound wave from a point source varies inversely as the square of the distance from the source. A point source in the open spreads sound energy in all directions, and the farther from the source, the greater is the spherical area through which the same sound energy flows per unit time. Because the area of a sphere increases as the square of the distance from the centre, the intensity of the wave, which is the average flow of energy per unit area, must decrease inversely with the square of the distance. Actual observation indicates that the falling off of intensity with distance in the open is always greater than this purely geometrical decay. This effect is attributed to a variety of ways in which energy is dissipated—reflection, refraction, diffraction, and scattering—as well as to the absorption of the sound caused by the viscosity and other similar properties of the medium. The diminution of sound intensity from a localized source is further complicated if the source is on the ground or underwater near the surface, and even more so if it is inside a closed room or auditorium.

Frequency, pitch, and wavelength. An air siren produces sound waves by interrupting jets of air at different rates, by means of a rotating disk with orifices in its periphery. The greater the number of separate puffs of air per unit time, the higher is the perceived pitch of the sound. This leads to a consideration of periodic sound waves, which are much the easiest to study analytically (see the section on the properties of sound waves). Waves are called harmonic waves because they result from a source of sound in which the disturbance repeats itself regularly. The simplest periodic sound, a harmonic sound wave, is characterized by a definite frequency. The frequency of the sound is defined as the number of times per second the disturbance at any point in the medium transmitting the wave is repeated.

The relation between pitch, as perceived by the human ear, and physical frequency is more complex in the case of the human voice or ordinary musical instruments than it is for a pure harmonic wave as emitted by a carefully made tuning fork or loudspeaker, activated by an electronic oscillator under precise frequency control. The fork or speaker is said to produce a pure tone of definite single frequency. In such a harmonic wave the time for one complete cycle of pressure change at any point in the medium is called the period of the wave, which is the reciprocal of the frequency, *i.e.,* one divided by the frequency in cycles per second, or hertz.

Meaning
of pure
tone

The sounds of all frequencies are not heard equally well by humans. An object vibrated in air at a frequency of less than about 20 hertz will create a compressional wave that will not be heard as sound by most normal ears. Such waves are said to be infrasonic. An object vibrating at a higher frequency than this will tend to cause audible sound if the intensity is sufficiently great. When the frequency of a compressional wave in air is increased beyond 20,000 hertz (20 kilohertz), the audibility ceases for most people, even if the intensity becomes great, though an uncomfortable nervous sensation is experienced by many. With advancing age of the listener, this upper frequency threshold decreases markedly. These inaudible high-frequency sound waves, commonly called ultrasonic, are periodic stress waves that produce a deformation of the medium through which they are passing. They play an important part in modern acoustical science and engineering. By means of appropriate sources (see below *Sources of sound*) it has been possible to produce ultrasonic radiation of frequency of about 10^{10} hertz. Such waves, however, are rapidly absorbed in fluid media.

Every harmonic sound wave is characterized by a definite wavelength; that is, the distance between identical successive points in a spreading wave at which the disturbance is exactly the same and doing the same thing (*i.e.,* increasing or decreasing the compression of the fluid). At two such successive points the disturbances are said to be in phase with one another. Because the disturbance (*e.g.,* the onset of compression) in a wave travels a distance of one wavelength in one period, the product of wavelength and frequency equals the wave velocity. This important relationship between wavelength, frequency, and velocity, which holds for harmonic waves of all kinds (*i.e.,* sound waves, light waves, water waves, etc.), means that, the velocity of all acoustic waves being constant in any medium, high-frequency acoustic waves in a given medium have smaller wavelengths than low-frequency sounds. Thus, in air at 20° C (68° F), a wave of frequency 1,000 hertz has a wavelength of 0.344 metre (1.128 feet), whereas a wave frequency 10^6 hertz or one megahertz has a wavelength of 0.344 millimetre (0.0135 inch). The corresponding values in water are, respectively, 1.5 metres (five feet) and 1.5 millimetres (0.06 inch).

Few sounds in daily experience are pure tones characterized by a single frequency; most can be considered as arising from more or less complicated combinations of harmonic waves, each with its own frequency. One or more frequencies in such a mixture will often predominate (*i.e.,* have greater intensity) and help give the sound its observed quality.

THE PROPERTIES OF SOUND WAVES

Wave functions. The simplest of all wave motions to imagine are waves that travel along a flexible rope when one end is flicked. This type of wave is called transverse because the disturbance of the medium (in this case the rope) takes place at right angles to the direction of travel of the wave itself. A transverse wave differs physically from a sound wave, in which the disturbance (compression) takes place in the same direction as the wave itself; thus sound waves are called longitudinal. Each particle of the medium through which a longitudinal wave passes is displaced from its equilibrium position along the direction of the wave propagation.

The longitudinal nature of sound waves

Though sound waves are not necessarily periodic, those of greatest interest are characterized by frequency and wavelength (*i.e.,* are periodic). Like transverse waves, these longitudinal sound waves can be expressed in a harmonic equation in which the displacement, or disturbance (ξ), is equal to the amplitude (A) times the sine of two pi (π), times a quantity equal to the frequency (f), times the elapsed time (t), minus the distance (x) from the origin divided by the wavelength (λ), or:

$$\xi = A \sin 2\pi\,(ft - x/\lambda). \qquad (1)$$

The amplitude of the wave represents the maximum value of the disturbance. This equation can be used to determine the size of the disturbance for any value of the time (t) or position (x). In the equation, the quantity inside the parentheses can take successive values of 0, 1/8, 2/8, 3/8, etc., so that the disturbance is equal to the amplitude times the sine of 0, $\pi/4$, $2\pi/4$, $3\pi/4$, etc.; that is, the disturbance is alternately equal to plus and minus the amplitude and passes through zero in between.

Concept of excess pressure

For a harmonic sound wave in a fluid, the displacement ξ is more conveniently replaced by the excess pressure (p_e) produced by the wave. Excess pressure may be defined as the difference between the actual pressure (p) in the medium at any place and time and the normal pressure (p_0) in the originally undisturbed medium. Thus, symbolically, $p_e = p - p_0$. This excess pressure for a harmonic wave varies in space and time like ξ as shown in equation (1).

Velocity of sound waves in fluids and solids. A mathematical study of the propagation of a compressional wave in a fluid (gas or liquid) also shows that the velocity (V) of the wave is equal to the square root of the ratio of excess pressure to excess density:

$$V = \sqrt{p_e/\rho_e}, \qquad (2)$$

in which p_e is the excess pressure in the medium produced by the wave and ρ_e is the associated excess density (difference between the actual and normal, $\rho_e = \rho - \rho_0$). Whenever a fluid is squeezed by the application of extra pressure, the density goes up, and when the excess pressure becomes negative the density goes down. Equation (2) can be used to provide expressions from which the velocity V may be calculated in special cases. Hooke's law of elasticity (see MECHANICS) states that elastic stress is always proportional to elastic strain; that is, stress divided by strain is equal to a quantity B, in which B represents a proportionality constant. In the case of a compressional wave in a fluid, provided these quantities are not too great, one can write:

$$\frac{p_e}{\rho_e/\rho_0} = B, \qquad (3)$$

in which ρ_0 is the average density of the undisturbed medium. The quantity B, called the bulk modulus of the medium, is the coefficient of proportionality between the excess pressure p_e (the stress) and the fractional excess density ρ_e/ρ_0 (the strain). By substituting equation (3) in (2), one obtains the formula for the sound velocity: the velocity is equal to the square root of the ratio, bulk modulus divided by the average density of the undisturbed medium, or,

$$V = \sqrt{B/\rho_0} = \sqrt{1/\rho_0 K} \qquad (4)$$

in which K, called the compressibility, is the reciprocal ($K = 1/B$) of the bulk modulus. For example, the bulk modulus of water has been measured to be (around room temperature) 2.1×10^{10} dynes per square centimetre. Because the density of water under the same condition is about one gram per cubic centimetre, one has a numerical value for the velocity of sound in water under these conditions (from equation [4]),

$$V = \sqrt{\frac{2.1 \times 10^{10}\ \text{dynes/cm}^2}{1\ \text{gram/cm}^3}} = \sqrt{2.1} \times 10^5$$

centimetres per second, which comes out to be 1,450 metres (4,756 feet) per second, a figure in good agreement with the experimentally measured velocity. As a matter of fact, confidence in equation (4) is such that it is commonly used to determine the bulk modulus of liquids from the observed velocity of sound, in place of the static compression method that demands great precision, because liquids are hard to squeeze.

It can be noted from equation (4) that the velocity is expressed in the form of the square root of the ratio of a factor (B) that depends on the elasticity to a specific inertia factor (ρ_0). This type of dependence holds for elastic waves in general (*i.e.,* in solids as well as in gases and liquids). It means that for a given specific inertia factor (such as density), the more elastic a medium is the greater is the sound velocity. It may be that the two factors are interdependent, and then the sound velocity formula can take on a different form. This is true, for example, for gases, as is made clear in the next paragraph.

To use equation (2) for gases requires the so-called equation of state, which connects pressure, density, and temperature. For an ideal gas (see MATTER) the equation of state says that the pressure (p) divided by the density (ρ) is proportional to the absolute temperature, or

$$p/\rho = RT, \qquad (5)$$

in which T is the absolute temperature (Celsius temperature plus 273°) and the constant of proportionality, R, is the ideal gas constant per gram. It can be shown that if the temperature remains constant the ratio of the excess pressure to the excess density is also equal to RT—that is, $p_e/\rho_e = RT$—and therefore that the velocity of sound is equal to the square root of the gas constant times the absolute temperature and also equal to the square root of the ratio of the pressure to the density; *i.e.,* $V = \sqrt{p/\rho}$. As noted above, this is the result obtained by Newton. It does not agree with experiment. If one follows Laplace's assumption that the pressure and density changes that take place in a gas in sound transmission are adiabatic instead of isothermal—that is, the temperature in a gas through which a sound wave passes does not remain constant—one must use the equation in which the pressure divided by the density raised to the γ power is a constant, or $p/\rho\gamma$ = constant, in which γ (gamma) is equal to c_p/c_v, the ratio of the specific heat of the gas at constant pressure (c_p) to the specific heat at constant volume (c_v). Then, the velocity of sound is equal to the square root of gamma times the ratio of the actual fluid pressure to its density, or

Newton's law of sound propagation

$$V = \sqrt{\gamma p/\rho}. \qquad (6)$$

This formula agrees so well with experiment at sonic and low ultrasonic frequencies that it is often employed to determine the value of gamma from measurement of sound velocity. It might at first be supposed that if the sound-induced changes in pressure took place adiabatically at audible frequencies, they would continue to behave adiabatically as the frequency increases. That conclusion would follow from the reasoning employed by Laplace to deduce equation (6). Laplace's reasoning was faulty, however, and at very high ultrasonic frequencies (perhaps of the order of 10^{12} hertz) the velocity of sound follows the Newtonian formula.

For non-ideal gases and vapours—*i.e.,* those that do not obey the equation of state expressed by equation (5)—the simple formula (6) must be corrected, by methods given in advanced treatises.

In any case, for an ideal gas at what may be called ordinary frequencies, equation (6) shows that the velocity of sound does not depend on pressure or density but only on the absolute temperature. Thus equation (6), with the use of the equation of state, states that velocity is equal to the square root of the product of gamma, the gas constant, and the absolute temperature, or $V = \sqrt{\gamma RT}$. This equation can be rewritten in more convenient form. If the velocity of sound in dry air at 0° C (32° F) is denoted by V_0 that at temperature $t°$ C becomes equal to the velocity at 0° C times a factor, the square root of the quantity 1 plus the Celsius temperature divided by 273, or

$$V_t = V_0\sqrt{1 + t/273}.$$

For dry air under standard conditions (*i.e.,* 0° C), $V_0 = 331.3$ metres (1,086.7 feet) per second. For room temperature (20° C or 68° F) the equation then yields a velocity of 344 metres (1,128 feet) per second or 0.214 mile per second.

sence of persion

As the preceding equations indicate, under what may be termed ordinary conditions, the velocity of sound even in an ideal gas and indeed in all actual gases is independent of the frequency—*i.e.,* there is no so-called dispersion (see MECHANICS). If, however, the ratio of frequency to pressure (*i.e.,* f/p) increases sufficiently, the velocity of sound, even in a relatively permanent gas like hydrogen, also shows an increase. In the case of hydrogen, for example, as the ratio of frequency to pressure increases from one megahertz per atmosphere to 30 megahertz per atmosphere, the velocity of sound waves increases by about 8 percent from its standard low frequency value of 1,284 metres (4,212 feet) per second.

In the case of liquids to which equation (4) applies, strictly speaking, the adiabatic bulk modulus or compressibility must be used. For liquids, however, these values as a rule differ little from their isothermal counterparts. Because the dependence of the compressibility of a liquid on the temperature is a complicated affair, it is not possible to derive simple equations to express the dependence of the sound velocity on temperature. Empirical formulas found in advanced treatises show that the sound velocity in water, for example, increases as the temperature rises to 74° C (165° F) and thereafter decreases. For most other liquids, decrease of velocity with increase in temperature is the rule.

Compressional waves in solids are usually considered to be sound waves. Their velocities follow the pattern described in the beginning of this subsection, viz., they are expressible as the square root of the ratio of an elasticity factor to a specific inertia factor. The velocity (V) of a compressional wave in a long solid rod, for example, is expressible in the following way: the velocity (V) is equal to the square root of Young's modulus (the "stretch" modulus of elasticity, Y) divided by the equilibrium density (ρ_0) of the rod material; *i.e.,* $V = \sqrt{Y/\rho_0}$. For most hard solids the velocity here has a range of values from about 1,000 metres (3,300 feet) per second to 6,000 metres (19,800 feet) per second, depending on the elasticity and density. Thus the compressional wave velocity in a steel rod is many times that in a lead rod, as the density of steel is less than that for lead, and the Young's modulus for steel is about 13 times that of lead.

For a three-dimensional solid—*i.e.,* one not in the form of a long thin rod—the compressional wave velocity is similar to equation (4), except that the bulk modulus is increased by a quantity connected with the shear modulus (see MECHANICS). It is $V = \sqrt{(B + 4\mu/3)/\rho_0}$, in which B is as usual the bulk modulus, ρ_0 is the equilibrium density, and μ is the shear modulus.

Wave fronts and rays. It has already been pointed out that an audible sound wave tends to spread out in all directions from the source. It is helpful in describing this situation to introduce the concept of wave front: an imaginary surface of such a kind that at any instant the disturbance characterizing the sound wave is the same at every point of it.

Spherical wave front

An example is provided by a bird radiating a sound wave into the surrounding air. Since the sound wave propagates in all directions at a constant velocity, at any instant the state of compression of the air in this wave will be the same (to a good approximation) on the surface of a sphere with its centre at the bird. This sphere is a spherical wave front. The whole space surrounding the quasi-point source is filled with such wave fronts, each characterized by its appropriate phase of the disturbances. The propagation of the wave can be thought of as equivalent to the spreading of the sphere throughout space as its radius grows larger at a rate equal to the speed of sound in the medium. The expanding ripples on the surface of water produced by a dropped stone provide a clear analogy. A flat plate, on the other hand, vibrating in a direction perpendicular to its surface alternately compresses and rarefies the air and so produces a wave front that, if confined in a tube of uniform cross section, is approximately a plane wave front. Plane and spherical wave fronts are the two most important varieties in the study of sound-wave propagation.

Though wave fronts are an essential element in the description of sound-wave propagation through three-dimensional space, it is often convenient to describe the propagation in terms of rays—*i.e.,* lines perpendicular to a wave front. The ray passing through a particular point at a given instant indicates the direction in which the wave transmission is taking place at that time and place. Sound rays can often be useful, though not so much so as light rays, as it is harder to produce a genuine beam of sound concentrated in one direction because the wavelength of sound is in general so much larger than that of light. At ultrasonic frequencies, however, it is possible to approximate sound beams, and then ray acoustics becomes of value.

Reflection and refraction. The existence of an echo when a sound is emitted in the neighbourhood of a large, hard surface is a manifestation of the fact that sound waves can be reflected when they strike a surface separating media of different properties. Moreover, sound-wave fronts are bent (refracted) when sound crosses such a boundary; *e.g.,* in going from air to water or vice versa.

Huygens' principle of wave propagation

The laws governing the reflection and refraction of sound waves are best understood in terms of a fundamental principle stated by a Dutch scientist, Christiaan Huygens. According to this principle each point on a wave front (F_1 in Figure 1, for example) may be assumed to be the source of a hemispherical wavelet that moves outward from F_1 in the direction of propagation. The new wave front after a short time is the mathematical envelope of all these wavelets—*i.e.,* the surface tangent to them all (see F_2 in Figure 1). A simple but important illustration of Huygens' principle is the establishment of the laws of reflection and refraction of a plane wave at a plane interface separating two fluid media with different densities and different sound velocities. In Figure 2, SS' represents the trace (intersection) in the plane of the page of the plane interface between a fluid medium (I) in which the velocity of sound is V_1 and the mean density is ρ_1, and a medium II in which the corresponding quantities are V_2 and ρ_2, respectively. AB denotes the trace of a plane wave front (perpendicular to the plane of the page) approaching the interface SS' from medium I. If the wave front AB is incident on SS' as CD with angle of incidence i, as in the figure, then AC and BD are incident rays that make an angle i with the normal CN to the interface. The reflected and refracted wave fronts can be constructed by means

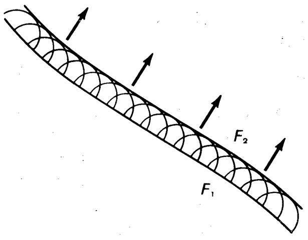


Figure 1: Huygens' construction showing points on wave front F_1 generating wavelets to produce new wave front F_2 (see text).

of Huygens' principle. At the instant the disturbance at one end of the incident wave front has reached C on the interface, the disturbance at the other end is still at D in medium I. In a time equal to the distance DE divided by the velocity (DE/V_1) that it takes the disturbance at D to reach E on the interface, the disturbance will travel out in all directions from C. In particular, in medium I it will traverse a distance equal to DE.

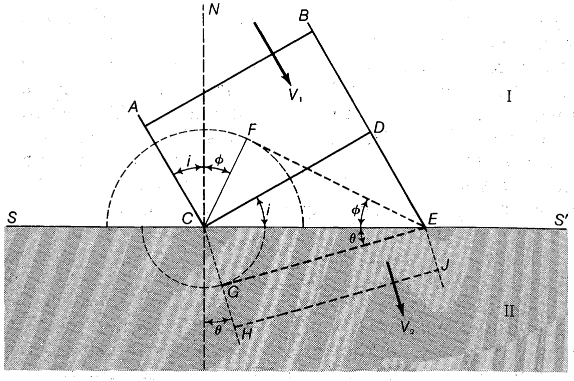


Figure 2: Huygens' construction showing wave fronts *AB* and *CD* that are reflected as *FE* at the boundary between media I and II and refracted as *GE* and *HJ* to give the law of reflection and Snell's law of refraction (see text).

If, with C as centre, a semicircle is drawn in medium I with radius DE and a straight line from E is drawn tangent to this, the result is EF, which makes the angle $\varphi = i$ with SS'. That EF is the trace of the reflected wave front can be confirmed by making a Huygens' construction for other points on the incident wave front. It is seen that the reflected wave front makes the same angle with the interface as the incident wave front. This, combined with the fact that both wave fronts are perpendicular to the same plane (plane of the page), constitutes the law of reflection. Equally well expressed in terms of the incident and reflected rays, this law is the same law that holds for light, namely, that the angle of reflection is equal to the angle of incidence.

The same kind of Huygens' construction shows that, after transmission across the interface, the incident wave front CD is changed to the refracted wave front GE, making the angle θ with SS', such that the ratio of the sine of the angle of incidence to the sine of the angle of refraction is equal to the ratio of the velocity in the first medium to that in the second—*i.e.,*

$$\sin i/\sin \theta = V_1/V_2 = n_{12}. \qquad (7)$$

This is Snell's law for the refraction of plane waves at a plane interface, with n_{12} called the index of refraction of medium II with respect to medium I. A similar law, in somewhat modified form, holds for the transmission of wave fronts through a medium with continuously varying properties. The law as expressed in equation (7) accounts well for the passage of sound waves from air to water, or vice versa, when the media are at rest.

In Snell's law, equation (7), if the velocity in medium I is greater than that in medium II ($V_1 > V_2$), the index of refraction is greater than one ($n_{12} > 1$) and refraction takes place for every angle of incidence. If the velocity in medium I is less than in medium II ($V_1 < V_2$), however, the index of refraction will be less than one ($n_{12} < 1$), and when the incident ray makes an angle i such that $\sin i = n_{12}$, Snell's law gives $\sin \theta = 1$; *e.g.*, the refracted ray just grazes the interface in the second medium—that is, the refracted wave front moves so as to remain perpendicular to the interface. For sin i greater than n_{12}, as may well be possible since the largest value of sin i is equal to one, sin θ becomes greater than one, which is mathematically impossible. The physical result, under this condition, is that no refraction takes place; instead, total reflection ensues. The angle at which this occurs (arc sin n_{12}) is called the critical angle of incidence, and for angles greater than this, all the sound is reflected, analogously to the similar phenomenon in light.

Total reflection

When the two media—*i.e.,* air and water—are moving relative to each other through the action of wind and water currents, the refraction of sound becomes more complicated and so the result will be given only for a relatively simple case in which the velocity of flow of medium I is U_1 parallel to the interface and that of medium II is U_2, also parallel to the interface. Then the law of refraction becomes:

$$V_1/\sin i - V_2/\sin \theta = U_2 - U_1. \qquad (8)$$

If the two velocities are equal ($U_2 = U_1$)—*i.e.,* the two media are moving together with no relative velocity—equation (8) reduces to equation (7). Equation (8) also applies, as more elaborate analysis shows, when the sound velocity and flow velocity vary continuously in a stratified or layered medium in which the flow is parallel to the layers. Then all quantities in (8) vary from place to place in a direction perpendicular to the stratification. Examination discloses that even if velocities V_1 and V_2 are constant, the effect of the motion of the medium is to bend the wave front and ray in the direction in which the motion takes place. Thus, if the wind velocity increases upward from the ground, sound-wave fronts are lifted to windward and depressed to leeward. This tends to decrease the range to windward and increase it to leeward. When the velocities V_1 and V_2 also vary with distance above the ground, as is commonly the case in the atmosphere, the wave-front propagation becomes even more complicated. Nevertheless, many practical cases have been worked out successfully. These considerations are of great importance in sonic gun-ranging.

Energy propagation: intensity of sound. As has already been pointed out, the propagation of sound may be considered as a form of transmission of energy through the material medium. When, for example, a fluid is disturbed at any point, work is done, and energy is expended. The reappearance of the disturbance at a distant point by wave propagation corresponds to the transfer of the original energy across the intervening distance. The average rate of energy transfer per unit time per unit area of the wave front is, by definition, the intensity of the wave. This may also be expressed as the average flow of power per unit area, the practical unit of which is the watt per square centimetre (watt/cm^2).

The intensity in a plane sound wave may be evaluated by finding the time average of the product of the excess pressure (p_e) and the flow velocity ($\partial\xi/\partial t$) in the medium. Inasmuch as pressure is force per unit area, the product just cited is valid as far as the consistency of dimensions is concerned. (The units of pressure times velocity are dyne centimetre^{-2} · centimetre second^{-1}, whereas the units of power per unit area are erg second^{-1} centimetre^{-2}. These are equivalent, as an erg is a dyne centimetre.) It is necessary to take the average of the product in question because both the excess pressure and the flow velocity vary in time and space as the wave progresses. Hence the flow of energy actually fluctuates, and to get a constant effective measure of the flow of energy per unit area, an average must be taken. The appropriate mathematical analysis shows that the intensity of a plane sound wave is given by:

$$I = p^2_{e,\,max}/2\rho_0 V. \qquad (9)$$

in which $p_{e,\,max}$ is the maximum excess pressure in the wave, ρ_0 is the average equilibrium density of the medium, and V is the sound velocity. This expression is independent of frequency. The use of the decibel notation in the representation of intensity has already been described in the section on the general nature of sound waves, and examples were given. It is sufficient to recall here that for ordinary conversational speech, at a distance of about one metre from the speaker's mouth the maximum excess pressure is one dyne per square centimetre (dyne/cm^2). For dry air at 20° C (68° F) and normal atmospheric pressure, the equilibrium density (ρ_0) is 0.001205 gram per cubic centimetre, and the velocity (V) is 344 metres (1,128 feet) per second. Substitution of these values into equation (9) and subsequent division by 10^7 to convert to watts per

Intensity of speech

square centimetre yields an intensity (I) of about 10^{-9} watt per square centimetre, or 70 decibels above the minimum audible intensity of 10^{-16} watt per square centimetre.

From equation (9) it follows that the intensity for given excess pressure depends on the medium and in particular on the product equilibrium density times sound velocity ($\rho_0 V$). The latter quantity is known as the specific acoustic resistance of the medium for a plane wave. For water, for which the specific acoustic resistance is about 3,800 times that for air under standard conditions, it takes an excess pressure some 60 times that for air to produce the same intensity. It will be seen below in the section on sound sources that for given frequency and excess pressure a solid sound source is a more efficient radiator in water than in air.

The intensity of a spherical wave diverging from an effective point source can still be written in the form given by equation (9), provided it is remembered that for this type of wave the maximum excess pressure is no longer constant but varies inversely as the square of the distance from the source. Hence the intensity as given in equation (9) still applies if the maximum excess pressure is the excess pressure amplitude at the place where the intensity is to be evaluated.

Flow of sound energy across a boundary. When a sound wave strikes the interface separating two media of different physical properties (*e.g.,* air and water), as was already seen, some of the wave disturbance is reflected and some transmitted (and in general refracted) through the second medium. The transmission of the sound energy is thus affected by the presence of the boundary.

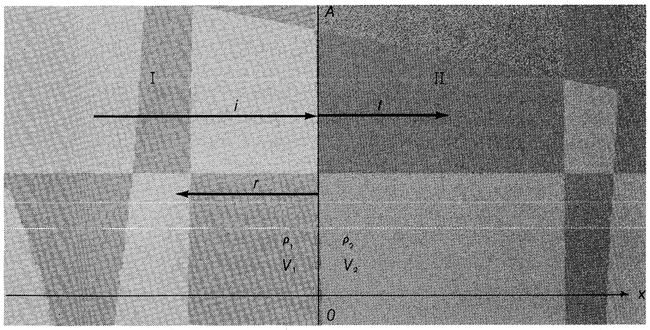


Figure 3: Direction of incident, transmitted, and reflected waves at boundary *AO* between media I and II (see text).

Figure 3 represents a plane boundary *AO,* separating two media I and II, that is perpendicular to the direction of plane wave propagation along the x axis. In medium I the mean equilibrium density is represented by ρ_1 and the sound velocity by V_1. The corresponding quantities in II are ρ_2 and V_2, respectively. The arrow i denotes a plane harmonic sound wave travelling from left to right in medium I in the x direction and striking the boundary at x equal to zero. Similarly, r denotes the wave in I reflected from the interface, whereas t denotes the wave transmitted across the boundary into medium II.

The amplitudes of the excess pressure in the incident, reflected, and transmitted waves may be represented by p_i, p_r, and p_t, respectively. These quantities are not independent inasmuch as the sum of the incident and reflected excess pressure amplitudes must equal the transmitted excess pressure amplitude. There is a similar condition connecting the flow-velocity amplitudes in the incident, reflected, and transmitted waves. The result of applying these two conditions can be expressed by the following two equations:

$$\frac{p_r}{p_i}=\frac{1-\rho_2V_2/\rho_1V_1}{1+\rho_2V_2/\rho_1V_1}, \qquad (10)$$

$$\frac{p_t}{p_i}=\frac{2}{1+\rho_1V_1/\rho_2V_2}. \qquad (11)$$

In the case in which the density times velocity (ρV value) is the same in both media ($\rho_2V_2=\rho_1V_1$), equations (10) and (11) give p_r equal to zero and p_t equal to p_i; in words, all the sound energy is transmitted across the boundary, and none is reflected. This would be realized if the two media were precisely the same: the interface would then of course disappear. It may well be possible, however, to find two different media for which the ρV values are the same and hence secure complete transmission without reflection. Thus there is a certain kind of soft rubber whose ρV value is about the same as that of water. A water tank lined with this material would appear to have no reflecting walls, and this is helpful in the calibration of certain acoustical instruments.

Transmission of sound without reflection

The greater the difference between the ρV values of the two media, the greater is the percentage of reflection at the interface. The actual power-transmission ratio (P_r), which is the ratio of the intensity of the transmitted sound wave to that of the incident sound wave, also depends only on the factors mentioned above. Calculation shows that the power-transmission ratio is given by equation (12):

$$P_r=\frac{4\rho_2V_2/\rho_1V_1}{(1+\rho_2V_2/\rho_1V_1)^2}. \qquad (12)$$

Again from equation (12) it can be seen that when the ρV value is the same in both media, the power-transmission ratio is equal to one, meaning that 100 percent of all the incident sound energy or power is transmitted. The specific acoustic resistance, ρV, is clearly an important criterion for the transmission of plane sound radiation across boundaries.

Formula (12) together with the corresponding formulas for the case of oblique incidence have had important practical application, such as to acoustic transmission in the atmosphere between layers of air at different temperature and humidity and to transmission in seawater with temperature gradients.

Acoustic impedance. The significance of the product ρV in sound transmission prompts further attention to it. Referring to this quantity as the specific acoustic resistance of a plane harmonic wave inevitably suggests the analogy of electrical impedance in alternating electric current, defined as the ratio of the alternating electromotive force to the alternating current (see ELECTRICITY AND MAGNETISM). It has therefore been thought helpful to define the acoustic impedance as the ratio of the excess pressure to the rate of volume flow. The volume flow is obtained from the flow velocity ($\partial\xi/\partial t$) by multiplying by the area (S) of the wave front. Hence the impedance (Z) has a value equal to the excess pressure divided by the product of the area and the flow velocity, or $Z=p_e/(S\partial\xi/\partial t)$. In words, the acoustic analogy of electromotive force is taken to be the excess pressure, and the acoustic analogy of the electric current is the volume flow rate.

In the mathematical theory of alternating currents the impedance is a complex quantity (so called because it involves the square root of minus one, an imaginary number) of which the real part is the resistance of the circuit and the imaginary part (having a factor $\sqrt{-1}$) is the reactance. In the acoustic wave case the quantity Z for a plane wave is real and equal to the density times velocity divided by wave front area ($\rho V/S$). Hence it is taken to correspond to an electrical resistance. The acoustic resistance for unit area of wave fronts (not per unit area) is the product ρV and this is justifiably called the specific acoustic resistance. Analysis shows that the acoustic impedance for a spherical wave has both real and imaginary components, corresponding to resistance and reactance, respectively. The physical meaning of this is connected with the spreading of the wave in all directions. The reactance of a diverging spherical wave turns out to be positive and behaves like an electrical inductance (*i.e.,* an inertia factor), whereas the reactance of a spherical wave converging to a point is negative and thus behaves like an electrical capacitance (*i.e.,* the reciprocal of a stiffness).

Specific acoustic resistance

The impedance notation has the advantage of permitting acoustic transmission calculations to be replaced by analogous electrical computations. Many acoustical effects can be readily estimated in terms of known values of acoustic impedance at strategic points. For example, the effective-

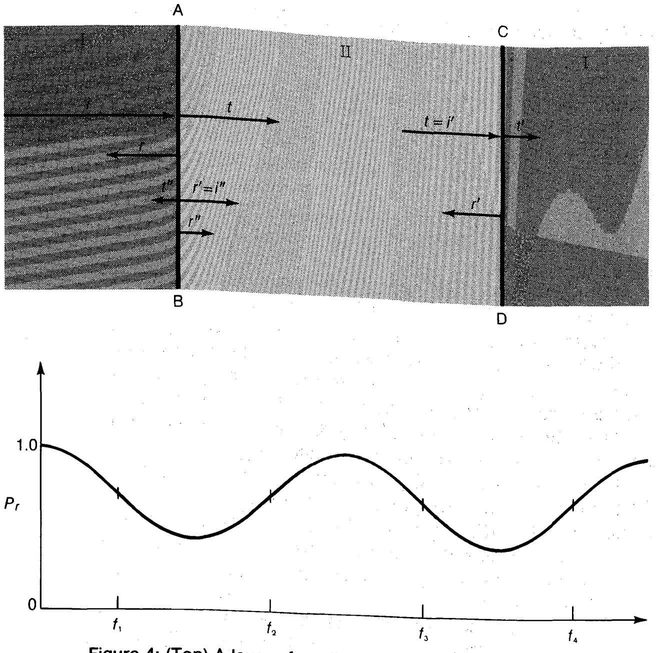


Figure 4: (Top) A layer of medium II situated between layers of medium I (see text). Waves (i, i', i'') incident on both faces AB and CD are reflected (r, r', r'') and transmitted (t, t', t''). (Bottom) Plot of power-transmission ratio (P_r) as a function of frequency (f) transmitted through medium I as above. ($P_r = 1$ represents total transmission.)

ness of a horn (*i.e.*, a tube of variable cross section) as both a receiver and a transmitter of sound waves is directly proportional (for a given volume current) to the value of the acoustic resistance at the horn's throat. The effective acoustic transmission through a tube in which a branch line is inserted at some point can be shown to depend in a relatively simple fashion on the acoustic impedance at the branch point. Problems in room acoustics involving the reflection of sound from the walls of a room can often be simplified by the use of the impedance notation.

Selective transmission: filtration. An important case of the transmission of acoustic energy arises when a finite layer of different acoustic material (medium II) is inserted in a medium (medium I) in which the propagation takes place. A simple example is illustrated in Figure 4 (top). To determine the influence of the layer of medium II on the transmission, a mathematical analysis can be made of the transmission of a plane harmonic sound wave as it passes from medium I through medium II and back into medium I. This yields a power transmission ratio (P_r)—*i.e.*, the ratio of the average power emerging from medium II to that entering it—which is found to depend on the frequency of the sound and the thickness of the layer of medium II, as well as on the ρV value for each layer. The plot of P_r as a function of frequency has the general form shown in Figure 4 (bottom). The transmission in this circumstance is said to be selective with respect to frequency and passes through a succession of maxima and minima as the frequency increases.

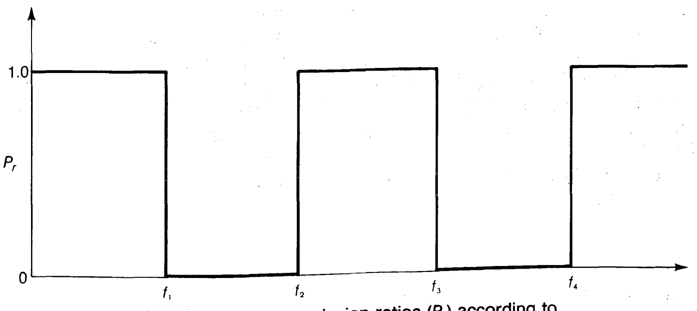


Figure 5: Power-transmission ratios (P_r) according to frequencies (f). Frequencies lying between 0 and f_1, f_2 and f_3, etc., called pass bands, are totally transmitted. Frequencies between f_1 and f_2, f_3 and f_4, etc., called attenuation bands, are totally absorbed.

This suggests the possibility of arranging a collection of media that will act as an acoustic filter—*i.e.*, transmit waves of certain frequencies well, but fail entirely to pass waves of other frequencies. Further theoretical study indicates that this result can be realized approximately by a structure consisting of a sequence of alternating layers of media I and II. As the number of layers increases the transmission ratio approaches more nearly the form shown in Figure 5. Here the power-transmission ratio is equal to one for frequencies lying between zero and f_1, equals zero for frequencies between f_1 and f_2, is unity again for frequencies between f_2 and f_3, etc. A frequency band such as $f_2 f_3$, in which the power-transmission ratio is one, is called a pass band, whereas a band such as $f_1 f_2$, in which the power-transmission ratio is equal to zero, is called an attenuation band. The structure of alternating layers thus acts as an acoustic filter with a series of pass and attenuation bands. Because the first pass band extends from zero to f_1, it is commonly called a low-pass filter.

Attenuation banc

The type of acoustic filter just described depends for its behaviour on the alternation of properties of two different media. This, however, is not necessary to achieve the same general result. Figure 6 (top) represents a succession of cylindrical tubes of cross-sectional areas S_1 and S_2 and lengths l_1 and l_2, respectively. Analysis and experimental tests show that such a structure acts as a low-pass filter for harmonic sound waves passing through it. The transmission pattern of pass and attenuation is indeed very similar to that shown in Figure 5. Here the ratio S_1/S_2 of the cross-sectional areas of the constricted and expanded tubes plays the same role as the ratio of the ρV values of the specific acoustic resistances in the iterated-media filter. The lengths l_1 and l_2 of the section also have a decided influence on the frequency range of the pass and attenuation bands.

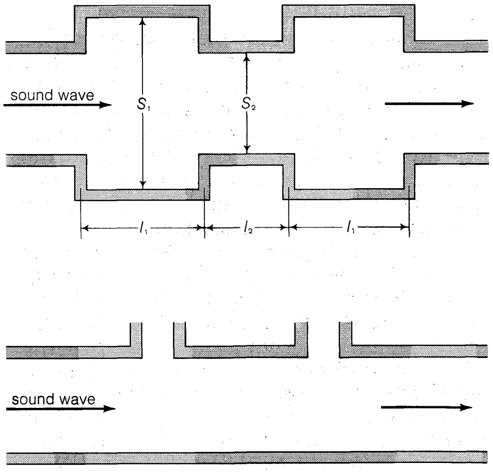


Figure 6: (Top) Low-pass acoustic filter, consisting of a succession of cylindrical tubes, cross-sectional areas S_1 and S_2 and lengths l_1 and l_2, for harmonic sound waves (see text). (Bottom) High-pass acoustic filter through which only high-frequency waves pass.

A tube with equally spaced side orifices, such as are shown schematically in Figure 6 (bottom), also acts as an acoustic filter, but in this case the lowest frequencies are attenuated, and hence it is called a high-pass filter. Many other types of sound filters have been constructed to suppress or silence certain frequency bands in blowers, mufflers, and other equipment producing noise by gas flow. Filters consisting of solid sections have also been constructed. Many musical instruments involve acoustical filtration in their behaviour.

Sound absorption and dispersion in fluids. In the previous discussion of sound filtration no account was taken of the dissipative absorption of sound radiation as it passes through any material medium and its transformation into heat. This leads to a loss in intensity with distance traversed that has nothing essentially to do with the selective transmission just treated. Also, as has already been mentioned, a decrease in intensity of a diverging spherical wave is caused by geometrical spreading. Superposed on this decrease is the absorption due to the interaction between the sound wave and the medium. The properties of the medium responsible for this absorption are shear viscosity, heat conduction and radiation, diffusion, and molecular energy exchange. Shear viscosity is a kind of internal fluid friction that manifests itself in the tendency of parallel moving layers of fluid to have their relative velocity reduced. The coefficient of shear viscosity (usually denoted by η) is a measure of this tendency (see MECHANICS).

To remain correct at all frequencies, the formula for the intensity of a plane harmonic acoustic wave in terms of the distance traversed must take account of the fact that, in propagation through a viscous medium, the sound velocity changes with frequency; *i.e.,* there is dispersion. Unlike the situation in light in which, except for anomalous dispersion, the velocity decreases as the frequency increases, the effect of the viscosity in a medium is to increase the sound velocity with increase in frequency. The effect is a small one, however, even in viscous media like the liquid, glycerin.

ct of
osity
ersion

The medium also produces sound absorption by conducting away heat produced in the compressional disturbance involved in the propagation of sound. For monatomic gases (*e.g.,* helium and argon) the contribution of the heat conduction to the absorption coefficient is about 70 percent of that due to viscosity. For polyatomic gases (*e.g.,* hydrogen sulfide and carbon dioxide) the influence of heat conduction is much less, and in liquids it is generally negligible. The effect of diffusion is also negligible save in mixtures of gases with considerably different molecular weights; *e.g.,* helium and krypton.

Experimentally determined values of the absorption of sound in fluids, especially polyatomic gases, are in general far in excess of those calculated above on the basis of viscosity and heat conduction. The explanation for this disparity has been sought successfully in a closer look at the energy exchanges among the molecules composing the fluid. After all, the propagation of sound is really a molecular phenomenon, which is somehow masked in the usual macroscopic theory of disturbances in elastic media. When a fluid is compressed during the passage of a sound wave, the average translational velocity of the molecules is increased, and it is the transfer of the associated momentum and kinetic energy by collision to the nearby molecules that provides for the transmission of the wave. But this is not all that happens. In a gas made up of polyatomic molecules, the molecules have internal energy of rotation and vibration as well as translational energy of their centres of mass. In the collision of the molecules, this internal energy is continually being exchanged with the external translational energy. When the translational energy is increased by the passage of a sound wave, some of the increase goes into the internal energy states. There is, however, a lag in its return to the translational form. Hence changes in the translational energy fail to keep pace with the propagation of the wave, and some of the original energy of the disturbance is rendered unavailable for transmission during each frequency cycle. It is therefore lost in dissipation and the result is absorption. This type of mechanism, called relaxation absorption, has proved adequate to account for the experimentally determined values. It has led to the establishment of a 20th-century branch of the theory of sound called molecular acoustics, by which sound measurements are able to give insight into the interaction of molecules in fluids.

lecular
ustics

Diffraction and scattering. When a sound wave strikes an obstacle of finite size, some of the sound energy is reflected, some is transmitted into the obstacle, and some gets around the obstacle and is said to be diffracted. It is customary to say that the original beam of sound is scattered by the obstacle. The scattering pattern depends on the relation between the wavelength of the primary incident wave and the dimensions of the obstacle. In the case, for example, of a rigid sphere of radius a as the obstacle, if the wavelength (λ) of the incident radiation is small compared with its circumference ($2\pi a$) the scattered wave on the side of the sphere opposite to the oncoming radiation combines with the latter to produce zero intensity by destructive interference (so called because, being out of phase with one another by 180°, their displacements are annulled). Hence the spherical obstacle forms a shadow that is sharper the smaller the ratio of wavelength to circumference ($\lambda/2\pi a$). This is analogous to an optical shadow in the case of light and justifies the use of sound rays (geometrical acoustics) in place of the propagation of wave fronts. For audible waves (λ in air greater than about two centimetres or 0.8 inch), however, this type of shadow-forming scattering is important only for relatively large obstacles.

At the opposite extreme, if the wavelength is much greater than the circumference ($\lambda >> 2\pi a$), the primary wave is scattered in all directions and there is no distinct shadow. The intensity of the scattered wave is directly proportional to the square of the volume of the scattering particle (in this case a sphere) and inversely proportional to the fourth power of the wavelength. This is the Rayleigh scattering law in acoustics; its optical analogue helps account for the blue colour of the sky (see LIGHT). In the scattered radiation the higher frequencies will produce the higher intensities. Thus it is observed that sound waves scattered by a grove of trees appear to be raised in pitch. A similar phenomenon occurs when underwater sound waves are scattered by air bubbles.

An interesting case of scattering is that in which light is scattered by sound. Optical diffraction by means of a diffraction grating is well-known. A sound wave can serve as a kind of diffraction grating in a material medium (a liquid as well as a gas) and hence diffract light. Study of the light-diffraction spectrum can actually provide a measure of the intensity of the sound.

The fluctuations in density in any material medium (at any temperature above absolute zero) due to molecular motion and ionic vibrations may be considered a form of sound radiation at extremely high frequencies—*i.e.,* about 10^{12} hertz and higher. This is sometimes called the pretersonic, or ultrahigh, ultrasonic range. Such sound waves also scatter light and the result is the so-called Brillouin scattering, which in the 1960s and 1970s showed great promise in the study of material structure, particularly of liquid structure. It has proved advantageous in the study of crystal lattices (a crystal is said to have a lattice structure because its atoms are arranged in an orderly array) to associate a sound "particle" with the ultrahigh-frequency waves travelling back and forth in the solid as a result of the lattice vibration. This hypothetical particle is called the phonon. It plays the same role with respect to ultrahigh-frequency acoustic radiation that the photon, or light particle, plays with respect to electromagnetic radiation.

The
phonon

Another example of the interaction of light and sound is the use of laser light (a narrow beam of coherent light) to stimulate the production of high-frequency sound waves in a crystal. A similar technique can also serve to amplify existing ultrasonic waves.

Waves in bounded media: standing waves in a tube. When sound waves travel in a bounded medium with reflecting surfaces, the superposition of waves travelling in many directions produces important and interesting effects. The simplest case is that of a cylindrical tube of length l and cross-sectional area S, closed rigidly at both ends. A disturbance produced anywhere in the air in the tube will be reflected from both ends and produce in general a series of waves travelling in both directions along the tube. From the geometry of the situation and the constant value of sound velocity, these waves must be periodic with frequencies determined by the boundary conditions at the ends of the tube; *i.e.,* zero resultant air displacement. The allowed frequencies turn out to be a multiple of the sound velocity in the air in the tube divided by twice the tube length, or

$$f_n = nV/2l, \qquad (13)$$

in which V is the sound velocity and n is any integer. These frequencies f_n in which $n = 1, 2, 3, \ldots$ are called the characteristic frequencies or normal modes of vibration of the air column. The fundamental frequency, or first harmonic, is described by an integer value of one and is $f_1 = V/2l$ with the associated wavelength $\lambda_1 = 2l$. The frequency f_2 is called the second harmonic and so on for $n = 3, 4, 5, \ldots$. The harmonics beyond the fundamental are also often referred to as overtones.

End correction for open pipes

Approximately the same set of normal modes holds for a cylindrical tube open at both ends, though the boundary conditions are different. It is assumed that the excess pressure p_e is zero at an open end, since it is open to the surrounding air. The justification of this assumption cannot be easily explained, but it can be confirmed experimentally by placing a probe microphone at the open end. Actually, it turns out that the point of zero excess pressure is never precisely at the open end but somewhat beyond it, producing an effect called an end correction to the normal mode frequencies. Application of the ideal boundary conditions just mentioned to a tube closed at one end and open at the other (the ideal case of a so-called closed organ pipe) leads to an equation for the normal mode frequencies that is similar to equation (13), but the whole integer is replaced by a half integer value:

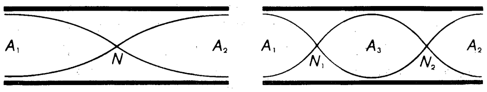
$$f_n = (n - \tfrac{1}{2})\ V/2l, \qquad (14)$$

in which n equal to one corresponds to the fundamental, and n equal to two to the second harmonic, etc. Because of the open-end effect, however, the length $(2l)$ in equation (14) must be modified by adding a quantity that depends on the tube diameter. Thus, the actual frequencies are

$$f'_n = (n - \tfrac{1}{2})\ V/2\,(l + S/c),$$

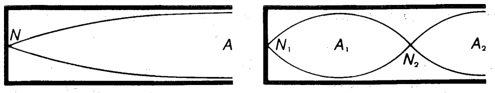
in which S is the area of cross section of the tube and c is a quantity with the dimensions of length of about the magnitude of the diameter of the tube. It is called the acoustic conductivity of the open end of the tube. The acoustic conductivity is difficult to calculate theoretically and is often evaluated by the actual measurement of the frequency f'_n.

In an ideal sound-wave disturbance corresponding to a given normal mode in a tube open at both ends, there are positions in the tube, namely, those at distances (x) from one open end that satisfy the equation, $\cos n\pi x/l = 0$, at which the displacement is zero at all times. Such a situation cannot exist in a progressive wave, as shown by equation (1). Hence the resultant wave disturbance corresponding to a normal mode is appropriately termed a standing or stationary wave. The positions of permanent zero displacement are known as nodes, whereas the intermediate positions for which $\cos n\pi x/l$ is equal to one instead of zero are known as antinodes or loops and correspond to maximum displacement. At a node the excess pressure is a maximum, whereas at an antinode it is zero. The distance between successive nodes is equal to a half wavelength of the particular mode. As an example, the fundamental in an open-ended tube is shown schematically in Figure 7 (left) by means of portions of sine waves (the maxima of a wave indicate the positions along the tube where the displacement is greatest). Displacement antinodes are indicated by A_1 and A_2, whereas N (at centre) is a node. In Figure 7 (right) the second harmonic is indicated with antinodes at A_1, A_2, and A_3, and nodes at N_1 and N_2. The latter are, respectively, 1/4 and 3/4 of the length of the tube from the open end at the left. In Figure 8 the corresponding situation is shown for a tube closed at one end and open at the other. In Figure 8 (left) the closed end is node N and the open end is an antinode A. The length of the tube is $\lambda_1/4$. In Figure 8 (right) the closed end is still a node N_1 and the open end an antinode A_2, but there now is another antinode A_1 and another node N_2. For the nodes in the tube closed at one end (at $x = 0$), the analytical condition that exists is $\sin (n - 1/2)\ (\pi x/l) = 0$, in which n equal to one corresponds to the fundamental, etc.



fundamental second harmonic

Figure 7: Nodes (*N*) and antinodes (*A*) in an open-ended tube (see text).



fundamental second harmonic

Figure 8: Nodes (*N*) and antinodes (*A*) in a tube closed at one end (see text).

Acoustic propertie of a roon

The preceding analysis can be applied to any confined space as, for example, a room with or without open windows. The reflections of a sound disturbance from the boundaries of the room produce a three-dimensional standing wave pattern that can be analyzed with a set of normal nodes, depending on the shape and dimensions of the room. Some of these nodes may be much more prominent and correspond to more energy flow than others and thus have a good deal to do with the acoustical properties of the room. Here absorption at the walls and in the air of the room plays an important role in diminishing the influence of these normal nodes.

Radiation pressure. For a plane harmonic progressive wave the excess pressure produced by the disturbance as it moves through a fluid is a sinusoidally varying function of time and space and hence vanishes if averaged over the time at any particular place. Associated with a progressive sound wave in any physically realizable fluid medium, however, there is also an excess pressure that does not become zero on the average, thus making a net addition to the static pressure prevailing in the fluid. This is known as the radiation pressure. It arises because the simple linear expressions for displacement and excess pressure, like that in equation (1), hold only for sound of relatively small intensity. For sounds of greater intensity, sine-squared terms and the like enter into the expression for the excess pressure, and it is these terms that, when averaged over the time, give a nonvanishing result, corresponding to the steady radiation pressure. The subject is admittedly a difficult one, and it turns out that there are at least two different kinds of radiation pressure, depending on definition, theoretical calculation, and experimental measurement, but the most practical from an experimental point of view is expressed simply as the ratio of the intensity (I) of the sound radiation to its velocity (V) and has the value $P_{rad} = I/V$. To get an idea of the order of magnitude of the radiation pressure, it is interesting to consider sound in air of intensity 10^{-2} watt per square centimetre, which is about 140 decibels above the minimum audible intensity. This intensity is at about the threshold of pain in the human ear and is equivalent to 10^5 ergs per square centimetre-second. In this case the radiation pressure comes out about three dynes per square centimetre.

Nonlinea acoustics

Radiation pressure is an illustration of what has come to be called nonlinear acoustics, resulting in effects from such large sound intensities that the ordinary linear wave equation no longer applies. Intensities of interest here are about 150 decibels above minimum audible or higher (greater than absolute intensity 1/10 watt per square centimetre). Such sound waves are now referred to as macrosonic. They have become of increasing importance in connection with practical applications of acoustics in industrial processing, as in metallurgy, in underwater sound for communication and detection, and in the biomedical field for the study of individual living cells.

Shock waves. In a macrosonic or high-intensity sound wave, (*e.g.*, that due to an explosion) a disturbance that starts out with a symmetrical profile—*i.e.*, a sinusoidal curve given by equation (1)—will soon develop a steepening of the wave front, which will then take the form of a sawtooth curve if the excess pressure at a given instant is plotted as a function of distance in the direction of the wave propagation. At successive points in this wave there are, ideally, discontinuities in pressure, density, and flow velocity. Empirically, these ideal discontinuities become large changes in these quantities over a very small space interval—*e.g.*, about 10^{-4} centimetre (4×10^{-5} inch). This transition interval travels through the medium faster than the normal acoustic velocity. Called a shock wave, it can result not only from an explosion but is also emitted from an object like a missile or jet airplane moving faster than the velocity of sound. In the case of an airplane breaking the sound barrier, as the phrase goes, it results in the sonic bang or boom on the ground. Such sonic disturbances have produced a serious problem in connection with the flying of supersonic airplanes.

Sound production and perception

SOURCES OF SOUND

Types of sources. Any local change in the density of an elastic medium can serve as a source of sound. This accounts for the great variety of acoustic sources because density changes may be produced in a great many ways, including mechanical, thermal, electrical, magnetic, and chemical actions. The most common sound waves are produced by the mechanical vibrations of solids, liquids, and gases. Solid vibrators include strings and rods, membranes and plates, shells (*e.g.,* bells), as well as three-dimensional extended objects like the Earth itself. Liquid sources are not as common, but the turbulent flow of water or air provides an example. Gaseous sources include organ pipes, whistles, singing flames, and explosions, as well as turbulent airflow.

Methods of producing local density changes in elastic media vary greatly. One of the simplest is the impact of one solid on another, such as the impact of a pile driver on a pile, a piano hammer on a piano string, a drumstick on a drum, a clapper on a bell, or a shoe on a floor. Still another common method is exhibited by the friction associated with the rubbing of two solids together, as in the bowing of a violin string and in the motion of air and other gases past the surfaces of solids, as in the sound of the wind and in aerodynamic noise.

More sophisticated sound sources became technically feasible when mechanical vibrations were made possible by means of magnetic and electrical effects. The attraction of a piece of iron or some magnetic alloy to the pole of an electromagnet is of course the basis for the electric bell and the telephone receiver. The motion of a wire carrying an alternating current in a magnetic field is the basis for the so-called electrodynamic sound source used in radio and television loudspeakers. They are appropriately called electroacoustic devices because they transform electrical oscillations into mechanical vibrations, usually of rods, membranes, or diaphragms.

ctro-
ustic
ices

Still more sophisticated in many respects are the magnetostrictive and piezoelectric sound sources. The former utilize the change in dimensions produced in a magnetic solid, such as nickel, by a changing magnetic field. The latter exploit the fact that many crystals, such as quartz, change their size and shape when placed in a changing electric field. These sources have received wide use in the production of underwater sound radiation and in medical diagnosis and therapy, particularly at ultrasonic frequencies. It is customary to call such sound sources transducers because they involve the transformation of one type of vibratory energy into another. These devices are usually suitable not only for the production of sound but also for its reception as well.

Explosions caused by the rapid combustion of organic materials produce another source of sound usually in the form of noise, as exemplified by the internal-combustion engine.

Mechanical vibration and radiation efficiency. The properties of a sound source depend essentially on the nature of its elastic vibrations. The analytical details will be found in the article MECHANICS. Here the discussions will be confined to general characteristics having to do with the ability of an elastic vibrator to radiate sound energy into the surrounding medium.

When a mechanical system vibrates in a vacuum, it loses energy at a rate equal to a damping factor times the average of the square of the displacement velocity of the vibrator—*i.e.,* $R\overline{\dot{\xi}^2}$ in which R is the damping factor and $\dot{\xi}$ is the displacement velocity. The origin of the damping may be found in the internal frictional forces inherent in the vibrator and its mounting. If now the vibrator is allowed to vibrate in a material medium like a gas or a liquid, it will be found to lose energy at a greater rate because of the production of mechanical radiation, which is transmitted through the medium. The average rate of radiation of energy by the vibrator as a source of sound can be written in the form $R_r\overline{\dot{\xi}^2}$, in which R_r now represents the damping due to the radiation. In order to maintain this radiation energy at a constant average rate, it is necessary to continue to supply energy to the vibrator at the same average rate by whatever is driving it. The radiation damping factor R_r that controls the rate of sound radiation by the source is found to depend on the size and shape of the source, as well as the density of the medium into which the radiation takes place and the velocity of sound in this medium. It also depends on the frequency of the radiation.

Radiation
damping
factor

The analysis of a pulsating, spherical shell, for example, indicates that its radiation efficiency for a given displacement velocity is much greater at higher frequencies than at low and this is one reason for the technological interest in ultrasonics.

A pulsating sphere is not usually a practical sound source. Similar considerations, however, apply to the membrane or diaphragm source in loudspeakers or sonar sources for underwater sound radiation. The mathematical expressions differ in detail, but in every case, other things being equal, the radiation efficiency increases with the frequency. At all frequencies the radiation efficiency depends directly on the density of the surrounding fluid medium. This dependence makes it difficult to get radiation from any source into a medium having a density that is small compared with that of the solid vibrator itself. Such a low efficiency accounts for the often-cited experiment of the difficulty of hearing a bell in an evacuated jar, commonly explained by saying that "sound is not transmitted through a vacuum." Actually, sound waves will still travel short distances without abnormal dissipation through gases at the low pressure prevailing in such an experiment; the difficulty is to get the wave into the gas from the source in the first place.

Sound beams. In view of the relatively long wavelength of audible sound, it is obviously difficult to produce beams of such sound without the use of extremely large acoustical mirrors and lenses. Nevertheless, at ultrasonic frequencies the sound wavelength becomes small enough to make sound beam formation practical, although never with the sharpness of light beams, even at the highest ultrasonic frequencies (*e.g.,* 10^3 megahertz). Such beams are basic to the operation of sonar, the underwater echo ranging system.

Practical simple sources: voice and siren. At the beginning of the section on sound sources various types of sound sources were mentioned; here are presented some specific examples. First to be considered as a source is the human voice (see SPEECH). It operates by forcing air from the trachea to vibrate the vocal cords. This in turn sets into vibration the air in the cavities of the throat and mouth; the resulting disturbance emerges from the lips. The part the vocal cords play in singing and speech has been established by motion pictures of their vibration.

Studies of the acoustical power of conversational speech have shown that in the case of an average speaker the overall level directly in front of the mouth can vary from 81 decibels at 15 centimetres (six inches) from the mouth to 65 decibels at 100 centimetres (40 inches). These decibel levels are relative to a standard reference level of 10^{-16} watt per square centimetre (the so-called minimum audible threshold). The mouth is only an approximate point source of sound emitting spherical waves, for at 100 centimetres directly back of the head, in the case just mentioned, the level is 62 decibels. This effect is due to diffraction by the head and is present to some extent in all practical sound sources of appreciable size.

The power in speech sound waves varies with the sound emitted, being much larger for vowels than for consonants. Ordinary speech involves the superposition of many waves of different frequencies. Moreover, the recognition of speech as mediated by the ear is not as simple as the detection of ordinary sound sources by appropriate electromechanical receivers. In the 1970s researches were directed toward improved methods of speech analysis and machine synthesis of speech. Much attention was also paid to measures of speech perception and intelligibility, as well as to better systems for the transmission of speech. Efforts were made to improve machine recognition of speech—*e.g.*, the phonetic typewriter, which types out spoken dictation. The coupling of sound recognition to computer technology was being developed.

Principle of the siren

The siren is one of the simplest sources of intense sound. In principle it consists of a disk containing a number of small holes placed at regular intervals around the same periphery. If this disk is rotated past an orifice lined up with the holes and compressed air is driven through the orifice, the result will be a succession of puffs of air. If the number of puffs per second (governed by the speed of rotation of the disk) is sufficiently large the result will be a periodic sound wave of frequency equal to the number of puffs per second.

Jet-edge tones and organ pipes. Another standard sound source operated by blown air is the jet-edge system coupled to an air column, as in the organ pipe. In the familiar type of pipe a jet of air impinges on the narrow edge of the bottom of the pipe. The resulting vibrations are coupled to those of the air in the pipe. The natural frequency of the edge tone coupled to the pipe increases approximately linearly with the velocity of airflow, but the coupling forces the fundamental tone of the pipe on the system for small velocities. When the flow velocity reaches the value for an isolated edge tone of frequency equal to the second harmonic of the pipe, the tone emitted rises to this harmonic, and the pipe is said to be overblown. The production of edge tones is associated with turbulence (vortex motion) in the airflow in the vicinity of the edge.

The large noise output of the jet-engine airplane in the 1960s and 1970s stimulated considerable interest in the theory of the generation of sound by airflow. The theory is intricate but leads to the result that the sound-power output of a jet of air is directly proportional to the fifth power of the mach number of the flow. The mach number is the ratio of the air flow velocity (relative to the stationary medium) to the velocity of sound in the stationary medium. Flow with mach number greater than one is known as supersonic.

Electroacoustic sources. The most important sound sources in technological use employ electrical energy. These may be divided into two main classes: irreversible and reversible. In an irreversible source the energy for sound radiation is supplied by a flow of fluid (air or water) under pressure and the electrical energy is used to modulate or control. A variable magnetic field, for example, can produce vibrations in a reed or diaphragm that will control the flow of compressed air through a variable orifice. Such a device can produce extremely intense sound waves with relatively high efficiency.

A reversible electroacoustic source transforms electrical energy directly into mechanical energy and thence into acoustical radiation. As the name implies, such a source can function in reverse to transform acoustical into electrical energy via some mechanical structure. It can therefore function equally well as a receiver of sound. As can be shown, under certain specific conditions, which hold approximately in practice, a good reversible source of sound will act also as a good receiver.

Transducers

Any system that transforms electrical to acoustical energy or vice versa is called an electroacoustical transducer. The important practical characteristics of a transducer used as a source are its acoustical output, its efficiency (ratio of the radiated power to the input power), its directivity, and its departure from linearity. The nonlinearity means the extent to which the mechanical force producing the vibrations leading to radiation is not strictly proportional to the electrical field or voltage applied to the input. Such nonlinearity leads to distortion in the radiated output. The directivity refers to the variation in intensity of the radiation with direction from the source.

To come to detailed examples, in the electromagnetic transducer, an alternating current of the desired frequency is passed through a coil that activates an electromagnet to produce an alternating force on a diaphragm of iron or some magnetic alloy. This is the type that has been employed in the conventional telephone receiver. The electrodynamic transducer is now more common. In this transducer the oscillating current passes through a coil suspended in a permanent magnetic field and undergoes the usual force. The coil, called the voice coil, is fastened mechanically to the sound-radiating surface. In the well-known loudspeaker, the latter is normally a cone attached to the voice coil at or near the vertex. The cone is the acoustical element, radiating directly into the surrounding medium. Various practical devices have to be incorporated in such a source to ensure large output and to diminish the distortion due to the resonant or normal mode frequencies of the cone.

The efficiency of an electrodynamic loudspeaker tends to vary directly with its linear dimensions. The speakers in common use in radio and television sets have efficiencies varying from 2 percent to 5 percent. By the use of an attached flaring horn this figure can be increased to about 50 percent. All these figures depend on the frequency range over which the speaker is driven. For a typical small-cone receiver 10 centimetres in radius, the efficiency is 2 percent only over the frequency range from 75 hertz to 1,000 hertz; at 10,000 hertz the efficiency drops practically to zero. The total acoustical power output of an electrodynamic loudspeaker can have maximum values ranging from 0.5 watt to 150 watts, depending on size and the associated electric circuit. Such a speaker at maximum output risks nonlinear behaviour—*i.e.*, large diaphragm motion that is no longer directly proportional to the applied electrical voltage.

Quartz crystal as an ultrasonic source

For ultrasonic radiation, the most important electroacoustic sources employ the piezoelectric and magnetostrictive effects already mentioned earlier in this section. A plate of quartz cut with its faces parallel to the optic axis of the crystal (a so-called *X*-cut), and used as the dielectric in a capacitor subjected to an alternating electric field, becomes an ultrasonic source with the frequency of the oscillatory field. It will, of course, vibrate with maximum intensity and radiate correspondingly if this frequency matches one of its own natural frequencies of vibration. The various resonance frequencies are inversely proportional to the thickness of the quartz crystal. Higher harmonics of the fundamental frequency of the quartz can be used to provide higher frequencies of radiation without using crystals too thin to be practical.

In practical use, a number of *X*-cut crystals are usually arranged in the form of a mosaic to provide a larger radiating surface. Many types of piezoelectric materials are now available in addition to quartz; these include Rochelle salt, barium titanate, and lead zirconate titanate. With such crystals, radiation of ultrahigh frequency has been achieved. By the mid-1970s it proved possible to reach frequencies of the order of 10^{10} hertz.

Magnetostrictive oscillators have also proved popular as ultrasonic radiators. If an alternating current passes through a coil surrounding a rod of magnetizable material already magnetized in the direction of its length, the rod will vibrate lengthwise with the frequency of the current. If this frequency coincides with one of the natural frequencies of the rod, large amplitudes can be obtained. A nickel rod 10 centimetres (four inches) long, for example, will radiate strongly at its fundamental frequency of about 24 kilohertz. Simple magnetostrictive sources are less efficient than piezoelectric devices, largely because of the loss of energy through the production of eddy currents. Magnetic materials, the so-called ferrites, are also in use in cutting down heating loss in magnetostrictive transducers. These ferrites are ceramic-type materials, such as ferric oxide mixed with nickel oxide and compressed so as to form solid rods.

Magnetostrictive transducers are often used to provide underwater acoustic beams, usually beyond the audible frequency range.

Doppler effect. In the discussion of sound sources it has been tacitly assumed that the source is at rest relative to the medium. Moving sources—*e.g.*, airplanes, submarines, and missiles of various kinds—however, are important in applications. A significant effect of such motion on acoustic reception is the perceived change in pitch in accordance with the Doppler principle. If a harmonic source of frequency f moves with linear velocity v_s with respect to a stationary medium and a receiver is moving with velocity v_r along the same line, the receiver will estimate the emitted frequency as a frequency f' that is given by

$$f' = f(V - v_r) / (V - v_s),$$

in which V is the velocity of sound in the stationary medium. In the formula v_r and v_s are reckoned as positive in the direction from the source to the receiver and negative otherwise. If, for example, the receiver is at rest and the source in motion, the apparent frequency at the receiver is lower than the emitted frequency if the source is receding and is higher than the emitted frequency if the source is approaching. Knowledge of the emitted frequency and measurement of the apparent frequency thus permit the determination of the velocity of the source, a matter of significance in military applications such as sonar.

RECEPTION OF SOUND

To be studied and applied, sound energy must be detected—*i.e.*, transformed to make it perceptible to people or sensible to cybernetic (communication and control) devices. This is the function of a sound receiver. Any reversible sound transducer can be used as a sound receiver.

The ear and hearing. The ear is the most important acoustical receiver in man's environment. Normal people can detect sound intensity as low as 10^{-16} watt per square centimetre and can stand intensities up to about 10^{-4} watt per square centimetre before pain ensues. These are figures averaged over representative samples of the population.

For the description of the structural details of the ear as a receiver of sound see the text article SENSORY RECEPTION. During the 1970s theories of hearing remained in a state of flux, with no one point of view commanding universal assent.

esholds
eeling
.
ring

The threshold values of hearing and feeling as a function of frequency are based on the work of a physicist in the United States, Harvey Fletcher, and confirmed by other investigators (Figure 9). The lower two curves relate minimum audible intensity to frequency. These curves rise at both ends of the audible range and have a minimum at about three kilohertz. The upper curve is the corresponding plot of the threshold of feeling. Intensities greater than those along this curve produce pain. Even so-called normal people show variation in thresholds and the curves represent averages over suitably large samples. Advance in age results in general in hearing loss, appearing in Figure 9 as a rise in the lower curves. The tendency is for the upper-frequency limit to decrease with age.

From *Speech and Hearing in Communications* by Harvey Fletcher (©1953); D. Van Nostrand Company, Inc.

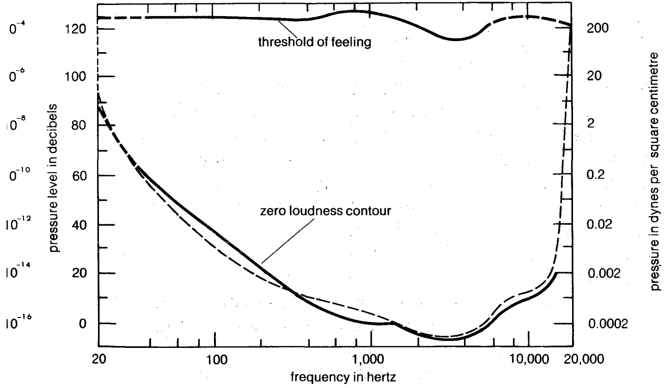


Figure 9: Auditory area between threshold values of feeling and hearing (see text).

Loudness is a sensation related to the intensity of the received sound but not directly proportional to it. Intensity is an objective physical quantity, whereas loudness is subjective and defined in terms of the so-called average listener. It is customary to distinguish between loudness level and true loudness. The former is measured by varying the intensity of a reference tone until it appears to a listener to be as loud as the sound stimulus under test. The number of decibels by which the reference tone has been raised above the minimum audible threshold is said to be the loudness level of the tone under test in phon units. This applies, strictly speaking, only if the reference tone has a frequency of one kilohertz and is placed at a distance of one metre from the head of the listener.

Measurement of true loudness

The measurement of true loudness demands the introduction of a scale such that when loudness of a certain number of units on this scale is multiplied by any arbitrary factor the magnitude of the new auditory sensation then produced will be heard by a listener as equivalent to the corresponding number of units times that of the original sensation. The unit of loudness introduced for this purpose is called the sone. It is defined as the loudness produced by a tone of one kilohertz at 40 decibels above the minimum audible threshold. By a statistical study of experiments in hearing, a loudness scale in sones has been established in which there is a logarithmic relation between loudness (L) in sones and loudness level (P) in phons, $\log_{10} L = 0.03 P - 1.2$, that holds for a reference tone of one kilohertz. This means that subjective loudness is equal to 10 raised to the power $.03P - 1.2$. The loudness level is of course a measure of acoustic intensity on the decibel scale.

Hearing with two ears, the so-called binaural effect, leads to the ability to detect the direction of the source of sound radiation. A source of sound on the right of the plane bisecting the line joining the ears and perpendicular to it will be recognized as on the right. This effect is not due primarily to difference in intensity at the two ears but rather to difference in phase or time of arrival at the two ears. This effect has proved useful in acoustic detectors for gun ranging and underwater sound. It is also of great importance in stereophonic radio broadcasting and phonograph reproduction. Still more important, indeed, than the directionality effect is the ability conferred by the binaural effect to distinguish different sounds; that is, for example, meaningful speech from noise. It seems that the difference between meaningful sound and noise is enhanced, by some means as yet not clear, by the difference in arrival time at the individual ears, so that the nervous system of the listener can make the necessary distinction.

Summation and difference tones

From the fact that the eardrum is an asymmetric vibrator (*i.e.*, its displacement for given applied pressure is not the same in the inward as in the outward direction) it follows theoretically that listeners should be able to hear summation and difference tones. What this means is that if two tones of widely separated frequencies f_1 and f_2 ($f_1 > f_2$) are presented to the human ear, in addition to these tones, the listener detects tones of frequency equal to their difference ($f_1 - f_2$) and equal to their sum ($f_1 + f_2$). As a matter of fact, tones of other combinations of frequency ($jf_1 + kf_2$), in which j and k are small arbitrary integers, have also been detected. If the original two tones, f_1 and f_2, are so close together that their difference is small compared with either, the usual phenomenon of beats is observed—*i.e.*, a periodic rise and fall in intensity with a frequency equal to their difference ($f_1 - f_2$).

(R.B.L.)

Microphones. In modern terminology the word microphone is applied to any electroacoustic transducer used for the detection of sound. Actually, it might legitimately be used to denote a transducer capable of transforming the excess pressure in a sound wave into any other more readily observable physical effect. For example, a sensitive flame formed by burning gas issuing at high speed from a very small orifice and kept just below the point of flaring was an early form of microphone. In the presence of even a rather weak sound wave, the flame makes a perceptible jump. It can be made particularly sensitive to high-frequency, inaudible sound.

Such early microphones have now been entirely superseded by the electroacoustic variety. Such an electroacoustic microphone transforms the excess pressure changes due to an impinging sound wave into alternating electric currents that can be amplified and visualized on an oscillograph.

The conversion, within a microphone, of sound-pressure variations in the air into corresponding electrical waves

Operation of the microphone

occurs in two operations that usually take place simultaneously. In the first operation, the sound wave impinges on a surface, generally referred to as a diaphragm, which is capable of slight movement. Variation of air pressure on this diaphragm causes it to move to and fro in a manner corresponding to the movement of the air particles. In the second operation, the diaphragm by its movement causes a corresponding change in some property of an electric circuit. Thus, for example, the displacement of the diaphragm may cause variations in the resistance of carbon granules or variations in the motion of a coil or conductor in a magnetic field. In each case, motion of the diaphragm produces a variation in the electric current.

Thus, the sound vibrations are first converted into mechanical motion, and then the mechanical motion is converted into variations in electric current. Hence, in studying the operation of a microphone, two things must be considered: (1) the mechanical movement of the diaphragm; and (2) the method whereby this movement sets up the desired variations of electric current. Either part of the process may be taken as a basis for classifying microphones.

The carbon microphone

Any microphone that has its diaphragm exposed to the sound wave on only one side is a pressure-operated type; that is, the displacement of the diaphragm is approximately proportional to the instantaneous pressure of the impressed sound wave. Outstanding among pressure types and of great historical importance is the carbon microphone, used mostly in telephony. Another pressure type, typically consisting of a thin stretched diaphragm separated by about one-thousandth of an inch (0.025 millimetre) from a parallel fixed plate, is the condenser microphone. The diaphragm acts as one plate of a variable capacitor (device for storing electricity) that is connected in series with a resistor and a direct current source. The motion of the diaphragm in response to sound waves varies the capacitance, producing a varying electric current. Condenser microphones are used for sound measurement and for high-quality recordings. Still another pressure type is the crystal microphone, used in hearing aids, sound recording, and office dictating machines. This microphone depends for its operation upon the piezoelectric effect (*i.e.,* the generation of a voltage by the deformation of a crystal by a moving diaphragm). Crystals of barium titanate and Rochelle salts exhibit great piezoelectric activity; they are commonly used in crystal microphones (see also ELECTRONICS: *Piezoelectric devices*).

The dynamic or moving-coil microphone falls within the general class of moving-conductor microphones—*i.e.,* those in which an electric voltage is generated by the motion of a conductor in a fixed magnetic field. In the dynamic microphone a coil is attached to the moving diaphragm; motion of the coil in a fixed magnetic field induces a voltage in the coil. Dynamic microphones are used quite widely in recording, broadcasting, and public-address systems, especially when ruggedness is essential. Among the other types of microphones are the magnetic microphone and the ribbon microphone.

By proper design a microphone can be made to pick up sound from a single direction only, resulting in what is called a cardioid pattern. Such a microphone can be useful in discriminating against reflected and unwanted sounds and noises. Bidirectional microphones, in which the pickup is essentially the same from either front or back but zero from the sides, also are useful for specific applications. Microphones with approximately equal pickup from all directions are called nondirectional or omnidirectional. (E.I.G.)

CRITERIA FOR GOOD HEARING CONDITIONS

Good hearing conditions for an audience listening to speech or music, whether in an auditorium, courtroom, church, outdoor theatre, or living room, depend on four basic conditions. The space must be quiet, the sounds to be heard must have adequate loudness, there must be a good distribution of sound, and the sounds must be properly blended, with adequate separation for good articulation of music or speech.

Absence of noise. Perhaps the single most important factor in providing good hearing conditions is the absence of any interfering noise. Background noise that can be useful in providing privacy and freedom from distraction in the office or apartment has no place in the room in which there is serious listening to speech or music. Not only must the air-handling system be inaudible, but there must be no noise transmitted from other spaces in the building or from outside (see also below *Noise control*).

Adequate loudness and distribution. Obviously, the sounds to be heard should be loud enough and should be uniformly distributed. People sitting in the front of a room should not be subjected to greater volumes of sound than those in the rear. Hot spots, where the sound is too loud, and dead spots, where the sound is too faint, can be as unsatisfactory as seats in which everything is heard twice due to delayed echoes. Adequate loudness and good distribution of sound are determined largely by the shape and surface finish of the room.

Figure 10A illustrates how sound is distributed when an audience is seated on level ground outdoors. In a free field, a sound wave decreases in intensity by a factor equal to the square of the distance from the source. As the sound wave grazes the heads and clothing of the sound-absorptive audience, however, there are additional losses. People seated far back in such an audience receive less sound energy not only because they are far away from the sound source but because some of it is absorbed by the people in front of them. In such a seating area, loudness and distribution requirements are poorly met.

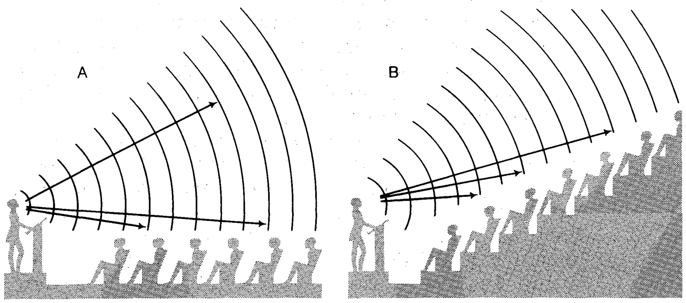


Figure 10: (A) Distribution of sound when audience is seated on ground level outdoors. (B) Distribution of sound in an outdoor theatre of the classical Greek design.

Sound mirror

The ancient Greeks and Romans solved the problem by placing audiences on steep hillsides, which resulted not only in much better sight lines but in significantly smaller losses by absorption of the intervening audience (Figure 10B). Good sight lines almost always mean good hearing lines. An alternative to the steep hillside is to raise the source of sound high above the audience, but this is not usually convenient. A more satisfactory indoor arrangement is to let the ceiling of the room act as a sound mirror, bringing reflected sound down on top of the audience to reinforce the direct sound (Figure 11). In many auditoriums more sound energy is received from the ceiling and other reflectors in the room than directly from the source of sound itself. The ceiling is the most important surface in the room in assuring adequate loudness and good sound distribution. A ceiling in any room that must have good hearing conditions for an audience must always be hard and sound-reflecting, not sound-absorbing. This principle is often overlooked; in fact many auditoriums and other

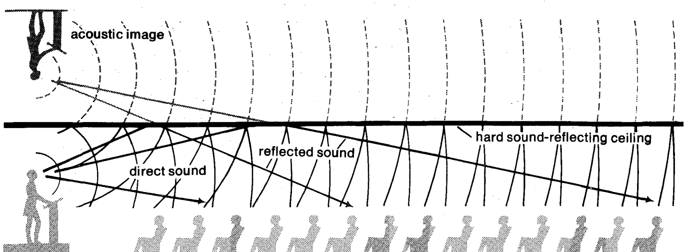


Figure 11: Distribution of sound in a well-constructed auditorium.

buildings for audience use have been deliberately but mistakenly equipped with sound-absorbing ceilings. Such use of sound-absorbing material inevitably produces bad hearing conditions because the room has been deprived of its most useful sound reflector. Sound-absorbing material is useful in rear wall surfaces to control echo and reverberation.

An equally important function of the sound-reflective ceiling is the mixing of sounds that emanate from various parts of a large performing group; *e.g.*, a congregation in a church or an orchestra or a chorus on a stage. Whenever people perform together, it is essential that all members of the group hear each other. An orchestra cannot perform satisfactorily in a conventional stagehouse surrounded by velour draperies, because the various instrumentalists cannot hear each other. Some type of overhead reflector must be used.

If a listener is to receive sound energy by reflection from the ceiling, he must be seated directly under the ceiling; seats in a deep underbalcony area are usually the poorest in the house. The undersurface of such a balcony should always slope upward toward the front, with a sufficient opening to permit the rearmost seats a line of sight to part of the ceiling.

The reflections from ceiling and wall surfaces useful to the listener in improving loudness, clarity, and articulation must arrive at the listener's ear not more than about 30 milliseconds (thousandths of a second) after the initial sound reaches him. This means that the sound will have travelled over a path not more than about 30 feet (nine metres) greater than the distance between source and listener. Reflections delayed longer than this tend to detract from clarity and add confusion. If a reflection is delayed longer than about 70 milliseconds, it begins to be audible as a separate sound or echo.

To provide the members of the audience in the front half of the orchestra, or main floor, of an auditorium with these early reflections, the ceiling near the performing area cannot be more than 30 or 35 feet (nine or 11 metres) above the floor. Ceiling heights of 60 or 70 feet (20 metres) subject much of the main-floor audience to poor sound quality and lack of clarity. In the balconies and toward the rear of the main floor, reflections with a small time delay are received, even if the ceiling is rather high. (R.B.N.)

Sound recording and reproduction

Sound recording consists of transcribing those vibrations in air that are perceptible as sound onto a storage medium such as a phonograph disc. In sound reproduction the process is reversed, and the variations stored on the medium are converted back into sound waves. The degree of accuracy achieved in these two processes is referred to as the fidelity of the sound system and may vary widely with the intended application. For business dictation, and many other purposes where the spoken word is recorded and reproduced, mere intelligibility is sufficient. For full enjoyment of music the fidelity requirement of a sound system may be very high, demanding precision equipment and techniques at every step in the recording–reproduction process.

Over the years a variety of storage media has been used for sound recording and reproduction, but the three principal systems that have survived may be loosely grouped into the mechanical (the phonograph disc), the magnetic (recorded tape), and the optical (sound tracks for motion pictures and the digital Compact Disc, or CD). All three today share a common reliance on electronic circuits, such as amplifiers, but each developed sufficiently independently to warrant treatment in separate sections in this article. Concepts and techniques specific to high-fidelity sound systems and contemporary developments in digital recording and playback techniques are treated in the final two sections.

MECHANICAL SYSTEMS

A phonograph disc stores a replica, or analogue, of sound waves as a series of undulations in a sinuous groove inscribed on a rotating surface by the vibrations of a stylus. When the record is played back, a stylus responds to the undulations, and its motions are then reconverted into sound.

Early history. *The Edison phonograph and its successors.* A precursor of the phonograph appeared in 1857, when the French inventor Léon Scott constructed a device he called the phonautograph. This device recorded the movements of a diaphragm or other vibrating body in response to sound as a wavy line on the smoked surface of a rotating cylinder. The first machine that could both record and reproduce sound was invented in the United States by Thomas A. Edison in 1877. The original Edison recordings took the form of indentations embossed into a sheet of tinfoil by a vibrating stylus attached to a diaphragm. The tinfoil was wrapped around a cylinder that was rotated as the sounds were being recorded. The sound waves were recorded as variations in the depth of the groove, a process later referred to as hill-and-dale recording. For reproduction the groove was again run under the stylus tip, with only enough pressure to maintain contact with the groove bottom.

Edison's phonograph

In 1885 two U.S. inventors, Chichester A. Bell and Charles Sumner Tainter, patented a machine similar to Edison's except that it cut, or engraved, the groove in a cylindrical surface of wax with a sharp tool carried by the vibrating diaphragm. This method gave much better sound reproduction than Edison's indented tinfoil. Another inventor, the German-born Emil Berliner, was responsible for a radical departure in sound recording. During the recording process Berliner's mechanism traced its wavy line as a spiral on a flat disc rather than as a helix on a cylindrical surface. The disc was made of zinc and was coated with a thin layer of a fatty substance that protected it from the action of acid except where the scribe had traced its line. An acid bath then left a groove, the width and depth of which could be controlled by the duration and strength of the acid treatment. Berliner thus obtained a record that could be played on an appropriately designed reproducing mechanism, which he called the Gramophone.

Berliner's Gramophone

Berliner did not use the etched master as the record to be played, however; rather, a negative was made from the master by electroforming (see below *Electroforming*), and the negative was then used to mold records in a thermoplastic material, a procedure that subsequently became standard. Berliner also developed lateral recording, in which vibrations are recorded as sidewise deflections of a groove of uniform depth rather than as variations in the depth of the groove. Starting in about 1900 lateral recording attained wide acceptance both in the United States and in Europe.

As the phonograph and record business grew steadily in Europe and the United States, various technical innovations were introduced that improved the quality of sound reproduction. Methods of molding cylinder records of a relatively hard thermoplastic were introduced in 1901, and in 1908 cylinders made of a new material called Amberol enabled recording grooves to be spaced 200 to the inch, thus doubling the playing time to four minutes per cylinder. In 1912 Edison introduced Blue Amberol, a further improvement in material, and the following year he began making discs and machines to play them. He continued for some years, however, to issue cylinder recordings as well. The 78-rpm (revolutions per minute) record, which virtually replaced cylinders by about 1915, had a playing time of approximately 4½ minutes per side.

The first phonograph machines had to be hand-cranked continually, although some speed-stabilizing effect was provided by flywheels. In one of his early patents Edison stressed the importance of correct and uniform speed, and experiments with battery-operated electric motors were undertaken. These proved sensitive to changes in load and voltage, and various mechanical governors were used to reduce speed variations. The most satisfactory of these governors were those with a friction disk, the friction of which increased rapidly if the speed exceeded the desired value. While Edison was reluctant to abandon the electric motor, he ultimately used a spring-driven governed motor developed by Frank Capps, a pioneer in many phonograph improvements. Spring motors were also developed by the Columbia Phonograph Company and by Berliner, so that by the end of 1896 all three companies were using governed spring motors. Extensive use of electric motors was delayed until radio–phonograph combinations were introduced in the 1930s.

Acoustical horns. The most serious limitation on early phonograph developments was imposed by the need to rely entirely on acoustic methods of recording and reproduction. Performers huddled close to large, clumsy recording horns, and only the most powerful instruments and voices recorded well. At the small end of the horn the sound waves were concentrated onto a diaphragm to which the cutting stylus was attached. During playback only the acoustic output from the stylus–diaphragm combination was available to be fed into the small end of the horn. This severely limited the output volume and led to many attempts to increase the efficiency and improve the tonal quality of phonograph horns.

The easiest horn to construct was of simple conical shape. As experience accumulated, the sizes of horns increased and their shape began to change. In 1906 the Victor Talking Machine Company marketed a machine with the horn contained within the cabinet. This arrangement facilitated increasing the horn size and also made it desirable to keep down the size of the middle part of the horn, thus favouring a shape somewhat like that of an exponential horn. In an exponential horn the cross-sectional area doubles with each increase of x inches in distance along the axis from the small end. If it is a round horn, the diameter doubles with each increase of $2x$ in axial distance. Horns can be bent, or "folded," to conserve space without too seriously impeding their performance, provided the difference between the shortest and longest paths around the bend is not more than about one-third the shortest wavelength.

The exponential horn

After about 1920 the exponential horn became the nearly universal choice of engineers. A theoretical analysis of its characteristics was published in 1919 by a U.S. physicist, Arthur Gordon Webster, and the subject was further elucidated by two U.S. electrical engineers, Clinton R. Hanna and Joseph Slepian, in 1924. It was not until the introduction of the Orthophonic design in 1925, however, that a definitely planned exponential expansion was followed, with considerable increase in length and with an end opening occupying practically the whole front of a large cabinet.

Electrical recording and reproduction. Shortly before World War I the U.S. scientist Lee De Forest invented the Audion (triode) vacuum-tube amplifier, a device that could multiply sound power a thousandfold or more with little loss of quality. Although work toward applications of the Audion amplifier began as early as 1915, World War I intervened, and it was not until the early 1920s that the transition from acoustical to electrical methods took place. With the advent of electrical recording, performers could work under more natural conditions; microphones collected the sound, which was then amplified electrically to the necessary level for recording with an electrical disc-cutting head. The Audion amplifier was also applied in sound reproduction, leading both to the replacement of the acoustical horn by the electric loudspeaker and to the development of the electrical phonograph cartridge.

The vacuum-tube amplifier

Record materials and production processes. Tinfoil and the other earliest recording media were supplanted by materials of two general types: the wax disc (actually a heavy-metal soap with wax gel) intended for subsequent processing and duplication; and the lacquer-coated disc, usually on an aluminum base. The latter was originally intended primarily for instant playback, but improvements in formulations and cutting techniques eventually made lacquer blanks the choice for making master discs suitable for duplication. Since about 1950 electrically heated jewelled styli with precisely ground cutting edges and a special burnishing facet have been used to achieve a smoothly cut, low-noise groove in the master lacquer disc.

Electroforming. To make copies of the master disc, dies must be made whose surfaces are negative replicas of the master-record surface. Electroforming techniques, pioneered by Berliner, are now used to reproduce the surface to an accuracy of better than 0.000001 inch (0.000025 millimetre). In order for this process to be used, the master-record surface is first made electrically conductive by a wet-silvering method similar to that used for silvering mirrors. It is then plated, either with copper or (now more commonly) with nickel, creating a negative mold of the silvered-lacquer master.

After separation from the master record, the electroformed metal master disc can be used directly in a plastic-molding press to produce records if only 100 or so are desired. For records in commercial quantities many pressing masters, called stampers, are required. It has been the practice, therefore, to electroform a metal positive from the metal master. This positive, called a "mother," can be played as a record for quality approval and then used to generate a number of stampers by electroforming. By using a metal master to generate a number of molds and each mold to generate a number of stampers, it is possible to make many stampers, and thus thousands of records, from each master disc. The records themselves are molded by placing a "biscuit" of thermoplastic material between two stampers in a heated hydraulic press that melts the material sufficiently so that it will conform under pressure to the shape of the grooves. The press is then cooled to set the thermoplastic material, and the record is removed for finishing (*e.g.,* removal of the excess material, or "flash," around the edge), packaging, and sale.

The master, "mother," and stamper

Until about 1930 records were made of shellac, a natural thermoplastic, combined with a finely ground filler, usually limestone and clay. This rather abrasive compound served to extend record life, given the heavy weight (100 to 200 grams, or 3.5 to 7 ounces) of early phonograph pickups. These pickups normally used steel needles, which were quickly ground down by the abrasive records to conform to the groove shape. This increased the contact area between the stylus and groove significantly, thus decreasing the force of the stylus per unit area, which reduced record wear. (Steel was later replaced by harder stylus tips, made of sapphire or diamond, which resisted being ground into conformity with the groove, thus increasing record wear.) In the 1930s synthetic thermoplastic resins that used little or no filler material became the choice for record material.

Phonographs and record changers. The advent of commercial radio broadcasts, beginning in 1919 with Pittsburg's KDKA, rapidly captivated public interest and stimulated the next major developments in record-playing equipment. Radios came equipped at first with headphones, and later with amplifiers and electric loudspeakers; they did not need hand cranking, someone else changed the records every 4½ minutes, and, of course, the music was free. For a while inexpensive single-play phonographs that could be plugged into a socket on the back of the radio stimulated the growth of record collections by those who wanted to hear their own choice of music. As collections expanded and longer and longer pieces of music were recorded, however, the need grew for a device designed to permit a stack of records to be played in sequence without human intervention.

While a number of early patents were granted on automatic record-changing devices, very few machines reached the hands of the public until after the mid-1930s. By 1940 many manufacturers were equipping their phonographs with record changers, despite the fact that the edges and centre holes of the heavy 78-rpm records were frequently damaged by the changer mechanisms.

Record changers

The use of electric motors both to turn the phonograph platter and to operate an automatic changer mechanism became extensive with the introduction of radio–phonograph combinations. Typically, a small high-speed induction motor was used, which communicated its rotation to the inner rim of the turntable platter through one or more rubber-edged idler wheels. When the tripping mechanism at the end of the record side was triggered, the motor turned a series of gears connected to a mechanism that lifted the pickup arm above the record just played, returned the arm to the side, dropped the next record down from the stack, and placed the tone arm down on the new record in its lead-in groove.

After World War II many improvements were made in record-changer mechanisms. The use of vinyl record material considerably reduced damage to the record by the selector mechanism. In addition, improvements were made in the turntable drive system, tone arm, phonograph cartridge, and tone arm tripping mechanisms. These, in turn, made it possible to use a lighter tracking force, thus reducing wear on the stylus and record alike.

Modern sound reproduction. *The long-playing disc.* The beginning of the modern era in sound recording and reproduction came in 1948 when Columbia Records introduced the Long-Playing (LP) Record made of Vinylite. Twelve inches (actually, 11⅞ inches, or 30.2 centimetres) in diameter, with a rotational speed of 33⅓ rpm, LPs used fine-groove (often called "microgroove") recording techniques to achieve up to 30 minutes of playing time. Shortly afterward the RCA Victor Company introduced the seven-inch 45-rpm disc with the same groove size, which played for up to eight minutes. These new records offered such improvements in cost, convenience, and performance that they rapidly supplanted the old 78-rpm discs. The necessary accuracy of the recorded groove can be judged by the requirements established for long-playing records: the bottom radius of the groove must be less than 0.00025 inch (0.0064 millimetre) and the top width no less than 0.0022 inch, giving a depth of about 0.00125 inch. The standard also specified a V-shaped groove with an included angle of 90 degrees, a fact that was to become important in the development of stereophonic records.

reo-
ɔnic
tems

Stereophonic phonograph records, with two separate channels of information recorded in a single groove, became a commercial reality in 1958, less than a year after the system had been publicly demonstrated. In a live musical performance the sounds of various instruments and voices come from different directions. Stereophonic systems, by reproducing sound with spatial perspective, produce a far greater sense of realism than single-channel, or monophonic, systems (see below *High-fidelity concepts and systems*). A stereophonic system requires two independent channels, with separate microphones for recording and separate loudspeakers for reproduction. On a stereo disc the outer groove wall contains the left-channel signal, the right channel being recorded on the inner groove wall. The standard stereophonic playback system employs a single stylus with its axes of movement inclined at 45 degrees to the record surface, the pickup being designed to produce two practically independent signals in accordance with vertical and horizontal motion components. Operation at 45 degrees provides a symmetrical arrangement that offers advantages both in the design of disc cutters and reproducers and in quality of reproduction, because the electronic characteristics of the two channels are alike. In addition, good reproduction is possible over a monophonic system because information from both stereophonic channels is contained in the resultant lateral modulation.

Phonograph cartridges. A phonograph cartridge, or pickup, is a device that when actuated by the record groove produces a usable electrical signal proportional to the mechanical signal (music) recorded in the groove. It is typically mounted in a tone arm, which allows its stylus to track the spiral groove on the record. Although steel, and even cactus, needles were used in early phonographs, a great improvement was made possible by the development of jewelled stylus tips that could be ground to precise dimensions.

lus
pe
l
:

For proper tracking it is important that the stylus retain contact with the two sides of the V-shaped record groove, and its "point" should ride above the bottom of the V (where there is no musical information to be extracted, but where there may be grit that can generate noise). Until the development of stereo records (and even today in inexpensive equipment), all playback styli were conical in shape, with a precisely rounded point. For the 78-rpm record a conical stylus with a 0.003-inch (0.076-millimetre) tip radius was desirable and for long-playing records, a 0.001-inch radius. For stereo records a slightly smaller (*e.g.,* 0.0007-inch) tip radius was recommended, but high-quality cartridges today use biradial, or elliptical-shaped, tips that are narrower from front to back (the direction of groove motion) than from side to side (the direction of the groove undulations). The tip dimensions of a biradial playback stylus are typically 0.0002 inch by 0.0007 inch. While sapphire tips were once used, modern styli are made from natural or industrial diamonds. All playback styli are subject to wear, which can create edges that literally chisel away the high frequencies recorded in the groove. A modern diamond stylus should be checked for wear with a specialized microscope after approximately 500 hours of playing time.

Phono cartridge cantilevers, to which the stylus is attached, are usually made of a thin aluminum tubing, although various other materials, such as magnesium, boron, and titanium, are sometimes used. The cantilever must be rigid enough not to bend, twist, or break, but should not resonate at any audio frequency. The effective mass of the stylus tip, cantilever, and generating element(s), together with the compliance of the pivot, contributes to the mechanical impedance of the pickup. Low mechanical impedance has received considerable attention because its attainment permits a low value of stylus tracking force. The less the stylus force, the less is the wear on the stylus and the record groove. Concentrated effort on the part of pickup designers has reduced the mechanical impedance to the point where high-quality pickups can maintain proper contact with the walls of the record groove with only 1.0 to 1.5 grams (0.035 to 0.053 ounce) of stylus pressure.

Modern
phono-
graph
cartridges

To generate an electrical signal from the motions of the stylus–cantilever assembly, phonograph pickup designers have explored nearly every known technique. These include the piezoelectric effect, variable resistance, variable capacity, and variable light-beam photoelectric effects. In general, it is possible to design a pickup using any principle that will produce an electrical signal from a mechanical movement. The most popular types of high-quality phono cartridges employ miniature magnets and coils. In order to produce an electrical audio signal, the stylus assembly either moves the magnet(s) in relation to the fixed coils, moves the coils in relation to the fixed magnet(s), or moves a magnetically sensitive element between the magnet(s) and coils. The purpose in each case is to cause a magnetic flux field to impinge on the coil that varies with the movement of the stylus, which in turn reflects the undulations in the record groove walls.

Stereophonic pickup design follows the same basic principles used for monophonic pickups except that it is necessary to produce two independent signals using a single stylus actuated by the independent signals recorded on the two walls of the record groove. Thus the pickup has two independent electric-generating units connected to and actuated by a common stylus.

Motors and drive systems. Although numerous turntable drive systems have been used since the earliest days of the phonograph, today's quality record players are likely to use either of two systems: a belt drive that couples a high-speed motor to the slow-moving turntable platter; or a direct-drive system employing a special semiconductor-controlled motor whose shaft rotates at the playing speed of the record. In the latter type of system the heavy metal platter that supports the record sits directly on the motor shaft. Precision machining and careful attention to design details are necessary to minimize "wow" and "flutter," low-frequency and high-frequency periodic changes, respectively, in signal frequency that result from momentary variations in turntable speed. "Rumble," a low-frequency noise picked up by the cartridge, is caused primarily by vibrations from the phonograph motor that are transmitted through the motor supports or the drive system (see below *High-fidelity concepts and systems*).

MAGNETIC SYSTEMS

An audio tape deck is an instrument for electromagnetically imposing or detecting varying magnetic patterns, which correspond to sound waves, on a moving magnetizable surface.

Poulsen's
tele-
graphone

Early history. In 1898 Valdemar Poulsen, a Danish inventor, applied for a patent on a device that stored electrical information by means of magnetizing areas of a steel wire. With later improvements his telegraphone, as it was called, was capable of recording continuously for 30 minutes on a length of piano-wire steel that moved at a speed of 84 inches (213 centimetres) per second. The device did not find wide application, however, and for many years interest in magnetic recording was largely confined to the laboratory.

Early attempts at magnetic recording were plagued both by high distortion and noise and by the lack of a suitable medium on which to record. In 1927, however, the scientists W.L. Carlson and G.W. Carpenter patented the use of ac (alternating-current) bias (see below) to reduce distortion and increase the amount of signal that could be recorded in relation to noise. In the same year the first U.S. patent was issued for magnetic tape made by drying a liquid containing magnetic particles on the surface of a strip of paper. A German patent was issued in 1928 for a similar process. In England, Germany, and the United States during the 1930s work progressed on a magnetic recording system to record and synchronize sound with motion pictures. In 1936, at the Berlin Radio Fair, the Magnetophon Company of Germany demonstrated the potential of tape recording by using a Magnetophon recorder and a plastic-base magnetic tape manufactured by BASF to reproduce a performance by Sir Thomas Beecham and the London Philharmonic Orchestra. During World War II Magnetophon machines served the interests of Nazi propaganda organizations, and following the war the machine's principles were imitated and improved upon in the United States and in other countries.

Recording media. Modern recording tapes consist of a plastic-base film coated with magnetic powder that is held in place by a binder. Mylar and other polyester plastics are widely employed as the tape base, which is sometimes "back-coated" with a conductive material to eliminate the buildup of static electric charges and to improve its winding characteristics. For open-reel recording plastic-base films of 1.5 and 1.0 mils (38 and 25.4 micrometres) are commonly used. For cassettes the most common base thicknesses are 0.47 and 0.31 mils.

Common tape materials

The magnetic material most widely used in coating tape is a type of iron oxide, Fe_2O_3, in needle-shaped particles roughly 0.6 to one micrometre (one micrometre equals 0.000039 inch) in length and one-tenth or less that in diameter. The way in which the magnetic powder is prepared, the impurities deliberately introduced or accidentally present, and the physical size and shape of the particles have much to do with the audio characteristics of the finished tape. To meet the demand for extended high-frequency response at the low tape speed of 1⅞ ips (inches per second [4.56 centimetres per second]) used by cassette recorders, Du Pont developed a new magnetic material, chromium dioxide (CrO_2), in the late 1960s. In the mid-1970s TDK introduced a cobalt-adsorbed ferric oxide particle, Super Avilyn, that matched the high-frequency response of CrO_2 but had a higher overall output. Pure metal-particle (nonoxide iron) tape was introduced in 1978 by the 3M Company to provide still better high-frequency response at high recording levels.

The magnetic powder is mixed with a binder, a type of glue that must not only bond the magnetic material to the base material but also hold the particles apart from each other when wet, so that they will remain evenly dispersed throughout the coating. Once dried, the binder must adhere to the base material and retain its flexibility, yet it must not adhere to adjacent layers even when the finished tape is tightly wound and stored at high temperatures. Finally, it must retain its characteristics without change throughout much usage and over many years.

The wet mixture of binder and magnetic particles is precisely coated onto large rolls of the film and passed over an intense unidirectional magnetic field to orient the needle-shaped particles uniformly, and the coated film is then passed through drying ovens. Following this, the tape is pressed between sets of calendering rollers to polish its surface and ensure a uniformly thick coating. Typically, tapes intended for open-reel use may have a magnetic coating thickness of 0.56 mil, while those for cassettes may use a magnetic coating of 0.24 mil. The large rolls of finished tape are then slit to proper widths for their application and are wound onto appropriate reels (see below *Recording formats*).

Recording and playback processes. *Recording.* The electrical signal to be recorded is usually applied to the tape by means of a "head," consisting of a coil wound around a core of magnetic iron, that has a gap at the point where the tape moves across its surface. The current in the coil produces a magnetomotive force across the gap, magnetizing the particles in the tape. The strength and direction of final magnetization of a particular place on the tape is determined by the strength and direction of the magnetic field at the moment that place leaves the influence of the gap. Thus the tape receives a magnetic record containing direction (north or south pole), amplitude, and linear dimension (along the tape) corresponding to the direction (plus or minus), amplitude, and time of the original electrical signal.

In modern ac-biased recording the tape is passed across an erase head before it comes to the record head. Erasure does not remove the magnetism of the tape particles; it simply removes any previous pattern in their north–south or south–north orientation, so that the distribution of north-seeking and south-seeking particles is random. In this condition the sum of the particles' magnetic fields is zero at all measurable points. When the neutralized, or "demagnetized," tape reaches the record-head gap an ultrasonic alternating current is added to the signal current; it is much higher in amplitude than the signal current and several times greater than the frequency of the highest frequency signal to be recorded. While crossing the gap of the record head, the magnetic particles on the tape are subjected to many cycles of the bias current and leave the head with a magnetization that is proportional to the signal. The purpose of the ac bias is to assure low distortion and a high signal-to-noise ratio.

Losses are encountered in recording through eddy current and resistive and capacitive losses in the windings and core of the head. Space between the head and tape causes very large losses, particularly at high frequencies (short wavelengths). The gap must be straight and true, and the surfaces in contact with the tape must be very smooth in order to obtain the best possible transfer of the signal pattern onto the tape. Optical polishing methods are commonly used for finishing the surfaces of the gaps and faces of the head that will be in contact with the tape.

Record-head gap lengths

The gaps used in record heads vary in length (direction of tape motion) depending on their application. For open-reel recording a typical record-head gap length is 100 to 500 microinches (2.5 to 13 micrometres). For slow-speed cassette operation record-only head gaps may be as long as 275 microinches, though cassette heads intended for recording and playback are typically 50 microinches. The tolerances on the trueness of the faces and edges of the gap are usually only a few microinches. The width of the cores and the resulting recorded tracks on the tape may vary from approximately 0.021 inch (0.53 millimetre) for stereo cassettes to 0.25 inch for professional monophonic recording.

Playback. In reproduction the tape is either passed over the same head used in recording (as in most cassette decks) or over one similar to it but with a smaller gap length designed for optimum playback performance (almost universal in open-reel recorders and increasingly used in high-quality cassette machines). The magnetized patterns on the tape passing over the gap in the reproducing, or playback, head cause the magnetic flux in the core to change, generating a voltage in the coil proportional to the speed of the tape and the amplitude and frequency of the recorded signal.

Because the output from the playback head varies with the number of changes per unit time in the magnetic patterns on the tape, its output rises with increasing frequency. For example, there are half as many north–south pole alterations in the octave from 20 Hz (hertz) to 40 Hz as there are in the next higher octave, from 40 Hz to 80 Hz, so the output (given equal recorded levels) from the playback head doubles. This process continues until the gap length of the playback head begins to approach one-half the recorded wavelength of the signal. In order to equalize the frequency response of the playback machine at all frequencies and assure that tapes can be freely exchanged from one recording machine to another reproducing machine, detailed international equalization standards for tape reproducers have been established.

ıyback-
ad gap
ıgths

Maximum playback head output is reached when the gap length (popularly called gap width) equals one-half the recorded wavelength, although some losses have already begun to occur. At one wavelength there is no difference in the magnetic field at the two sides of the gap, so the output from the head is zero. The recorded wavelength (λ) equals the tape speed divided by the frequency. The wavelength of a 20,000-Hz signal at a tape speed of 15 ips (38.1 centimetres per second) is thus 750 microinches (19.05 micrometres), and the requisite playback gap length should thus be significantly less than half of this. At the cassette speed of $1\frac{7}{8}$ ips the same 20,000-Hz signal has a wavelength of 93.75 microinches, so a playback gap length of less than 39 microinches is desirable.

Tape-recording machines. A wide variety of mechanisms has been devised to make use of tape recording in many different applications. These include open-reel machines, which employ 7- or $10\frac{1}{2}$-inch (18- or 27-centimetre) reels of tape, the cassette, and the microcassette. (The once-popular eight-track cartridge format has largely become obsolete.) Each of these is available in a wide variety of equipment, and cassette and microcassette players and recorders are frequently built into home, portable, or automobile AM/FM stereo radios. One of the advantages of magnetic recording is the ability to erase and reuse the recording medium thousands of times with little or no loss in the recording quality. Erasure may be accomplished either with the erase head in the recorder or by subjecting the entire tape to a strong alternating magnetic field provided by an external device and slowly withdrawing the tape from the erasing field. This is called bulk erasing and is similar to the method used by a jeweller when demagnetizing a watch. Professional sound-recording engineers frequently use both methods to make sure all the previous recording has been removed from a tape before it is reused.

The most common drive mechanism found in tape recorders is the capstan drive, in which the tape is pulled forward by the friction between a pinch roller, consisting of a cylindrical piece of rubber, and a motor-driven cylinder (capstan), whose rotational speed and diameter determine the speed of the tape. In many machines both the feed reel, which holds the tape before playing, and the take-up reel, on which the tape winds as it plays, are driven by separate motors to maintain the tape in tension and to permit rapid winding of the tape in either direction. The general mechanisms for magnetic recording and playback described earlier apply to all the formats.

Recording formats. *Open-reel.* High-quality professional-type open-reel magnetic tape recorders are used for recording the master programs for phonograph discs, prerecorded tapes, FM broadcasts, and motion pictures. Such recorders normally operate at 30, 15, or 7.5 ips (76.2, 38.1, or 19.05 centimetres per second) and often use 2-, 1-, or $\frac{1}{2}$-inch (50.8-, 25.4-, or 12.7-millimetre) tape widths to record 24, 16, 12, eight, or four separate tracks on a single reel. A single voice or musical instrument or small group of instruments is recorded on a track. In transfer to a disc, prerecorded tape, or sound motion-picture master record, the material recorded on the tracks is mixed down (combined and rerecorded) onto two or four tracks on $\frac{1}{4}$-inch or $\frac{1}{2}$-inch tape for stereo or quadraphonic mastering.

pe deck
›eds

Consumer open-reel recorders have come to be principally used in semiprofessional applications or where the ability to edit tapes or achieve uninterrupted playing times of longer than 45 minutes is important. Most such tape decks operate at tape speeds of $7\frac{1}{2}$ and $3\frac{3}{4}$ ips, although speeds of 15 and $1\frac{7}{8}$ ips are also available. These machines use $\frac{1}{4}$-inch tape and usually record in a quarter-track or half-track stereo format. While they were once highly popular, tape "recorders" with built-in amplifiers and speakers are no longer generally available; instead, a tape "deck" is connected to the home amplifier and speaker system.

The cassette. Introduced by Philips of The Netherlands in the mid-1960s as a dictation device for business use, the Compact Cassette (usually abbreviated to cassette) has developed so remarkably into a high-fidelity medium that by the early 1980s sales of prerecorded music cassettes exceeded those of phonograph records.

Construction of the cassette

The cassette consists of a feed reel of tape and take-up hub, plus guides and a pressure pad, enclosed in a sealed, roughly rectangular package. The tape is 0.15 inch (3.8 millimetres) wide and travels at a speed of $1\frac{7}{8}$ ips (4.56 centimetres per second). When recording and being played, the tape begins on the feed reel at one side, passes through guides, and then travels across an opening in the cassette shell, where a pressure pad forces the tape into contact with the record–playback head mounted in the recorder. At another opening a rotating capstan and pinch roller pull the tape toward the take-up reel, onto which it is wound after passing through another set of guides. After play is complete, the cassette can be inverted and played in the opposite direction. Cassettes are generally sold in C-60 and C-90 lengths (30 and 45 minutes per side, respectively), though some longer and shorter lengths are available.

For portable use cassette recorders may be battery-powered and are often equipped with a built-in microphone. Most portables also have a low-power amplifier and loudspeaker, and jacks are provided for the use of headphones for private listening. Both stereophonic and monophonic models are available.

Microcassette recorders, using tape $\frac{1}{8}$-inch wide and moving at a speed of $1\frac{5}{16}$ ips in a package only 5.5 by 3.3 by 0.7 centimetres ($2\frac{3}{16}$ by $1\frac{5}{16}$ by $\frac{9}{32}$ inches), have been developed. Once again the design intention is for business dictation, but many models have been improved for music listening. (C.L.S./L.Kl.)

OPTICAL SYSTEMS

Early efforts to produce motion pictures with sound involved attempts to synchronize phonograph recordings with the images on the screen. One inventor actually tied the phonograph to the projector motor through a system of belts and pulleys that ran the length of the theatre from booth to screen. As theatres grew larger, however, it became apparent that amplification as well as synchronization was a problem.

The amplification problem was solved first with the invention of the Audion amplifier (described above) and its application to loudspeakers. De Forest turned his attention to the problem of synchronization and experimented with recording sound directly onto the motion-picture film. By 1923 he had developed techniques for transcribing the varying impulses of sound waves into varying impulses of light that could be photographed on a strip of film. The developed film, when passed between a light source and a photoelectric cell in the projector, could transform these areas of varying density back into electric voltages that could be amplified and converted into sound by a loudspeaker system. Although there were variations in detail, De Forest's methods were fundamentally duplicated by subsequent sound systems in Germany, the Soviet Union, and the United States.

Stereophonic sound tracks

A fortuitous discovery in optical recording was that distortion in the sound-track area of the film could be lowered by recording the light impulses as two separate but identical sets of patterns read simultaneously by a single photoelectric cell in the projector. Thus, when stereophonic motion pictures were produced, beginning in 1976, the sets of optical tracks could easily be converted into left and right channels. In theatres equipped only for monophonic reproduction the two channels are automatically summed into one signal by the existing photocell, while more modern houses with dual-cell projectors can benefit from stereo sound. The sound tracks for stereophonic releases are also normally encoded using the Dolby noise-reduction system (see below *High-fidelity concepts and systems*). The two stereo tracks are also electrically "matrixed," so that in actuality four different sound channels may be obtained. The musical score is heard in full stereophonic perspective, while dialogue emerges from centre front, and special rear effects (*e.g.,* "surround sound") are time-delayed and produced from speakers along the sides and in the rear of the theatre. The rear channel is also reproduced with reversed phase so that it reinforces the perception of reverberation. While the high-frequency response of monophonic optical recording is typically limited to approximately 7,000 Hz, the greater emphasis on sound quality has led to the development of stereophonic optical recorders and reproducers capable of response to 12 or 13 kHz (kilohertz). (See also MOTION PICTURES.)

AUXILIARY EQUIPMENT

A transducer, a device that converts one form of energy into another, is fundamental to all modern sound-recording and reproduction systems. A microphone converts acoustic energy into electrical energy. A phono cutting head converts electrical energy into mechanical energy to drive the stylus. A tape record head converts electrical energy into magnetic energy. Similarly, in the playback process the mechanical motion of the phono stylus or the magnetic variations presented to the playback head are converted into electrical energy and amplified. At the end of the record–reproduce chain electrical power is converted back into acoustic power by a loudspeaker. Amplifiers, frequency equalizers, compressors, noise-reduction systems, and the like, however, all operate within the electronic realm and so are signal-processing devices, not transducers.

Microphones. When sound waves strike the surface of a microphone diaphragm, they cause it to move correspondingly, and this motion causes changes in some property (current, capacitance, or resistance) of an electrical circuit. The electrical variations produced by the microphone are generally very small (in the order of a few millivolts), but they can be electronically amplified to any desired level suitable for recording, broadcasting, driving loudspeakers in public-address systems or earphones in hearing aids, and other uses. The first successful instrument for converting acoustic power into electric power was the electromagnetic telephone transmitter used by Alexander Graham Bell in 1876. The term microphone was first used in 1878 when the concept of the variable resistance of a loose contact was elucidated. In Bell's device variations in sound pressure alternately compressed or expanded the contact between loose granules of carbon in a carbon telephone transmitter, thereby introducing a variable resistance into the circuit controlling the energy delivered by a battery. As a result the sound waves were converted into corresponding but amplified variations in electric current.

Despite their historical importance carbon and crystal microphones (which utilize pressure on a piezoelectric material to generate an electric voltage) are no longer generally used in sound recording. The most common types used in recording music and speech are the condenser, the dynamic, and the ribbon microphones.

Condenser microphones. Used extensively both for studio recording and for laboratory sound measurements, the condenser microphone consists typically of a thin stretched diaphragm separated by about 0.001 inch (0.025 millimetre) from a parallel fixed plate. The diaphragm acts as one plate of a variable capacitor (a device for storing an electric charge) that is connected in series with a resistor and a direct current source. The motion of the diaphragm in response to sound waves varies the capacitance, producing a varying electric current. Condenser microphones require the use of an external power supply (typically a "polarizing voltage" of more than 48 volts [direct current]) and have a small preamplifier built into the microphone case. Because of their high cost they are generally restricted to professional applications.

In recent years a new type of condenser microphone, the electret condenser, has achieved wide use. In this type of microphone the polarizing voltage is permanently stored on a metalized fluorocarbon foil at the time of manufacture, and small penlight batteries are used inside the microphone case for the preamplifier. Electrets are usually far less expensive than their conventional counterparts, and so have found wider application.

Dynamic microphones. Dynamic microphones are used widely in recording, broadcasting, and public-address systems, especially when ruggedness is essential. In a dynamic microphone a coil of wire is attached to the diaphragm, and the motion of the coil within a fixed magnetic field generates an electrical voltage in the coil. This type of microphone is available at a wide variety of prices, depending on the level of quality required.

Ribbon microphones. In this type of microphone a very thin, lightweight corrugated ribbon is suspended in a fixed magnetic field and exposed to sounds coming from either side. The motion of the ribbon in the field generates an electric voltage. The "figure-eight," or bidirectional, response pattern of the ribbon microphone makes it ideal for broadcast interviews, where its rejection of sounds coming from off-axis is important.

Directional patterns. By proper design a microphone can be made to pick up sound from a single direction only, resulting in what is called a cardioid, or unidirectional, pattern. Such a microphone, usually of either the dynamic or the condenser type, discriminates against reflected and unwanted sounds and noises and is the most frequent choice of soloists in popular recording. Bidirectional microphones, in which the pickup is essentially the same from either front or back but zero from the sides, are also used in an X–Y (crossed) configuration or, pointing to the sides and in conjunction with a front-aimed cardioid microphone, for stereophonic music recording of large ensembles. Microphones with approximately equal pickup from all directions are called nondirectional, or omnidirectional, and generally offer more nearly flat frequency response than do cardioids of equal price.

The dynamic loudspeaker

Loudspeakers and speaker systems. Speakers are classified in much the same way as microphones, and their function is exactly complementary. The most widely used type is the dynamic loudspeaker, which was introduced in 1925. An electrical coil of wire is suspended in a fixed magnetic field provided by a permanent magnet. Sound currents (*i.e.,* the electrical impulses into which sound waves have been transformed) flow through the coil. These currents produce a magnetic field that interacts with the fixed magnetic field, causing the coil to move. A cone-shaped diaphragm fastened to the coil alternately pushes and pulls the air in front of it, creating sound waves. Dynamic speakers are normally mounted in an enclosure or against a large baffle to prevent the air compressed by the front surface of the diaphragm from simply circulating around the edge of the speaker to fill the rarefaction created at the back surface, thus neutralizing the acoustic output. This is a particular problem at the low frequencies, where the cone moves back and forth relatively slowly. The performance of a dynamic loudspeaker depends heavily on the enclosure in which it is mounted. An enclosure with a port (an opening to the outside) precisely dimensioned to match the acoustic characteristics of the speaker mounted in it enhances efficiency and bass response, though some find the bass quality "boomy" and unnatural. An acoustic-suspension speaker system is one in which the volume of air within a sealed cabinet, rather than a conventional mechanical suspension, supplies much of the restoring force to centre the cone after its excursions. This type of design is particularly popular in home sound systems because of its relatively small size and smooth bass response.

A single dynamic loudspeaker cannot fully cover the entire frequency range of recorded sound, so it is customary to divide the frequency spectrum into parts that are reproduced by appropriately designed speakers. The low-frequency speaker is called a "woofer" and the high-frequency speaker a "tweeter." In many systems a third, or midrange, speaker is also employed, and in a few there are separate "subwoofers" and "supertweeters" to reproduce the extremities of the audible spectrum.

The electrostatic loudspeaker

The counterpart of the condenser microphone is the electrostatic loudspeaker. In this type of speaker a thin metalized foil is suspended between two acoustically transparent, fixed electrodes and is polarized by a high ac voltage. A step-up transformer feeds the audio signal to the fixed electrodes in a push-pull arrangement. In this way the movable foil diaphragm is alternately attracted and repelled, and its motions are transferred to the air, creating sound waves. Electrostatic loudspeakers are exceptionally well suited for the reproduction of treble frequencies, for their moving mass (in comparison with the coil and cone of a dynamic speaker) is exceedingly small. They are generally less successful in reproducing high-volume bass frequencies, however, as their limit of excursion is small. For this reason they are usually produced as very large panels, permitting a relatively large air mass to be moved by small diaphragm motions. Because of their power requirements and their capacitive nature, electrostatic loudspeakers often require special amplifier circuitry.

Amplifiers. Amplifiers are used at many different points and for various purposes throughout the sound-recording and reproduction process. The output levels of such transducers as microphones, phono cartridges, and tape heads are very small and must be amplified with low-noise, high-gain amplifier circuits. In addition, phono and tape transducers must be "equalized" to obtain overall uniform response in conformity with the standard frequency curves shown in Figure 12, and this process requires an amplifier stage. The midrotation setting of a volume control in an audio system involves inserting a 20-dB (decibel) loss in signal level, and an amplifier following the control is required to restore the loss. Bass and treble controls require still further amplifiers. While all the foregoing amplifier stages operate at low power levels, the signals necessary to drive a loudspeaker system may require very high power, requiring still further amplification.

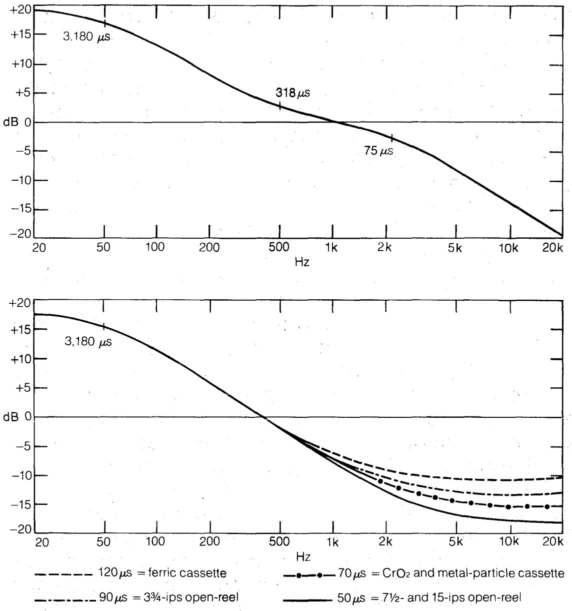


Figure 12: The playback equalization curves for (top) phono and (bottom) tape in the cassette and open-reel formats. Measured in microseconds (µs), playback equalization adjusts both low and high frequencies to obtain overall uniform response.

Amplifiers are not perfect devices. In addition to performing their intended function they also create distortion products, and when it is necessary to link together many amplifier stages, distortion can rise to very significant levels. The principal technique (apart from proper circuit design) used to reduce amplifier distortion is called negative feedback. In a negative-feedback amplifier part of the output signal from the amplifier is fed back into its input, but with reversed phase. Through the use of negative feedback the percentage of distortion can be reduced to negligible proportions, though when used excessively, negative feedback can itself be a source of distortion.

From the days of De Forest's Audion until the early 1960s the vacuum tube provided the only practical means of audio amplification. During the 1960s, however, the transistor gradually took over, and by the early 1970s integrated circuits, each containing a dozen or more transistors, began to be widely used. Integrated circuits are now used not only as amplifiers but also to contain whole circuit sections, such as the noise-reduction circuitry found in nearly all cassette tape decks.

Transistors and integrated circuits

In broadcasting and recording studios amplifier circuits are also used for other forms of signal processing. If the dynamic range (the loud-to-soft ratio) of a sound source is too great to record on a tape or disc or to broadcast, a compressor amplifier is used to reduce the range to a level that can be accommodated. If it is desired that certain portions of the audio spectrum be augmented or cut back, a multiband equalizer amplifier is employed. Popular recording techniques usually involve taping in a "dry" (nonreverberant) studio, so synthetic reverberation devices are used to add a feeling of spaciousness to the sound. Such signal processors all involve some actual or possible loss of fidelity, but they are often considered necessary.

HIGH-FIDELITY CONCEPTS AND SYSTEMS

From the mid-1930s to the mid-1950s the typical home music system consisted of a nicely finished console containing a record changer, AM radio, loudspeaker, and amplifier, augmented in some later models by an FM radio and a television set. For most people music was a form of relaxing entertainment, not a driving passion, and if the sound was pleasant it was good enough.

But a few approached home music reproduction with all the manic intensity of a Toscanini conducting Beethoven's *Ninth Symphony.* They wanted to hear Artur Rubinstein as if he were actually playing in their own living rooms. The LP disc, the FM broadcast, and the rapidly developing art of tape recording were newly at hand and had a music-reproducing potential that was mocked by the poor frequency response, low power, and high distortion of the department-store consoles, whose very brand names became derisive epithets. The enthusiasts of the early 1950s who launched the "hi-fi revolution" did not simply want music to sound good; it had to sound right. This commitment to sonic accuracy is the commitment to high fidelity.

Since the sound waves generated by the original musical source go through a multitude of transformations before they emerge from a loudspeaker, a systematic approach demands examining each link in the reproducer chain. If one manufacturer makes the best phonograph turntable, another the best amplifier, and a third the best FM tuner, it makes sense to interconnect these specific components, however unaesthetic may be their visual appearance together. Moreover, particularly in the early days of high fidelity, improvements were rapid, so it was desirable to be able to change one link in the reproducing chain without disturbing the rest. This dictated a component approach to assembling a home music system. Typically, too, the early audiophiles had at least some background in electronics, usually from training in electrical engineering or from amateur radio. They were not intimidated by, but rather welcomed, the idea of using a soldering iron and a schematic diagram to modify and improve upon a commercially built design.

Component systems

Fortunately, no one today need have the technical orientation of these early hobbyists. The overall quality of high fidelity components (as opposed to prepackaged systems) from established manufacturers in the field has risen to a point where major sonic differences are only likely to be heard between loudspeakers or between systems costing considerably different amounts.

Absolute sonic accuracy would demand that the sound waves arriving at one's ears in the listening room be a facsimile of those heard at a live music performance (minus, perhaps, the noise of the audience). However perfect the electronic part of a high-fidelity system may be, the goal of facsimile reproduction can never be achieved. This is both because microphones do not "hear" selectively (as the ear does) and because home loudspeakers radiate into a much smaller acoustic space than do the instruments in an actual performance. A living room 19 feet 8 inches long by 9 feet 3 inches wide with an 8-foot ceiling (6 by 2.8 by 2.4 metres) has the same dimensional ratios as Boston's Symphony Hall, on a 1:8 scale. For a centrally located listener in Symphony Hall the first reverberant echo from the rear wall arrives 140 milliseconds after the direct sound is heard; in the living room it arrives in only 17.5 milliseconds and can bounce back and forth seven additional times before the first echo is heard in the concert hall. The multiplicity of closely spaced reverberant echoes experienced in home music reproduction seems to give rise to an exaggerated sense of loudness that has been called "echo clutter." With foresight in microphone placement and various tricks (see below) to fool the ear (as the eye is fooled into seeing continuous motion from a series of still pictures in a motion picture) it is possible to create a very plausible illusion of "concert-hall sound" in the home. To do even this, however, demands much of the high-fidelity components.

The requirements for realism. To achieve realism in a sound-reproduction system, at least four fundamental conditions must be satisfied as follows:

First, the frequency range must be such as to include without frequency discrimination all the audible components of the various sounds to be reproduced. Conventionally, the human hearing system is said to extend from approximately 20 Hz to 20,000 Hz, although the ability of the ear to respond to the very highest frequencies tends to diminish with age, particularly among males and those who have had long exposure to high-intensity sounds (*e.g.,* members of a rock group). FM broadcasts are limited to the range from 30 Hz to 15,000 Hz, and a modern FM tuner should be able to cover this range in monophonic or stereophonic modes within ±1 or 2 dB. Very few discs and tapes contain much energy in the lowest musical octave (16 Hz to 32 Hz), and only a few loudspeakers can produce much undistorted output below 40 Hz. At the high-frequency end of the spectrum the weakest link is still the cassette. While many decks can claim a frequency response (±3 dB) to 18,000 Hz and a few go to 20,000 Hz or slightly beyond, the response measurement in this case is made at a low (−20 dB) signal level. At a 0-dB recording level even premium-quality ferric and CrO_2-type cassettes begin to reach tape saturation at about 7,000 Hz; with metal-particle cassettes this high-frequency saturation point is extended nearly an octave higher (14,000 Hz). (It should be noted that, in assessing the claim of any component to reproduce all frequencies from x to y Hz, if no qualifying $\pm z$ dB tolerance is stated the claim is without meaning.)

Frequency range

Second, the volume range must be such as to permit subjectively noiseless and distortionless reproduction of the entire range of sound intensities. A full symphony orchestra or an amplified rock group can produce sound intensities that measure 100 dB or a little higher. Possibly owing to the echo-clutter phenomenon (see above), music reproduced at this level in the home sounds much louder than live music; even during the loudest fortissimo passages in a concert hall one can make a comment to another person nearby without shouting, and this cannot be done in a living room when reproduced music is at so high a measured level. In one's living room the limit of tolerable loudness is probably in the area of 90 dB, with some variations among individuals. With high-powered amplifiers (about 100 watts per channel) and moderately efficient speakers this can be achieved, and the signal-to-noise ratio of a high-quality power amplifier today is adequate to the task.

Dynamic range

The same cannot be said for the sound sources typically available to the audiophile. Few FM broadcasts utilize a dynamic range of more than 50 dB, and if the music exceeds this it is compressed or (worse) limited. New record-cutting techniques such as direct metal mastering (an innovation of Teldec) and direct-to-disc cutting raise the usable dynamic range on discs to perhaps 70 dB or more (justifying turntable manufacturers in reducing their rumble ratings to better that figure). Studio master tapes can be made with the professional Dolby-A noise-reduction system to approach 80 dB, and the consumer Dolby-C system can raise measured signal-to-noise figures for cassettes to about 70 or 72 dB. (A different noise-reduction system, dbx, can achieve a dynamic range of nearly 100 dB, but its incompatibility with the widely available Dolby-B and Dolby-C systems has tended to limit its availability.) The solution to obtaining signal sources adequate to full musical demands probably awaits the wider dissemination of digital recordings (see below *Digital versus analogue recording*).

Third, the spatial sound pattern of the original sound should be preserved in the reproduced sound. The early high-fidelity systems of the 1950s were limited to monophonic reproduction and so could not even begin to project a sense of spatial perspective. Since the arrival of stereo—when careful microphone placement is used—the same order of spatial separation in the reproduced musical instruments and voices can be obtained as in the case of firsthand listening to the live sound sources. Moreover, image-enhancement devices have become available that can yield an astonishing degree of perceived realism in both breadth and depth.

Spatial perspecti

Fourth, the reverberation envelope, or acoustic ambience, of the original sound should be approximated in the reproduced sound. It was expected that the development of quadraphonic sound systems in the early 1970s would solve the final problem in accurate home music reproduction. By feeding separate rear- (or side-) mounted speakers with the reverberant information provided by the hall, it was hoped that the feeling of constricted space often experienced in the listening room could be overcome.

Reverberant sound

For many reasons, including commercial bickering over which matrix four-channel system to adopt for disc recordings and the unwillingness of Philips (the licenser of all cassette machines) to permit four-channel cassettes that could be played on only one side, the quadraphonic idea never captured popular attention, and the medium became dormant. Too few people felt that the benefit was worth the additional cost involved, and demonstrations of the systems tended to be anything but convincing. No one seeking realistic reproduction of music wants, for example, to listen to a string quartet surrounded by the four players.

More promising, however, are reverberation amplifiers that use digital circuitry and time-delay techniques to make the dominant component of the reverberant sound more nearly approximate the ambience invariably present in the concert hall. With good stereo microphone techniques and one of the more advanced "dimensional synthesizers" a very plausible sense of acoustic space can be obtained.

Digital versus analogue recording. The sound-recording and reproducing systems discussed above are all analogue systems. That is to say, they attempt to preserve a complete correspondence between the continuous variations of the sound waves and the methods by which they are stored. The electrical signals from the microphone(s), the undulations in the record groove, the magnetic patterns recorded on tape, and the optical patterns of a film sound track all are types of replicas or pictures, *i.e.,* analogies, of the original sounds. The closer this correspondence is maintained, the greater the fidelity. Since practical devices and storage media are never perfect, absolute fidelity can never be achieved, but in an analogue system closer and closer approximations to the continuously changing sound waves are always possible.

In digital recording, by contrast, no attempt is made to copy all of the infinite number of variations present in even a simple audio tone. Instead, sound waves are sampled at defined time intervals, and the value of each sample is converted into a binary (base-2) number that is recorded on tape as a series of pulses. Each pulse may, for example, be assigned a value of 1, and the absence of a pulse in the sampling sequence be given the value of 0, the two quantities universally used by computers. The 1s and 0s, known as "bits," are grouped into digital "words" to permit larger numbers to be represented. With a 4-bit digital word there are 16 possible combinations, from 0000 to 1111; in general, the highest number that can be represented is 2^n, where n is the number of bits per word. In digital audio recording 16-bit words are used for professional mastering; for consumer applications 14-bit words are nominally standard, although some pressure continues for a 16-bit home format.

Digital sampling

The highest numbers that can be represented in the

16- and 14-bit word formats are 65,536 and 16,384. If each number is used to represent the electrical voltage, or current level, in each sample of the sound wave, the available dynamic range of the professional system is 96 dB and that of the consumer digital system is 84 dB. Both of these ranges far exceed the capabilities of the best available analogue recording and playback machines. Distortion is correspondingly lowered to 0.0015 or 0.006 percent, respectively, again far superior to the performance obtainable in analogue recording.

The frequency-response capability of a digital audio system is determined by the number of samples (the 14- or 16-bit words) taken each second. It can be shown mathematically that any waveform, such as a sound wave, can be reconstructed from samples if the sampling rate is at least twice as high as the highest frequency component in the waveform. To process fully all frequencies in the nominal (20 Hz to 20,000 Hz) audio spectrum thus requires a sampling rate in excess of 40,000 Hz; typically, 44.1 kHz is used for consumer digital audio and 50 kHz for professional recording.)

Figure 13 shows the principal stages involved in a typical digital audio recording. The electrical output from the microphone(s) is amplified to a nominal "line level" of approximately one volt (still in the analogue domain) and passed through a sharp filter that rejects any components above 20,000 Hz. (Without this anti-aliasing filter ultrasonic signals higher than half the sampling frequency can appear within the desired audio spectrum as musically unrelated "aliases" of themselves.) The signal is then processed by an analogue-to-digital (A/D) converter. (Figure 13 shows this "quantization" as a staircase-shaped waveform. In actuality, the various amplitudes shown by the heights of the samples are recorded at the same level, with the pattern of 1s and 0s representing the proper binary number.) After various additional bits are added to correct for any loss of information that may occur in recording or playback from imperfections in the physical tape, the entire pulse code is recorded, or "written," at saturation level onto a tape and stored.

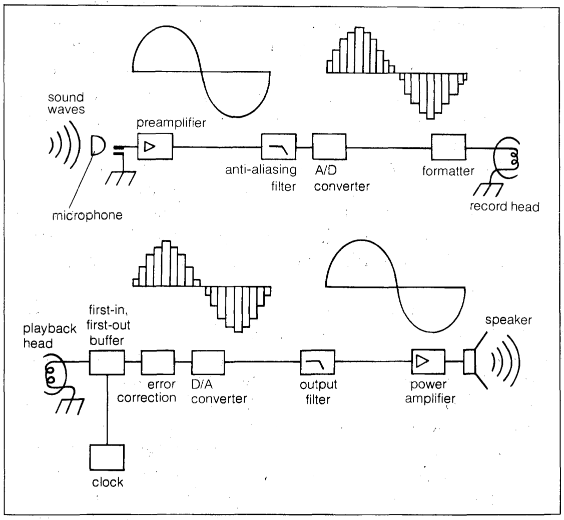


Figure 13: The analogue-to-digital-to-analogue process.

In playback the detected pulses are fed into a "first-in, first-out" buffer memory and are then released at precisely clocked intervals for subsequent decoding. This eliminates the effects of any mechanical wow and flutter, which, to some degree, always degrade the performance of analogue recording and playback systems. After all of the error-correcting procedures have been implemented, the original 14- or 16-bit words are passed to a digital-to-analogue (D/A) converter, whose output passes through a second sharp filter that rejects all signals above 20,000 Hz. In this way the ultrasonic components generated by switching from one bit level to the next are eliminated. A conventional analogue power amplifier and loudspeaker are then used to transform the electrical signal back into sound waves.

By the 1970s a precursor of digital audio had already made its appearance in home music-reproduction systems in the form of digitally mastered recordings converted into conventional, widely acclaimed analogue discs. By the early 1980s more and more master tapes of music performances, particularly in the classical field, were being recorded digitally, even when their public release was in the form of conventional analogue discs and tapes. Further, although intended for a limited market, several companies produced PCM (pulse-code modulation) adaptors for home videocassette decks that recorded and played back audio digitally, and prototypes of digital audio cassette decks were demonstrated. Devices for digital editing, equalizing, and other techniques used in recording studios also became available.

Digital discs

By 1983 fully digital discs and players had been introduced as consumer products in Japan, Europe, and the United States. Developed jointly by Sony (Japan) and Philips (The Netherlands), the Compact Disc (CD), also known as the DAD (Digital Audio Disc), holds approximately 60 minutes of stereophonic, digitally encoded music on one side of a 12-centimetre (4¾-inch) disc. A 16-bit word format and 44.1-kHz sampling rate is used, and the digital information on the disc is detected optically for playback by a laser beam. Technically, these are the finest source of prerecorded sound ever introduced into the home market, for they have an available dynamic range of 90 dB, distortion of less than 0.004 percent, and a frequency response from 5 Hz to 20,000 Hz ±0.5 dB. In one form or another digital technology will play an increasingly large part in all aspects of sound recording and reproduction.

(C.L.S.)

Other applications of acoustics

ARCHITECTURAL ACOUSTICS

The production and reception of sound in closed spaces such as lecture and concert halls, theatres, church auditoriums, school classrooms, and home constitute the subject of architectural acoustics.

The sabin unit

When a sound source is excited in a closed room, the resultant sound intensity at any point depends not solely on the sound waves reaching that point directly from the source but also on the arrival there of sound that has been reflected from the walls and other surfaces of the room. For rooms of the size encountered in ordinary experience the sound intensity builds up rather quickly, and the phenomenon known as reverberation ensues. If the room surfaces were perfect reflectors and the source were to continue to emit sound at a constant power output, the sound intensity would increase indefinitely, except for the attenuation of sound caused by the transmission through the air in the room. Actually, sound is always absorbed to a certain extent by any surface, and this keeps the intensity from building up indefinitely. But if the source stops it takes time for the intensity to decrease so as to render the sound inaudible. The time taken for the average sound intensity in a room to diminish to one-millionth of its original value at the instant the source was turned off is called the reverberation time. It is found experimentally that the reverberation time (*T*) depends directly on the volume of the room and inversely on the absorption of the sound by its various surfaces. If the fractional amount of sound energy absorbed by a surface is called the absorbing power *a*, the total amount of absorption is the integrated value of *a* over all the surfaces of the room. Thus, if 30 percent of the sound energy incident on a particular surface is reflected back into the room, the value of *a* for that surface is 0.7. Each square metre thus acts as the equivalent of 0.7 square metre of a perfect absorber. The metric sabin (after Wallace Clement Sabine, a United States physicist) is the equivalent of one square metre of perfectly absorbing material. The corresponding English unit uses the square foot and is known simply as the sabin. If *a* is the total absorbing power of all the surfaces of the room, then integrating the fractional amount of sound energy absorbed over all the room surface,one gets $a = \int \bar{a}dS = \bar{a}S$, in which dS is the element of surface and S the total room surface. The average absorbing power per unit is $\bar{a}$. The Sabine reverberation law may then be expressed as $T=0.161v/\bar{a}S$, in which T is the reverberation time in seconds, v is the volume of the room in cubic metres, and $\bar{a}S$ is the absorption in metric sabins. The same formula applies if the volume is in cubic feet and the product $\bar{a}S$ is in sabins, though the factor 0.161 then becomes 0.049. Various modifications of the Sabine formula have been proposed to meet the situation posed by extremely high surface absorption ("dead" rooms), and to take account of the absorption in the air of the room.

The optimum acoustical conditions in a room depend on its intended use. For satisfactory reception of speech, a relatively short reverberation time is obviously essential for adequate articulation. With long reverberation time successive utterances overlap, unless the speech is abnormally slow. For auditoriums with volume of the order of 6×10^3 cubic metres to 14×10^3 cubic metres (2×10^5 to 5×10^5 cubic feet), experiments suggest a maximum reverberation time of about 2.5 seconds for satisfactory articulation. For best listening conditions in such rooms the reverberation time should be about one second. These figures indeed assume a low general noise level—*i.e.,* only about 30 decibels above minimum audible. In large rooms (of the order of 3×10^4 cubic metres or more) no speaker's voice can be heard without the use of a public address system, no matter how short the reverberation time. A room used primarily for music requires a longer reverberation time than one used for speech. For concert halls the optimum reverberation time for music is about 1.9 seconds (see Figure 14). Inasmuch as absorption varies more or less directly with the frequency, the reverberation time varies inversely therewith. The above figures are for an average frequency of about 500 hertz.

Sabine's law indicates that the reverberation time for a room of given size can be controlled by appropriate changes in the absorption by the surfaces. Many kinds of acoustic absorbing materials exist to meet a wide variety of needs, both for permanent installation to meet the original design requirements and for portable installations to provide for flexibility in use.

In addition to the overall reverberation, and indeed as a contributing factor to it, the transmission of the sound back and forth between the walls, floor, and ceiling of a room stimulates the development of the normal modes mentioned above. Hence a complicated sound pattern can result in which certain frequencies are accentuated, leading to discomfort and difficulty in hearing. The excitation of the most troublesome normal modes can be materially reduced by the use of absorbing material, and also by introducing slight irregularities in the walls, so as to get rid of parallelism. These precautions break up the regular sound pattern and serve to diffuse the sound energy more equally throughout the space.

Modern architectural acoustics technology uses many special techniques, including the free use of sound-reinforcing systems with loudspeakers placed at strategic points throughout the space. The optimum acoustic design of auditoriums as well as schoolrooms, business offices, residences, etc., reached a high level of efficiency in the 1960s and 1970s. Much of the success of modern room acoustic design is due to the use of computers. A simulated model of a given hall can be introduced into a computer program that will then yield a faithful representation of the acoustical properties of the actual hall. Thus, the modifications necessary to give optimum conditions are relatively easy to determine while the hall is still in the design stage. Acoustic models of this kind will undoubtedly play a role of increasing importance in room acoustics and indeed in applied acoustics generally.

Computers used in acoustic design

MUSICAL ACOUSTICS

Until relatively recent times, acoustical theory had little bearing on musical instruments, even though historically much acoustical theory was developed in the attempt to explain how such instruments work. With the advent of sound recording and reproduction, the attempt to apply acoustical principles to the design of more effective instruments has led to considerable progress.

Conventional musical instruments are complicated examples of coupled vibrating systems—*i.e.,* systems with many degrees of freedom. Many of them, such as horns and woodwinds, function as acoustic filters, emitting only certain frequencies or combinations thereof. The development of electroacoustics has encouraged the invention of many types of electrical musical instruments, such as electronic organs and related devices. Computer-composed music has also attracted considerable attention. (For details see MUSICAL INSTRUMENTS.)

MILITARY APPLICATIONS

The principal military use of acoustics has been the detection of distant objects on the ground or under water. If an impulsive sound wave from a distant source like a gun is received by at least three microphones placed at precisely known intervals along a line (more than three are used in practice), knowledge of the sound velocity in air and the differences in time of arrival of the sound signal at the various microphones is sufficient to yield the bearing of the source and its distance from the microphone range. This is the basis for sound ranging in air. There are, to be sure, many complicating factors, such as the variation of sound velocity due to temperature gradients and wind velocity. Hence meteorological measurements are always important in sound ranging.

Seawater is in general a very good medium for the transmission of sound radiation. Moreover, its acoustic impedance (see the section above on properties of sound waves) matches solid sound sources much better than that of air. Hence sound ranging under water has become a very effective method for the detection of underwater objects such as submarines. Passive detection based on the sound emitted by an underwater vehicle is fundamentally similar to sound ranging in air. Active detection, called sonar, is the more common method. In this, a beam of sound pulses sweeps the water horizontally; when it strikes a solid object, an echo is returned, providing a bearing on the object. The distance may be found from the knowledge of the sound velocity in water. This method is the acoustical analogue of radar. To secure a reasonably sharp beam from a not-too-large transducer surface (employing

By courtesy of Bolt Beranek and Newman Inc.

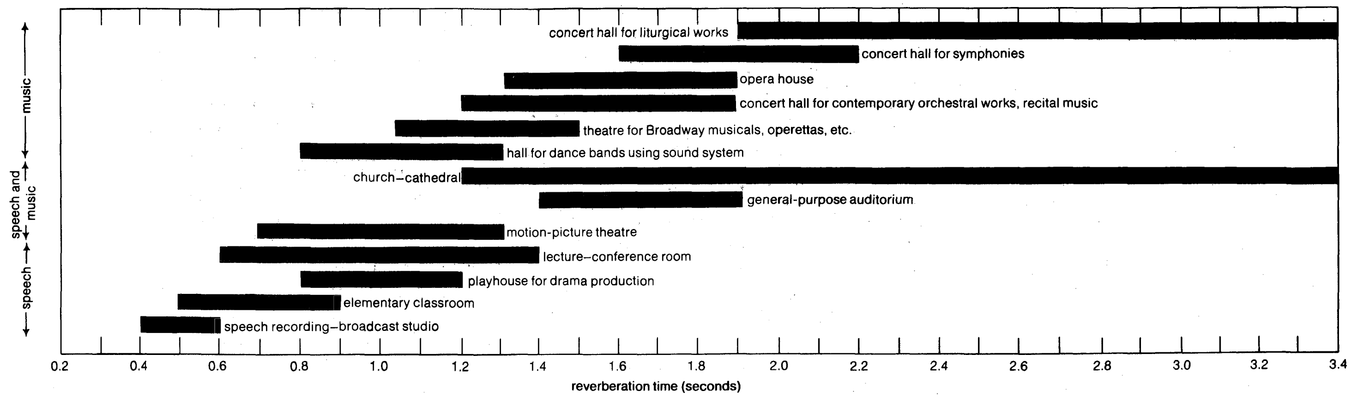


Figure 14: Optimum reverberation time for various types of facilities.

piezoelectric or magnetostrictive elements), a fairly high frequency must be used. On the other hand, because the absorption of sound in seawater increases more or less with the square of the frequency, there are limitations involved. It is usual to employ rather larger transducers with lower frequencies to secure long range. The use of sonar demands many correction techniques to take care of temperature gradients, changes in density, reflection from surface and bottom, and other error-producing effects.

Sonar has become increasingly important not only as a military weapon but as a navigational device and a means for exploring the properties of the sea.

NOISE CONTROL

Noise or unwanted auditory experience was an important aspect of practical acoustics in the 1970s. Noise in human life had reached the level at which it was considered a form of environmental pollution. Noise control then developed into a major branch of acoustical engineering.

Every advance in the efficiency of the transformation of heat energy into mechanical energy has involved a noise problem, and in general this increases with power production. The large-scale turbulence of exhaust gases streaming from a jet airplane is an unusually intense source of sound in which the total acoustic power is proportional to something like the fifth power of the jetstream velocity. The annoyance of this is evident to anyone who lives in the vicinity of an airport. When a plane moves faster than the speed of sound in air, the bow wave formed is shed as a shock wave (analogous to Cherenkov radiation in light), and this leads to the so-called sonic boom, a loud explosion as the shock wave hits the ground. This is one of the serious noise hazards of the supersonic jet plane.

(R.B.L.)

Criteria for noise-control design. The amount of sound energy that reaches the listener depends on the paths by which the sound has travelled, on the distance from the source, the power of the source, and the nature of such intervening barriers as walls, doors, windows, and sound-absorbing ducts. Qualities that characterize a desirable acoustic environment vary widely, depending on how the space is to be used, how particular the users are, and how the specific space involved relates to other parts of the building. A library reading room, for example, should be free of distraction. This objective can be achieved either by keeping the reading room silent or by providing a moderate, continuous, background sound level of unobtrusive, unrecognizable character, sometimes called white noise. Because the background sound hides or masks the minor noises that are unavoidable, the second approach is considered more realistic. In a large business office, a higher degree of background noise might be acceptable, up to the point at which it interferes with conferences and telephone conversations. In general, people usually tolerate a noise conveying no information better than they do one that tells them something, and an expected noise is usually more tolerable than an unexpected one of the same magnitude.

olerable
oise levels

The acoustical character of a space is as important as the amount of sound generated in it or transmitted into it. If a room is finished in hard, sound-reflecting materials, sounds in the room not only persist but seem to come from all directions.

There are optimum ranges for the reverberant character of occupied spaces. An aid to the acoustical engineer is the fact that the acoustical character does not need to be uniform throughout a building; on the contrary, variations may provide a pleasant contrast, from a large, reverberant, monumental space to a smaller, quieter, more intimate space in which communication is easy.

Controlling noise outdoors. The most effective method of noise control outdoors is to separate the source and receiver in space, but because this is not always feasible, some type of barrier is often introduced. Walls and earth banks can be used to control many types of outdoor noise, but the effectiveness of any barrier depends on how well it cuts off the line-of-sight path between source and receiver. A wall to control traffic noise, for example, can be effective at ground level, provided it is near either the source or the receiver, but a low wall constructed in front of a multistory apartment house will do nothing to control noise for those above the level of the wall. Trees and shrubs have little effect.

Controlling noise indoors. When a sound wave strikes a sizable obstacle such as a wall or ceiling, part of it is reflected, part of it is absorbed, and part of it is transmitted to some adjoining space. The relative magnitudes of these three parts of the original sound are determined by the physical properties of the obstacle. A hard-surfaced, dense plaster reflects a great deal of the sound that strikes it, while a soft surface such as a heavy carpet absorbs most sound and reflects little. Absorption and reflection are important only in connection with the space in which the sound originates. More often the problem is the portion of the sound that is transmitted to an adjoining space. Although the mechanisms governing sound absorption and sound isolation are complex, they must be understood in order to predict the resulting acoustical environment in a finished building.

Sound absorption. Useful sound absorption is provided by porous materials such as carpets, draperies, glass-fibre blankets, clothing, and the many specially made, sound-absorbing materials. An essential property of a sound-absorbing material is that it have a porous structure into which the molecules in the air carrying sound energy can dissipate, thus losing energy in the form of heat. Sound energy so dissipated is never recovered as sound. Because it is essential that the interstices of sound-absorbing materials be intercommunicating, a closed-cell foam material is not a good sound absorber. Porous sound-absorbing materials may sometimes be faced with hard-surfaced, refinishable perforated panels, wood strips, or the like, but these facings must be so designed that the sound can penetrate into the porous material behind the facings. In general, the solid parts of the facing material should not be larger than two or three inches (five to eight centimetres) and the openings between these elements to the sound-absorbing material should represent 10–25 percent of the area.

Sound-absorbing materials

When sound-absorbing materials are placed within a room, the reflection of energy is greatly reduced and the sound dies away rapidly instead of being reflected many times by the enclosing surfaces. When the sound dies away quickly, the room has what is called a short reverberation time. If surfaces are highly reflective and sound continues to reflect from surface to surface with almost no loss, the room has a long reverberation time. The "feel" or quality of a space as sensed by the occupants is determined largely by the amount and type of sound-absorbing finish.

When a room has a short reverberation time, sounds seem to come from their point of origin rather than from everywhere, surrounding the listener with many reflections. There is actually some reduction in the level of sound from a source in such a "dead" room when contrasted with a hard, reverberant, sound-reflecting "live" room. A nonreverberant space seems more comfortable, and speech communication is easier because the jumble of successive sounds is absent. The proper use of sound-absorbing finishes can improve the quality of almost any space in which a number of people must work, live, or play together. Large open-plan offices and schools work satisfactorily only when both the floor and ceiling are covered with a sound-absorbing finish.

Dead and live rooms

Sound-absorbing ceilings can be very useful for noise control in factories, but the source of the noise must also be considered. Since a sound-absorbing ceiling affects only the noise level in the overall space, it has no effect on the level of sound reaching the ear of a worker near a noisy machine. In areas where noise is less intense, as in restaurants and college dining halls, a sound-absorbing ceiling is more effective. The potential of sound-absorbing materials in private residences has been largely neglected. A sound-absorbing ceiling in a dining room, for example, can reduce the din at a dinner party (whether the room is carpeted or not) to a pleasant level that permits conversation at ordinary voice level. The kitchen, with its customarily hard finishes, can also be a more pleasant space in which to work if it has a sound-absorbing ceiling.

Sound-absorbing materials are used as linings for air ducts to reduce the transmission of sound from room to room and from the fan. Also, mufflers for internal combustion engines incorporate either resonant or dissipative sound-absorbing materials to reduce the noise transmitted with the exhaust gases.

Sound isolation. While sound-absorbing treatments are useful in the control of the character of occupied spaces and in reducing the magnitude of noise in an enclosure, they are useless as noise barriers. A curtain hung in a doorway or over a wall provides no improvement in sound isolation. Acoustic tiles nailed to the underside of floor joints do not improve the isolation between the living room and a downstairs playroom.

The sound isolation provided by a barrier such as a wall, floor, ceiling, or window, is determined almost entirely by its weight or inertia. Since a sound wave consists of the back-and-forth oscillating motion of molecules in the air, these moving particles, when they encounter an obstacle, tend to make the obstacle oscillate slightly. If the obstacle has a great deal of weight or inertia, it resists this attempt at being moved, but if it is relatively lightweight it may move quite easily. When the obstacle is moved by action of the sound wave on one side, it radiates a new sound wave on the other side. This new sound wave is the transmitted sound, the sound heard in the adjoining space. The sound energy is transmitted from one side of the partition to the other by the direct motion of the partition. The heavier the construction, the less sound gets through. A great deal less sound can be heard through an eight-inch brick wall, for example, than through a lightweight plasterboard partition.

Any construction must be completely airtight if it is to be useful as a sound isolator. An opening in a barrier (such as an open door) transmits much more sound energy than an area of wall of equal size. A wall of porous concrete block or construction with loose joints allows sound energy to be transmitted through the leaks and cracks. Doors and windows work at their best as sound barriers only when they are tightly weatherstripped. Probably the most difficult problem in modern building construction is to make airtight connections between the many dry, prefabricated elements that form the enclosure and subdivision of buildings. Every service penetration for piping, every shrinkage crack, and, in fact, any leaks at all seriously limit the ultimate performance of noise-isolating construction.

Sound-isolating construction

Since it is often impractical to use extremely heavy construction for walls and partitions in buildings, two separated elements, instead of a single heavy unit, are sometimes employed. Two layers of glass, for example, separated by a space of about six inches (15 centimetres), can be as effective as a heavy masonry wall. A double window of this sort is a good isolator provided it is completely weatherstripped. Ordinary sealed double glass, with an air space of 0.5 inch (1.25 centimetres) or less, is not really a double construction from the acoustical point of view, because the air space is too small to permit independent motion of the two pieces of glass. Double plaster walls, using separate framing for each face, are often used for party walls between apartments. This construction can be very effective, but its acoustical effect is seriously impaired by back-to-back convenience outlets or penetrations for heating pipes or plumbing.

Background noise. The amount of acoustical privacy between two spaces is related to the background noise level as well as to the sound isolation provided by the construction. Architects have often been surprised to find that a partition type that worked well in a city office building was far from satisfactory in a quieter location. In such cases, a small amount of white noise may be added to the ambient noise level.

Background noise should be bland, unnoticeable, and continuous. Steady traffic noise of the city or the characteristic sounds of many types of air diffusers are the kinds of background noise that people expect in a typical modern building. When there are no partitions, the background noise becomes even more important in determining privacy. A large, elegant, but sparsely filled dining room is often so quiet that patrons must speak in low tones to preserve privacy, while with a moderate level of conversation throughout an ordinary restaurant, good privacy can be had at comfortable voice levels.

Structure-borne noise and vibrations. Sound is also transmitted in a building when the structure sustains a direct impact or when a vibrating object is in rigid contact with it. The structure then radiates sound energy into the air. For this reason a neighbour's doorbell, if fastened directly to a common wall between apartments, can often be heard as loudly in one apartment as the other. The sound of movements from the floor above is a familiar structure-borne sound, as is the noise from a rigidly attached appliance, or from the great variety of mechanical devices that fill modern buildings. As is the case with airborne sound, the heavier the construction the less sound of impact and vibration is transmitted.

Resilient isolation

Combined with massive construction is the need for resilient isolation. The noise of a footfall overhead is considerably reduced when the floor is covered by heavy carpeting. Footsteps can be made almost inaudible if, during construction, an additional layer of concrete or a wood floor is "floated" on special resilient pads on top of the basic structural floor. A doorbell on a resilient mounting is inaudible in a neighbour's apartment. A garbage grinder, attached to the sink through a rubber collar, is relatively quiet. Water pipes and drains, isolated from walls and floors with resilient materials, perform their functions with little noise. Fans and compressors, mounted so that they are free to vibrate on springs or rubber mountings, create little noise, although in the absence of such mountings, the whole structure acts as a sounding board for these devices.

Not only must moving equipment be properly mounted, but it must also be connected to necessary electrical and plumbing services through flexible connectors. Any rigid ties to the structure of the building nullify the value of resilient mountings.

(R.B.N./Ed.)

APPLICATIONS OF ULTRASONICS

Until about 1910 ultrasonic waves were little more than a scientific curiosity. Following the successful development of the piezoelectric transducer, they were used in an early form of sonar to detect the presence of submerged submarines. Other attempted early applications were as a means of communication, and as light modulators in early experiments with television. Following the development of radar during World War II, ultrasonic techniques were used in such wide-ranging fields as the study of molecular properties of materials, detection of flaws in metals, ultrasonic cleaning, industrial and dental drills, measuring the thickness of the heart walls in man, and determining the presence of fluid in the sac around the heart.

Applications of ultrasonics can be divided roughly into two classes: low energy and high energy. In the former, the amplitude (that is, the height of the wave from its highest point to its lowest point) is sufficiently low so that the wave is not distorted in passing through the medium, and the medium is left unchanged except for a slight rise in temperature. High-energy waves, on the other hand, modify the medium in some irreversible manner, either by the generation of large stresses or high temperatures, which induces physical or chemical changes.

When a low-energy ultrasonic wave travels through a medium, its energy decreases rapidly because of a characteristic of the medium known as its absorption coefficient. Absorption is caused by a medium's viscosity and thermal conductivity, though recently it has been found that the absorption coefficient is often much higher than calculations based on viscosity and thermal conductivity alone would indicate. Much information about the molecular properties of materials has been obtained from studies of the variation of this excess absorption with frequency.

Relaxation effect

Another important factor in dealing with the passage of an ultrasonic wave through a material is the relaxation time, or relaxation effect, a measure of how rapidly the material changes its volume as a result of the pressure changes caused by the passage of the ultrasonic wave. If an appreciable lag takes place between the wave's passage and the completion of deformation, a relaxation effect is said to occur. Whenever the relaxation effect exists, the wave velocity through the material depends on the frequency of the wave.

Propagation of ultrasonic waves. The manner in which ultrasonic waves propagate, or spread, through different media determines their applications. Ultrasonic waves are absorbed in a gas because of its viscosity (resistance to flow) and thermal conductivity (the ability to conduct heat). Relaxation effects, defined above as the measure of how rapidly a material changes its volume as a result of the pressure changes caused by passage of an ultrasonic wave, are also significant. Though liquids absorb for the same reasons as gases, measured values in many cases are greater than would be predicted on the basis of their viscosity and thermal conductivity, indicating that relaxation effects predominate. The thermal conductivity of a solid, like that of liquids and gases, also causes absorption. Thermal conductivity is enhanced in a solid composed of a large number of crystals of microscopic size (microcrystals). In addition, the microcrystals tend to scatter the wave as it passes through. The relation between the fraction of incident energy thus scattered and microcrystal grain sizes is a function of the frequency of the ultrasonic wave, and measurements of scattered energy versus frequency can be employed to determine grain sizes.

mplifi-
ıtion of
ı ultra-
ınic wave

An ultrasonic wave produces strains in a medium as it passes through; if the medium is piezoelectric, the strains generate an electrical field. If the piezoelectric medium is also a semiconductor, the current carriers (if negative) are accelerated by a positive electrical field and decelerated by a negative field. In the presence of the ultrasonic wave the current carriers bunch together and drift through the medium with a velocity dependent upon an applied electrostatic field (voltage). If these bunches drift more slowly than the ultrasonic wave, the latter tends to accelerate the bunches. The energy required for this to happen comes from the ultrasonic wave, which is, therefore, absorbed. If the drift velocity of the bunches exceeds the ultrasonic wave velocity, the opposite occurs and energy flows from the current carriers into the ultrasonic wave, which is therefore amplified. Although amplification factors of 30 have been obtained in cadmium sulfide crystals, there have, as yet, been no practical applications of this amplifying technique.

Lower power applications. *Flaw detection and thickness gauging.* Ultrasonic waves are scattered when they meet an acoustical impedance mismatch—*i.e.,* a point at which the resistance to the passage of ultrasonic waves changes abruptly. This mismatch can occur at a flaw, such as a hole or a crack in a metal casting. The detection of such flaws, vital in metallurgy, can therefore be performed by using an ultrasonic version of radar. In this application, the pulses of the ultrasonic waves are scattered when they strike a flaw in, for example, a metal casting, with some energy returning either to the same or another transducer. By measuring the time required for the ultrasonic wave to pass through the material, be deflected, and return, it is possible to determine the location of the flaw; thickness can be gauged by similar techniques.

Ultrasonic delay line. In certain electronic systems, such as that in the colour separation section of a colour television receiver, it is necessary to delay the passage of an electrical signal from one point to another. This time delay must be precisely controllable, appreciable (that is, greater than a few millionths of a second, or microseconds), and reproducible. One form of delay unit employs two quartz transducers cemented to the sides of a glass polyhedron. Ultrasonic waves emitted by one transducer bounce from side to side of the polyhedron until they reach the other transducer where they are converted back into electrical signals.

trasonic
rsts

Measurement of mechanical stresses. The presence of a stress in a body rotates the plane of polarization of polarized shear waves (a situation in which the particles that propagate the wave are all vibrating in the same direction or plane), and some correlation has been found between the amount of rotation and the magnitude, or physical size, of the stress. When a material is stressed it emits bursts of ultrasonic waves at frequencies up to about 40 kilohertz (40,000 cycles per second) for ductile materials and up to 400 kilohertz for brittle materials. Materials undergoing plastic deformation emit signals of a lower amplitude than when the deformation is such as to produce cracks. The repetition rate of the bursts increases with increasing rates of stain. It is anticipated that the detection of these bursts of ultrasonic waves can provide an early-warning system for incipient mechanical failure.

Ultrasonic image converters. The data presented by the type of flaw detector considered above is essentially one-dimensional; interpretation is difficult and somewhat subjective. For this reason much effort has been devoted to developing an ultrasonic camera equivalent to the X-ray machine.

In the ultrasonic camera a transducer generates an ultrasonic wave in a liquid. This wave passes through the object to be investigated and then strikes the metallized surface of a quartz disk that forms the front plate of a cathode-ray tube. The variation in ultrasonic intensity across the surface of the disk produces a corresponding variation in the alternating charge pattern appearing at the rear surface of the disk. This rear surface is scanned by an electronic beam that converts the charge pattern into a visible image on the face of a television picture tube. The picture on the television tube will then be an ultrasonic "X-ray" of the object.

A hologram is a three-dimensional picture made (without a camera) on photographic film by the pattern of interference formed by laser light reflected from the object; the picture is viewed by passing laser light through the film. In recent years optical holography has appeared as a powerful tool for optical imaging. Some work is being carried out to develop an ultrasonic hologram. In one system, illustrated in Figure 15, two beams of ultrasonic waves at identical frequencies are directed at an angle to the underside of a liquid surface. One of these beams passes through the object to be investigated. At the surface, the two beams interfere to produce an ultrasonic wave pattern containing, effectively, an ultrasonic hologram of the object. This wave pattern is visualized by shining monochromatic (single-frequency) light at an oblique angle onto the liquid surface. The liquid wave pattern behaves like a diffraction grating and the first order diffraction image contains a picture of the ultrasonic cross section. In fact, the light beam is behaving in the same way as the beam used to reconstruct a picture from an optical hologram.

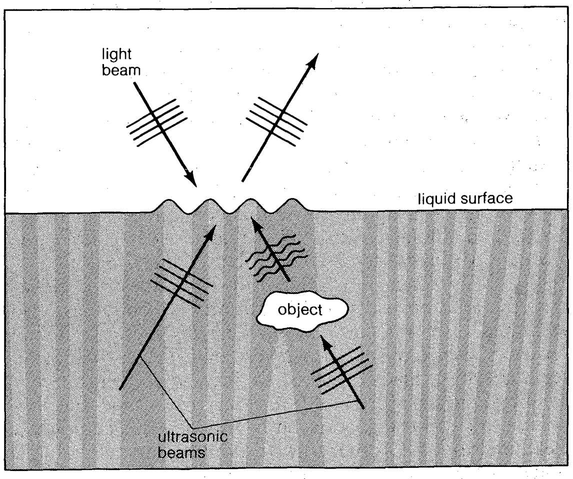


Figure 15: System for production of ultrasonic hologram (see text).

Ultrasonic holography, as so far developed, has a much higher resolving power than the ultrasonic camera.

Underwater applications. Underwater applications of ultrasonic waves are basically flaw-detection systems, in which an object such as a submarine, a shoal of fish, or the seabed acts as the flaw. Ultrasonic waves have been used for short-range underwater communication, but most transducers are so highly directive that their use is of limited value.

Applications in air. The applications of ultrsonic waves in air are similar to those in liquids and solids and are essentially forms of flaw detection. Ultrasonic waves have been used to count objects travelling along a belt, each object breaking the sound beam as if it were an optical beam. In the early 1970s a hand-held sonar blind-guidance device showed promise. In this device an ultrasonic beam is radiated; the presence of an obstacle causes this beam to be reflected, and the reflected beam is indicated by a tone in a headset worn by the operator. The pitch of the tone indicates the distance to the obstacle. Another interesting application is as an intrusion detector, in which a transducer sets up a pattern of ultrasonic waves in a room, producing a specific amplitude at a detector. An intruder alters the ultrasonic wave pattern and therefore alters the ultrasonic amplitude at the detector, setting off an alarm.

Measuring fluid flow rates

Ultrasonic viscometer. The acoustic impedance (resistance to the passage of acoustic waves) of a liquid for transverse or shear waves involves the liquid's viscosity. If the electrical impedance of a shear wave transducer is measured in air (which presents negligible resistance) and then in a liquid, the impedance difference can be related to the liquid's viscosity, enabling the latter to be determined conveniently.

Ultrasonic flowmeter. The apparent frequency of an ultrasonic wave travelling in a medium moving relative to the transducer in a direction parallel to that of the wave, differs from the vibration frequency of the transducer because of the Doppler effect. The frequency shift is directly related to the medium's velocity, and hence, if the frequency shift can be measured, the medium's velocity can be calculated. This principle has been applied to the measurement of fluid flow rates.

Medical applications. The use of X-rays has been well established for many years for medical investigations. X-rays, however, suffer from two defects—it is difficult to delineate soft tissues of the human body with them, and their use is not advisable in the early stages of pregnancy. Ultrasonic waves, however, are reflected by some soft tissues and, as far as is known, cause no harmful effects if the ultrasonic intensity is sufficiently low. In medicine, the main application of ultrasonics has been as a flaw detector in such applications as detecting shifts in the midline of the brain and investigations of the fetus. Fetal echoes are obtainable many weeks before the fetal skeleton is visible by radiography (see also DIAGNOSIS AND THERAPEUTICS).

High-power applications. The effects produced by high-energy ultrasonic waves are normally irreversible and arise from cavitation (see below), intense mechanical stresses, or intense localized heating.

If the ultrasonic pressure exceeds the ambient (normal average) pressure in a liquid, the pressure in the liquid will fall below zero at the ultrasonic wave pressure trough, or point of minimum pressure. If this occurs, a process known as cavitation takes place; the liquid ruptures and forms small cavities, which—if they have a suitable radius—will expand as the pressure rises and then, as the pressure starts to fall, will become unstable and collapse very rapidly. At the end of the collapse, the gas within the cavity will be highly compressed (measurements indicate pressures of several hundreds of atmospheres); these high pressures will be relieved by the radiation of shock waves. These intense shock waves can cause liquids to mix that normally would not do so—such as oil and water; they can break up giant molecules such as polymers, proteins, and viruses; force dirt and grease off surfaces for cleaning purposes; break up cells; and initiate chemical reactions. The violent agitation caused by cavitation has also been used to improve the quality of electrodeposited copperplate and the quality of optical glass by irradiating the glass, when melted, with highpower ultrasonic energy.

The rapid vibration of a gas or liquid under the influence of intense ultrasonic waves causes a net attractive force to develop between particles suspended in the fluid. In addition, small, light particles will take up more of the fluid's motion than large, heavier ones. This difference in motion will increase the chance of collisions between the particles. If the particles coalesce (adhere) on collision, irradiation by ultrasonic waves provides a means of coagulating the suspended particles. This has found applications in the removal of suspended particulate matter, as in the cleaning of factory exhaust gases.

Another useful application of intense ultrasonic waves is the ultrasonic soldering iron. The iron is designed to be a magnetostrictive transducer. The rapid vibration at its tip removes oxide layers from the surface of the metal to be soldered and not only improves the quality of the soldering but also permits the soldering of materials otherwise difficult to solder, such as aluminum.

Ultrasonics in brain surgery

Intense ultrasonic waves have also found an application in medicine, particularly in brain surgery. Four beams of ultrasonic energy are beamed into the brain through holes cut in the skull. The intensity of each beam is insufficient to cause damage, but when combined at some point within the brain the total intensity is sufficient to heat a small region to a high enough temperature to destroy the tissue that the surgeon wants destroyed. This technique has been used successfully in the treatment of Parkinson's disease.

INFRASONICS

Infrasonic waves are those whose frequencies are nominally below the lower limit of human hearing. In recent years, however, evidence has been accumulating that waves with frequencies down to one cycle per second can be detected by human beings, provided the intensity is sufficiently high. The detection process, not yet fully understood, is believed to be a disturbance of the inner ear leading to dizziness. These waves are generated wherever a large high-speed flow of air occurs into, or around, an enclosed box or room. Of particular current interest is their effect on automobile drivers driving with open windows. Some microphones operate down to these frequencies and can be used to detect the presence of infrasonic waves.

Other sources of infrasonic waves of even much lower frequency (sometimes below one cycle per second) are earthquakes and tidal motion. Earthquakes require special detectors that usually consist of a large, well-sprung mass acting as an inertial element, the movement of which relative to the Earth is monitored, either by a pen drawing onto a chart mounted on the Earth or by a light beam falling onto photographically sensitive paper. When the Earth moves, the chart moves with it, but the inertial mass remains stationary and, therefore, the deflection of the trace on the chart indicates the motion of the Earth relative to the stationary mass. For portable use in seismographic surveys, the inertial mass is formed by a well-sprung magnet, between whose poles is situated a coil connected to the Earth. The motion of the Earth relative to the magnet induces a voltage in the coil proportional to the speed of relative motion.

Use in underground exploratio

Underground explosions are used to generate infrasonic waves. From the traces provided by several seismometers of the paths traversed by the waves, the properties of the underlying rocks can be evaluated. This technique is used extensively in mineral and oil surveys. Recently, extremely sensitive seismometers have been employed to monitor infrasonic waves produced by underground nuclear explosions as part of the implementation of the nuclear test ban treaty.

Earthquakes are the main source of naturally occurring infrasonic waves. The main shock is often preceded by smaller shocks; by detecting these preliminary waves it may be possible to develop an early-warning system for large earthquakes. A similar early-warning system has been suggested for volcanic eruptions, which are usually preceded by seismological activity. (G.L.G.)

BIBLIOGRAPHY

General works: LORD RAYLEIGH, *Theory of Sound,* 2 vol. (1877–78; rev. ed., 1926; 1-volume ed., 1945), probably the most famous treatise on sound ever published, has remained an authority on nearly all aspects of theoretical acoustics since its publication. Twentieth-century accounts, particularly recommended, are WALLACE C. SABINE, *Collected Papers on Acoustics* (1922, reprinted 1964), a classic work on the development of the reverberation formula, still in use today, together with a fascinating discussion of solutions to acoustical problems, and a debunking of acoustical myths; F.V. HUNT, *Electroacoustics* (1954), a historical survey of the development of the application of electricity to acoustics; H.F. OLSON, *Acoustical Engineering,*

3rd ed. (1957) and *Music, Physics and Engineering* (1967), two books on various aspects of industrial acoustics, particularly with reference to sound recording and reproduction; R.B. LINDSAY, *Mechanical Radiation* (1960), an advanced treatment of the general principles of acoustical waves; L.E. KINSLER and A.R. FREY, *Fundamentals of Acoustics,* 2nd ed. (1962), a basic text in acoustics for advanced undergraduates; J.N. BRADLEY, *Shock Waves in Chemistry and Physics* (1962), a treatise on the properties and applications of shock waves; R.W.B. STEPHENS and A.E. BATE, *Acoustics and Vibrational Physics,* 2nd ed. (1966), a textbook for advanced undergraduates and graduate students; R.J. URICK, *Principles of Underwater Sound for Engineers* (1967), a practical treatise with emphasis on recent developments; P.M. MORSE and K.U. INGARD, *Theoretical Acoustics* (1968), an advanced and highly mathematical treatise; V.M. ALBERS, *The World of Sound* (1970), modern acoustics, popularly presented without mathematics; ALLAN D. PIERCE, *Acoustics: An Introduction to Its Physical Principles and Applications* (1981), a broad survey; RICHARD E. BERG and DAVID G. STORK, *Physics of Sound* (1982).

Sound recording and reproduction: THE INSTITUTE OF HIGH FIDELITY, *Official Guide to High Fidelity,* 2nd ed. (1978), provides a clearly written discussion of the components of high-fidelity sound systems for the layman. FRITZ WINCKEL, *Music, Sound and Sensation: A Modern Exposition* (1967; originally published in German, 1960), is an extraordinarily perceptive account of the physics and psychoacoustics of music. From a more traditional standpoint many of these topics are discussed by HARRY F. OLSON, *Music, Physics and Engineering,* 2nd ed. (1967). A discussion of some of the problems with facsimile reproduction is contained in CRAIG STARK, "O, O, O That Concert-Hall Realism Rag!," *Stereo Review* (November 1971). For the serious recordist JOHN M. WORAM, *The Recording Studio Handbook* (1976, reprinted with updates 1982), is an excellent manual. The more technical aspects of both analogue and digital recording are covered in CHARLES B. PEAR, JR. (ed.), *Magnetic Recording in Science and Industry* (1967). An insider's view of the development of disc recording and of those who have shaped that industry is contained in CHARLES A. SCHICKE, *Revolution in Sound: A Biography of the Recording Industry* (1974). A standard reference for all aspects of sound recording and reproduction is HOWARD M. TREMAINE, *Audio Cyclopedia,* 2nd ed. (1969, reprinted 1977); JOHN BORWICK (ed.), *Sound Recording Practice: A Handbook,* 2nd ed. (1980).

Applications of acoustics: J. BACKUS, *Acoustical Foundations of Music* (1969), a study of acoustical principles at the basis of music, suitable for students of music; and K.D. KRYTER, *The Effect of Noise on Man* (1970), an important discussion of the modern noise problem; C.M. HARRIS (ed.), *Handbook of Noise Control* (1957), a collection of 40 articles (each with bibliography) on the nature and control of noise in a variety of situations; L.L. BERANEK (ed.), *Noise Reduction* (1960; rev. ed., *Noise and Vibration Control,* 1971), a detailed text and reference work dealing with sound waves and their measurement, principles, criteria, and methods for noise control, and containing an extensive bibliography; *Music, Acoustics and Architecture* (1962), a critical analysis of factors that determine good room acoustics, including detailed drawings, photographs, and discussion of 54 halls in Europe and North and South America; R.B. NEWMAN and W.J. CAVANAUGH, "Acoustics," in J.H. CALLENDER (ed.), *Time-Saver Standards,* 4th ed. (1966), a useful reference covering the basic principles of sound control, data on acoustical properties of common building materials, charts useful in design, and drawings of many details to achieve good acoustics in buildings; W.P. MASON (ed.), *Physical Acoustics,* 6 vol. (1964–70), the most comprehensive survey of the whole field of ultrasonics ever published; G.L. GOOBERMAN, *Ultrasonics Theory and Practice* (1968), a short comprehensive textbook covering most of the theory of ultrasonics with some typical applications; L. BERGMANN, *Der Ultraschall und seine Anwendung in Wissenschaft und Technik* (1954), the classical text (in German) with only an elementary treatment of theory but an encyclopaedic coverage of applications and a complete list of references (two additional lists of references published); T.F. HUETER and R.H. BOLT, *Sonics: Technique for the Use of Sound and Ultrasound in Engineering and Science* (1955), a useful book particularly good on the design of transducers; K.F. HERZFELD and T.A. LITOVITZ, *Absorption and Dispersion of Ultrasonic Waves* (1959), an excellent book although limited to gases and liquids; B. BROWN and J.E. GOODMAN, *High-Intensity Ultrasonics: Industrial Applications* (1965), a good text covering most of the relevant industrial applications of ultrasonics; B. BROWN and D. GORDON (eds.), *Ultrasonic Techniques in Biology and Medicine* (1967), a useful collection of papers dealing with mainly diagnostic applications of ultrasonics in medicine. ARTHUR P. CRACKNELL, *Ultrasonics* (1980), an introduction to the theory and applications. See also HARVEY FEIGENBAUM, *Echocardiography,* 3rd ed. (1981), ultrasonic cardiography. Important journals include the *Journal of the Acoustical Society of America* (1929–); *Acustica* (1951–), the international journal of acoustics; the *Journal of Sound and Vibration* (1964–); and *Ultrasonics* (1963–). *Soviet Physics—Acoustics* (1955–), in English translation, is available through the American Institute of Physics. These journals are mainly devoted to the publication of archival research in acoustics, but most of them contain notes on modern developments of a more popular character as well as news about acoustical scientists and engineers. Useful acoustical data will be found in the chapter on acoustics in the *American Institute of Physics Handbook,* 2nd ed. (1963; 3rd ed., 1972).

South Africa

The Republic of South Africa (Afrikaans, Republiek van Suid-Afrika), the southernmost state on the African continent, has an area of 471,447 square miles (1,221,042 square kilometres). (If the territory of the republics of Transkei, Bophuthatswana, Venda, and Ciskei—all lying within South Africa's borders—is excluded, the area of South Africa is 434,700 square miles.) It measures almost 1,000 miles (1,600 kilometres) from north to south, as well as from east to west; a variety of landscapes, climates, and resources are to be found within its borders.

Of its four provinces, Cape of Good Hope Province, with an area (including the territory of the republics within its borders) of 278,381 square miles, is the largest, followed by the Transvaal (109,622 square miles), the Orange Free State (49,866 square miles), and Natal (33,578 square miles). The administrative capital is Pretoria, the legislative capital is Cape Town, and the judicial capital is Bloemfontein.

The self-administered republics of Transkei, Bophuthatswana, Venda, and Ciskei have been declared independent by South Africa, but they are not recognized by any other national government or by any international organization as being, in fact, independent. Namibia, which is situated to the northwest and has an area of 318,261 square miles, was formerly administered by the South African government, but negotiations held throughout the 1980s resulted in the territory's independence in 1990. Once a German colony, it was mandated to South Africa by the League of Nations; the United Nations, however, challenged South Africa's continuing administration of the territory on the grounds that the League's mandate was terminated, a position confirmed by an advisory opinion of the International Court of Justice in 1971. Walvis Bay, the only useful harbour of Namibia, was administratively attached to the Cape Province (then under British rule) during the latter part of the 19th century and continues to be administered by the South African government as part of the republic. South Africa also possesses two small subantarctic islands, Prince Edward and Marion, situated in the Indian Ocean about 1,200 miles southeast of Cape Town.

South Africa is bordered by Namibia to the northwest, by Botswana and Zimbabwe (formerly Rhodesia) to the north, by Mozambique and Swaziland to the east, by the Indian Ocean to the southeast and south, and by the Atlantic Ocean to the south and west. The Kingdom of Lesotho, a constitutional monarchy in the eastern part of the republic, is surrounded by South African territory.

South Africa's location, almost entirely south of the

Tropic of Capricorn and within temperate climatic zones, contributed to European settlement on a scale unknown elsewhere in Africa. Although South Africa is administered by whites and its political affiliations and most of its economic links are with the Western bloc of nations, its position—3,800 miles from South America to the west, 4,700 miles from Australia to the east, and about 6,000 miles from the markets of western Europe—is a relatively isolated one.

Since World War II, South Africa has been a frequent focus of attention. This attention intensified as the former Portuguese colonies of Angola and Mozambique and the former white-minority-ruled state of Rhodesia, all close to South Africa, gained independence under African majority rule. At the same time, the South African government, dominated by the minority white population, has maintained a policy of apartheid (meaning "apartness"), which maintains separation and separate development of the races and which has evoked vehement opposition from most countries in the world and in the United Nations. South Africa's peoples comprise a diversity of ethnic and racial groups. (A.Ne./D.F.G.)

This article treats the Republic of South Africa. For treatment of the geography and history of the other states of the region, see SOUTHERN AFRICA.

The article is divided into the following sections:

Physical and human geography

THE LAND

Relief. Geologically, South Africa contains some of the oldest rocks in the world. Dominating the topography is a plateau that covers the largest part of the country and is separated from the small coastal strip by what is called the Great Escarpment. The plateau is actually a massive basin reaching heights of 8,000 feet (2,440 metres) in the basaltic Lesotho region, dropping to heights of approximately 2,000 to 3,000 feet in the sandy Kalahari Plateau in the west. The central part of the plateau comprises the Transvaal and Free State Highveld, which is between 4,000 and 6,000 feet in altitude. The Witwatersrand Ridge forms the watershed between the Vaal and Limpopo rivers and their tributaries. The Great Karoo (Karroo) Plateau, south of the Orange River, averages 4,000 feet in height and contains a number of large salt pans such as Verneukpan.

The Great Escarpment

The Great Escarpment, known by a variety of local names, is the most continuous topographical feature in South Africa. Starting in the far northeast, at altitudes of up to 7,000 feet, it is known as the Transvaal Drakensberg ("berg" and "berge," in Afrikaans, meaning mountains). Continuing southward, the mountains become known as the Natal Drakensberg and reach heights up to 11,000 feet. Farther to the south, the mountains become less impressive, ranging from 8,000 feet in the Sneeuberg section to 6,000 feet in the Nuweveldberge and to 5,000 feet in the Roggeveldberge region. In the west the escarpment comprises the Bokkeveldberg and the Kamiesberg (5,600 feet), between which both the plateau and the escarpment are not clearly defined. In the far south is an area of folded mountains that includes ranges such as the Hottentots Hollandsberge, Drakensteinberge, Sederberge, Langeberge, Groot-Swartberge, Outeniekwaberge, Tsitsikamaberge, and the Hexrivierberge, all of which vary in height between 5,000 and 7,600 feet and are geologically older than either the Alps or the Himalayas.

The narrow South African coastal plain is poorly developed; it is less than 500 feet above sea level, and only in northern Natal does it reach a mentionable width—extending about 30 miles along the west coast and broadening to about 80 miles in northern Natal. South Africa has a straight, unindented, and somewhat monotonous shoreline; the only good natural harbour along its 1,836-mile length is at Saldanha Bay, the use of which has been restricted by a lack of fresh water.

Drainage. Because of insufficient and inadequate rainfall, South Africa has no rivers that form navigable waterways, even for short distances. The main drainage system is that of the Orange River, which drains an area of approximately 329,000 square miles; the Vaal and Caledon rivers are the main tributaries. The northern part of the plateau is drained by the Limpopo system, which empties into the Indian Ocean; the Limpopo's main tributaries are the Olifants, Marico, and Sand rivers. From the escarpment a number of smaller and shorter rivers, all of them irregular, flow seaward.

Soils. Three major soil regions may be distinguished. The soil region east of the 20-inch (500-millimetre) isohyet (*i.e.,* a line drawn on a map connecting points having equal rainfall) includes soils formed under conditions of a wet summer and dry winter climate, the more important types being laterite (red, leached, iron-bearing soil) and lateritic soils, unleached subtropical soils, and gley-like (*i.e.,* bluish-gray, sticky, and compact) podzolic soils (highly leached soils that are low in iron and lime). The second major soil region lies within the winter rainfall zone and the southern coastal area with all-season rainfall; gray sandy soils and sandy loams form the main types. West of the approximately 15- to 20-inch isohyet—an area that includes a large part of the interior and all of the western coastal area—the land is semidesert and desert, with rain falling primarily in summer and with a huge annual evaporation rate. The soils of this region as a whole are characterized by a top sandy layer, often a sandy loam, which is underlain by a layer of lime or an accretion of silica.

Climate. The climate is controlled by three main factors. First, South Africa's location between latitudes 22° and 35° S places it—except for a small part in the north—fully within the temperate zone, though it receives a great deal of sunshine. Apart from southwestern Cape Province, South Africa is out of reach of the rainy westerlies and next to a subtropical high-pressure belt of descending air where condensation and rainfall do not readily develop. Most of South Africa thus has a dry climate.

Second, South Africa is flanked on the east by the warm, southward-flowing Agulhas Current, and along the west coast by the cold, northward-flowing Benguela Current. The east coast is comparatively warmer, and the resultant warmer and less dense air rises more readily, facilitating the entry of rain-bearing clouds.

Finally, South Africa's inland plateau greatly tempers the influence of latitude on climate. From Cape Town in the south to the far north, the temperature hardly changes, the increasing intensity of the solar rays being moderated by the increase in altitude.

ıfluences
f the
ıbtropical
igh-
ressure
elt

Because South Africa lies in the subtropical high-pressure belt, the country, as well as the adjoining ocean areas, is influenced throughout the year by descending, divergent upper air masses that circulate primarily eastward, generally causing fine weather and low annual precipitation, especially to the west. During winter (June to August), cold polar air moves over the southwestern, southern, and southeastern coastal areas, sometimes reaching the southern interior of the country from the southwest. These polar masses are accompanied by cold fronts as well as by rain and snow. In summer (December to February) the Atlantic high-pressure system settles semipermanently over the southern and western parts of the country. Local heating of the landmass sometimes causes low-pressure conditions to develop, resulting in tropical air masses being drawn in from the Indian Ocean over the northeastern region and bringing rain.

South Africa is a generally semiarid country where farmers are constantly faced with the problem of a water shortage. About 21 percent of the country is arid, receiving less than eight inches of rainfall annually; another 47 percent is semiarid, receiving eight to 24 inches annually. Abundant rain, amounting to more than 40 inches annually, is received over only 6 percent of the country. From east to west there is a gradual decline in the rainfall. Natal receives more than 40 inches annually, Kimberley approximately 16 inches, and Alexander Bay on the west coast less than two inches.

Temperatures are generally moderate. The summers are moderately warm, with temperatures of between 70° and 75° F (21° and 24° C). Over the western plateau, summer temperatures vary from 75° to 85° F (24° to 29° C); in the Orange River Valley it becomes very hot, with temperatures rising to 90° F (32° C) and higher. South African winters are moderately cool; over the inland plateau July average temperatures fall below 50° F (10° C) and, in places, to less than 40° F (4° C). Along the east coast, in Natal and in the Transvaal Lowveld, even the winters are moderately warm. Influenced by the ocean currents, the temperature also declines from the east coast to the west coast. Durban, on the east coast, has an average annual temperature of 69° F (21° C), while Port Nolloth—on the west coast but approximately on the same latitude—registers 57° F (14° C).

Plant and animal life. In a semiarid country, forests, as such, rarely occur; the small Knysna Forest in southern Cape Province is perhaps the best example. Shrublike vegetation is found in the Karoo region, monotonous grassland with very few trees occurs on the Highveld, and savanna (parklike grassland) is found in a variety of areas such as the Transvaal Bushveld and Lowveld, the Kalahari, the deep valleys of the eastern Cape, and the Natal coastal region.

When the first European settlers arrived in the 17th century, South Africa had an unbelievable wealth of animal life, including lions, elephants, rhinoceroses, and a variety of antelopes; today, however, such wildlife exists only in the most marginal areas, such as the Kalahari. To preserve the larger animal species from extermination, a number of wildlife reserves came into being, of which the Kruger National Park in the northeast is the best known. There is a rich birdlife, and of the more than 100 species of snakes, about one-fourth are poisonous. (A.Ne.)

Settlement patterns. While South Africa could be divided into numerous small regions, known by a variety of local names, these may be grouped more conveniently into 11 large traditional regions.

The veld
regions

Three veld regions may be distinguished in the north. These are the Bushveld in the far northern Transvaal west of the Transvaal Drakensberg; the Lowveld east of the mountains; and the Highveld to the south, including the southern Transvaal, eastern Orange Free State, Lesotho, and an adjoining portion of Cape Province. The Bushveld and the Lowveld are still predominantly rural, but the prairie-like landscape of the Highveld on the eastern interior plateau contains the mining, industrial, and commercial heart of South Africa and the largest urban centres.

To the west and southwest of these veld regions the arid and semiarid landscape consists of three traditional regions. These are the Kalahari in the north, which includes most of Cape Province north of the Orange River but also extends into Botswana and eastern Namibia and which is still only sparsely populated; the Karoo, which lies south of the Kalahari and includes most of the interior of Cape Province and which is inhabited by rural Afrikaans-speaking people and a very few Africans; and Namaqualand, a rugged rural area in the extreme west, immediately south of the Orange River.

The final group includes the more humid regions bordering the South Atlantic and Indian oceans. These are the Western Province in the southwestern corner of the country, the area of the origin of the Cape Coloureds and the oldest white settlement in South Africa; the Eastern Province, centring on the port towns of Port Elizabeth and East London; the Southern Cape Districts (in South Africa itself perhaps better known as the Southwestern Districts), located between the Western and Eastern provinces; Transkei, inhabited by the Xhosa- (Xosa-) speaking people, situated to the northeast of the Eastern Province; and Natal, a garden province, whose character has been shaped by large numbers of English-speaking whites, most of South Africa's Indian population, and the indigenous African peoples.

Almost one-half of the total population still lives in rural areas. There has been a proportional decline, however, in the rural population in recent years. At the beginning of the 20th century, half of the white population was still rural, whereas today the proportion has dropped considerably. The African rural population has also declined greatly.

In the African areas settlement traditionally has taken the form of farming villages; the land belongs to the kraal (village) community, each family receiving from the chief or headman the right to build a home. Pastoral land surrounding the kraal is used communally. Among the whites of South Africa, settlement has taken the form of dispersed farms. The farms are often miles apart in the arid and semiarid regions but are closer together in the better watered parts.

Towns and cities

Urban settlement is much denser in the east and south. All of the largest cities and towns are located in this more humid half of the country. Also located there are the four main urban concentrations—the Pretoria–Witwatersrand–Vereeniging complex in the southern Transvaal; the Cape Peninsula in southwestern Cape Province; the Durban–Pietermaritzburg area in Natal; and the Port Elizabeth–Uitenhage region in eastern Cape Province.

Of these, the Witwatersrand conurbation is by far the largest and the most important for mining, industry, and commerce. Most South African towns originated as centres that supplied services to the surrounding rural areas. Others, such as Grahamstown and Simonstown, originally had a military function, while still others, such as Welkom, Virginia, and Kimberley, began as mining towns, sometimes—as in the case of Kimberley—shedding the mining function at a later stage. Some towns, like Sasolburg, where gasoline is manufactured from coal, developed around a specific industry. Germiston and De Aar are railroad centres, and Durban, Cape Town, Port Elizabeth, and East London are harbour towns. Pretoria is the main administrative centre, and Stellenbosch and Grahamstown are towns where important educational institutions are located. Johannesburg, the economic hub of the Witwatersrand area, is the major city.

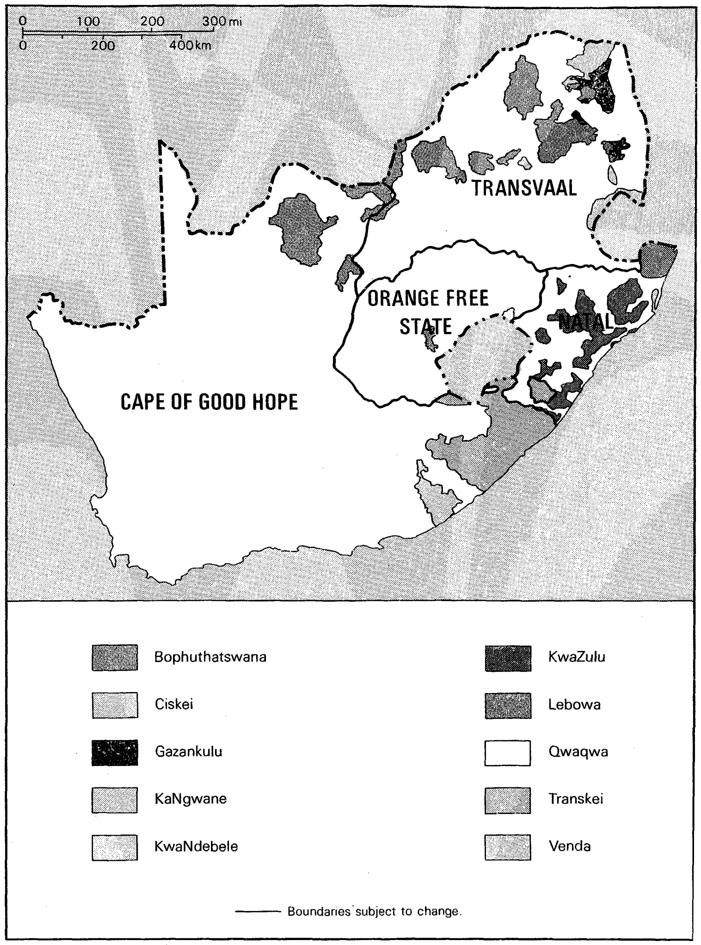
South African cities generally are marked by white residential districts near the city centre and African residential "townships" geographically separated in the outlying areas. Soweto (South-Western Townships), which adjoins Johannesburg, is the largest urban settlement in South Africa. Umlazi is the only town of more than 100,000 population without a substantial white population.

THE PEOPLE

Ethnic distribution. Four racial groups are designated under South African law, and each of these may be further subdivided into different linguistic and ethnic groups. Those classified as whites consist of two main groups—the Afrikaners, who are the descendants of Dutch, French, and, to a lesser degree, German forebears and speak Afrikaans; and the English-speaking group. The Afrikaners comprise more than half of the white population, and the English-speaking make up the remainder. Afrikaans and English are the official languages.

Nonwhites in South Africa outnumber the whites by approximately six to one; they comprise a variety of peoples classified as African, Coloured, and Asian. These classifications are highly arbitrary, however, and sometimes result in legal contestation. The Africans comprise about three-fourths of South Africa's population and have the fastest growth rate of all groups. The South African government emphasizes the ethnic differences among Africans, treating the African population as if it were composed of 10 separate nations, each with its own territory. The Africans more properly can be divided into four main ethnolinguistic groups. These are the Nguni, including the Xhosa, Zulu, Swazi, and Ndebele peoples, who comprise more than half of the African people of South Africa and live mainly in the eastern coastal regions; the Sotho, found mostly in the central and western African areas and constituting the largest minority of the Africans; and the Venda and Tsonga groups, who comprise quite a small percent of the Africans and who are found only in the northern Transvaal. Other Africans form the rest of the total population. At any given time about one-half of the African population is in the republics or the Black States, areas set aside for occupation by blacks, which comprise 13 percent of the country's geographical area. Of the rest, approximately two-thirds are in the cities, and one-third are in the rural areas set aside for white occupation, where Africans work mainly as farm labourers.

The African majority



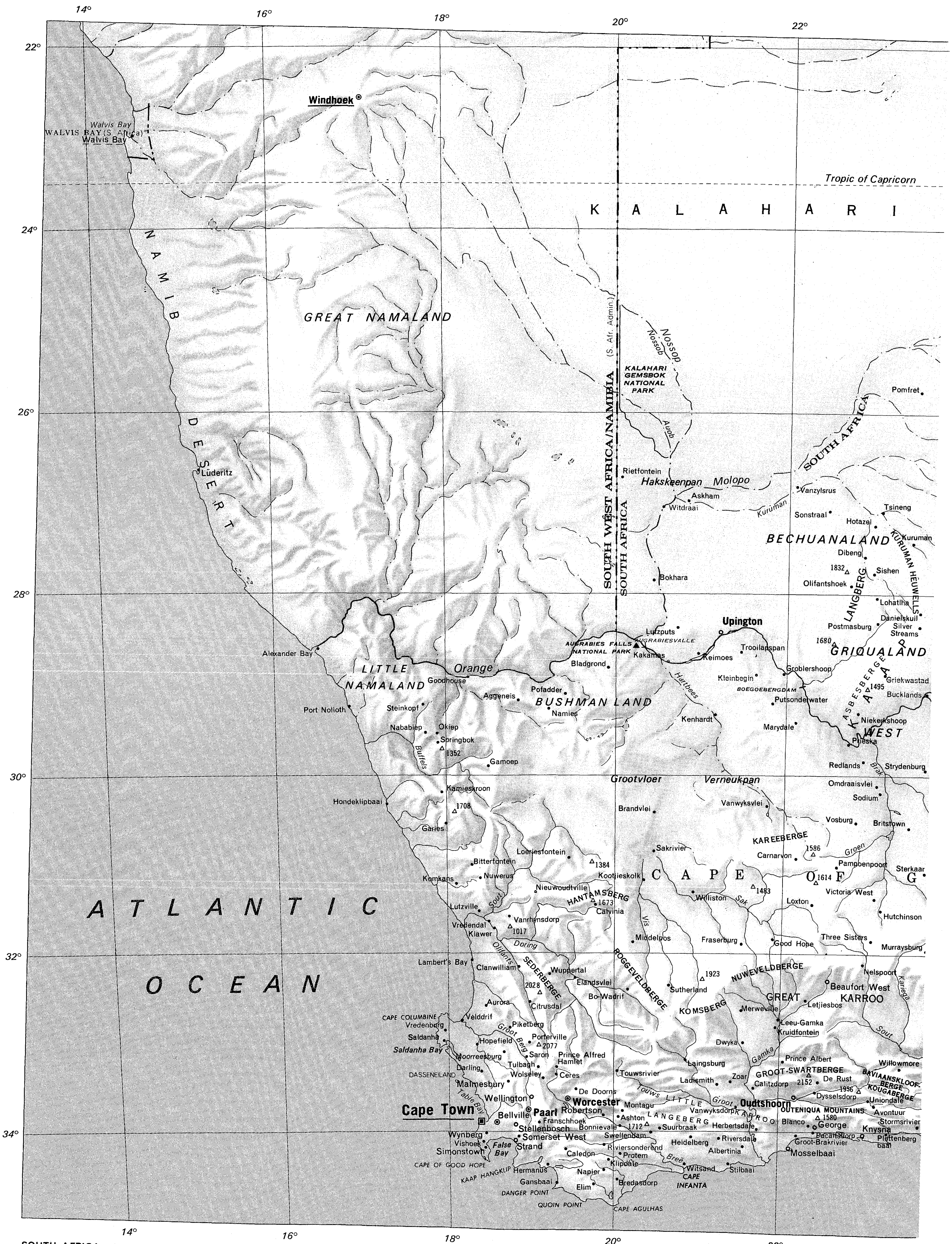
Republics and Black States within South Africa.

The smaller nonwhite groups are the Coloureds and the Asians—predominantly Indian—of whom the majority live in Natal Province.

South Africa's Coloured population has its roots in the early period (the 17th and 18th centuries) of the Dutch East India Company's regime at the Cape of Good Hope. Children were born of white colonists and sailors and Malay slave women (from the Dutch East Indies) and indigenous Khoisan women, and also of Khoisan women and Malay slave men, especially after the latter were freed in the early 19th century. Varying in skin colour, the Coloureds are culturally closest to the Afrikaners. Most speak Afrikaans and live in the western part of Cape Province, where they form a majority of the population.

The founding of an Asian community began in 1860, when a number of Indians were brought under contract to work on the Natal sugar plantations; the community grew, and by the early 1980s represented a tiny percentage of the population. The original Khoisan people (the Khoikhoin and the San, termed Hottentots and Bushmen by the whites) are almost nonexistent; their few descendants are found in parts of the Kalahari.

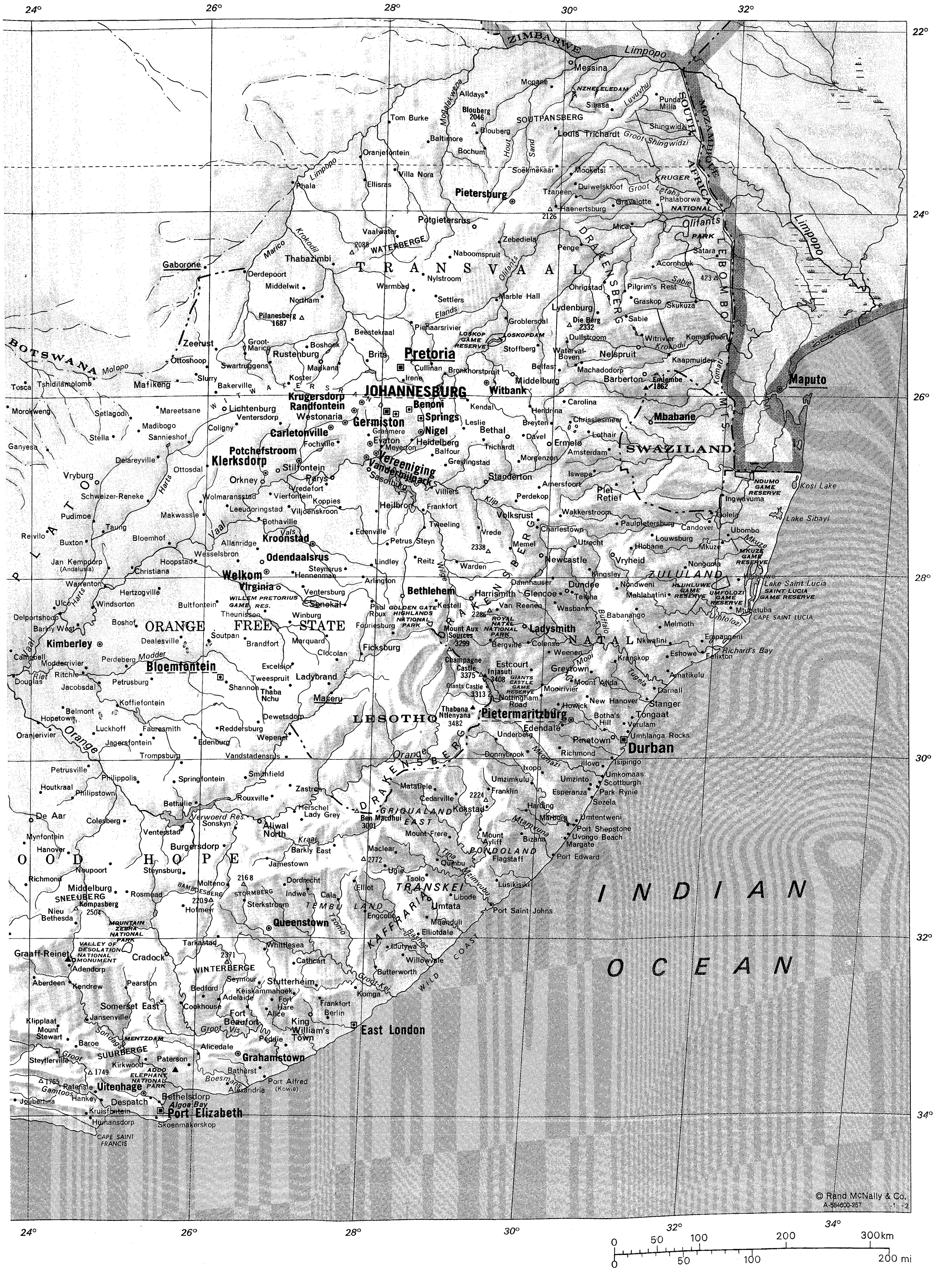
Religion. Among the whites and Coloureds, the Dutch Reformed Church has by far the most adherents of the organized religions, including a large number of African



SOUTH AFRICA

Size of symbol indicates relative size of town • ○ ◎ ▣ ▣ ■

Elevations in metres


24°
26°
28°
30°
32°
34°
22°
24°
26°
28°
30°
32°
34°
ZIMBABWE
Limpopo
Messina
Mopane
NZHELELEDAM
Sibasa
Luvuvhu
Punda Milia
Alldays
Mogalakwena
Blouberg 2046
Blouberg
SOUTPANSBERG
Tom Burke
Shingwidzi
Louis Trichardt
Groot Shingwidzi
SOUTH AFRICA
MOZAMBIQUE
Baltimore
Bochum
Hout
Sand
Oranjefontein
Villa Nora
Soekmekaar
Mooketsi
Limpopo
Phala
Ellisras
Pietersburg
Tzaneen
Duiwelskloof
Groot Letaba
KRUGER
Gravelotte
Phalaborwa
NATIONAL
Haenertsburg
2126
Potgietersrus
Vaalwater
2088
WATERBERGE
Zebediela
Mica
Olifants
PARK
Satara
Limpopo
Marico
Krokodil
Thabazimbi
Naboomspruit
TRANSVAAL
DRAKENSBERG
Penge
Acornhoek
Gaborone
Derdepoort
Middelwit
Nylstroom
Warmbad
Olifants
Ohrigstad
Pilgrim's Rest
Sabie
423
Northam
Settlers
Marble Hall
Lydenburg
Graskop
Skukuza
Pilanesberg 1687
Elands
Groblersdal
Die Berg 2332
Sabie
LEBOMBO MTS.
Beestekraal
Pienaarsrivier
LOSKOP GAME RESERVE
LOSKOPDAM
Dullstroom
Witrivier
Komatipoort
Zeerust
Groot-Marico
Boshoek
Stoffberg
Krokodil
BOTSWANA
Molopo
Ottoshoop
Rustenburg
Brits
Pretoria
Waterval-Boven
Nelspruit
Kaapmuiden
Swartruggens
Madikana
Cullinan
Bronkhorstspruit
Belfast
Machadodorp
Maputo
Tosca
Tshidilamolomo
Mafikeng
Slurry
Koster
Irene
Middelburg
Barberton
Emlembe 1862
Komati
Bakerville
WITWATERSRAND
JOHANNESBURG
Witbank
Morokweng
Setlagodi
Mareetsane
Lichtenburg
Krugersdorp
Randfontein
Benoni
Kendal
Hendrina
Carolina
Ventersdorp
Westonaria
Springs
Madibogo
Coligny
Germiston
Leslie
Breyten
Chrissiesmeer
Mbabane
Stella
Sannieshof
Carletonville
Nigel
Bethal
Davel
Lothair
Ganyesa
Grasmere
Heidelberg
Ermelo
SWAZILAND
Fochville
Evaton
Balfour
Trichardt
Amsterdam
Potchefstroom
Meyerton
Delareyville
Klerksdorp
Vereeniging
Greylingstad
Morgenzon
Ottosdal
Stilfontein
Vanderbijlpark
Vryburg
Orkney
Parys
Sasolburg
Standerton
Iswepe
NDUMO GAME RESERVE
Kosi Lake
Schweizer-Reneke
Harts
Wolmaransstad
Vredefort
Vierfontein
Villiers
Amersfoort
Piet Retief
Koppies
Perdekop
Ingwavuma
Pudimoe
Makwassie
Leeudoringstad
Viljoenskroon
Heilbron
Frankfort
Klip
Wakkerstroom
Lake Sibayi
Taung
Bothaville
Volksrust
Golela
Paulpietersburg
Candover
Reivilo
Buxton
Bloemhof
Vaal
Vals
Tweeling
Vrede
Charlestown
Ubombo
Allanridge
Kroonstad
Edenville
2338
Mkuze
Louwsburg
Jan Kempdorp (Andalusia)
Wesselsbron
Odendaalsrus
Petrus Steyn
Memel
Utrecht
Hlobane
Mkuze
MKUZE GAME RESERVE
Hoopstad
Reitz
Warden
Newcastle
Vryheid
Nongoma
Christiana
Steynsrus
Hennenman
Lindley
Wilge
Kingsley
Warrenton
Welkom
Virginia
Arlington
Dannhauser
ZULULAND
Hertzogville
Ventersburg
Bethlehem
Harrismith
Glencoe
Dundee
Nondweni
HLUHLUWE GAME RESERVE
Lake Saint Lucia
SAINT LUCIA GAME RESERVE
Harts
Ulco
Windsorton
Bultfontein
WILLEM PRETORIUS GAME RES.
Senekal
Paul Roux
GOLDEN GATE HIGHLANDS NATIONAL PARK
Kestell
Van Reenen
Talana
Mahlabatini
UMFOLOZI GAME RESERVE
Mtubatuba
Delportshoop
Theunissen
Wasbank
Babanango
CAPE SAINT LUCIA
Barkly West
Boshof
ORANGE FREE STATE
Winburg
Fouriesburg
2286
ROYAL NATAL NATIONAL PARK
Ladysmith
Melmoth
Umfolozi
Kimberley
Dealesville
Soutpan
Brandfort
Marquard
Mount Aux Sources 3299
Bergville
Colenso
Buffalo
NATAL
Nkwalini
Empangeni
Vaal
Campbell
Modder
Ficksburg
Weenen
Eshowe
Richard's Bay
Modderrivier
Perdeberg
Bloemfontein
Clocolan
Champagne Castle 3375
Estcourt
Mooi
Kranskop
Felixton
Riet
Ritchie
Excelsior
Injasuti 3408
GIANTS CASTLE GAME RESERVE
Greytown
Tugela
Amatikulu
Douglas
Petrusburg
Tweespruit
Ladybrand
Giants Castle 3313
Mount Alida
Darnall
Jacobsdal
Shannon
Thaba Nchu
Nottingham Road
Mooirivier
New Hanover
Stanger
Koffiefontein
Maseru
Howick
Tongaat
Belmont
Dewetsdorp
Thabana Ntlenyana 3482
LESOTHO
Pietermaritzburg
Botha's Hill
Verulam
Hopetown
Luckhoff
Fauresmith
Reddersburg
Wepener
Edendale
Underberg
Umhlanga Rocks
Oranjerivier
Orange
Jagersfontein
Edenburg
Pinetown
Durban
Trompsburg
Vandstadensrus
Orange
DRAKENSBERG
Mkomazi
Richmond
Donnybrook
Illovo
Isipingo
Petrusville
Philippolis
Springfontein
Smithfield
Ixopo
Umkomaas
Scottburgh
Houtkraal
Philipstown
Zastron
Umzimkulu
Umzinto
Park Rynie
Matatiele
Franklin
Esperanza
Sezela
Bethulie
Rouxville
Cedarville
2224
Harding
De Aar
Colesberg
Verwoerd Res.
Herschel
Lady Grey
GRIQUALAND EAST
Kokstad
Umtentweni
Venterstad
Sonskyn
Aliwal North
Ben Macdhui 3001
Marburg
Port Shepstone
Mynfontein
Mount Frere
Mtamvuna
Uvongo Beach
Margate
Hanover
Burgersdorp
Kraai
Mount Ayliff
Bizana
Barkly East
Maclear
PONDOLAND
Port Edward
GOOD HOPE
Jamestown
2772
Tina
Flagstaff
Noupoort
Steynsburg
Qumbu
Ugie
Mzimvubu
Richmond
2168
Dordrecht
Tsolo
Lusikisiki
Middelburg
BAMBOESBERG
Molteno
TRANSKEI
Elliot
SNEEUBERG
Rosmead
2209
STORMBERG
Indwe
Cala
Libode
Kompasberg 2504
Hofmeyr
Sterkstroom
TEMBU LAND
KAFFRARIA
Nieu Bethesda
Engcobo
Umtata
Port Saint Johns
INDIAN
MOUNTAIN ZEBRA NATIONAL PARK
Queenstown
Tsomo
Mqanduli
Elliotdale
Bashee
VALLEY OF DESOLATION NATIONAL MONUMENT
Tarkastad
Whittlesea
Idutywa
WILD COAST
Graaff-Reinet
Cradock
2371
OCEAN
Adendorp
WINTERBERGE
Cathcart
Willowvale
Aberdeen
Kendrew
Pearston
Stutterheim
Butterworth
Seymour
Keiskammahoek
Groot-Kei
Komga
Bedford
Somerset East
Adelaide
Fort Hare
Frankfort
Jansenville
Cookhouse
Alice
Berlin
Klipplaat
Fort Beaufort
King William's Town
East London
Mount Stewart
Groot
Vis
Baroe
Sondags
MENTZDAM
Peddie
Alicedale
Steytlerville
Groot
SUURBERGE
Paterson
Grahamstown
Kirkwood
1749
ADDO ELEPHANT NATIONAL PARK
Bathurst
1765
Boesmans
Port Alfred (Kowie)
Ratelsie
Uitenhage
Alexandria
Gamtoos
Hankey
Despatch
Bethelsdorp
Joubertina
Algoa Bay
Kruisfontein
Port Elizabeth
Humansdorp
Skoenmakerskop
CAPE SAINT FRANCIS
© Rand McNally & Co.
A-584600-257 -1 -2
0
50
100
200
300 km
0
50
100
200 mi

MAP INDEX

Political subdivisions

Cape of Good Hope	31.00s	23.00e
Natal	28.40s	30.40e
Orange Free State	28.30s	27.00e
Transvaal	25.00s	29.00e

Cities and towns

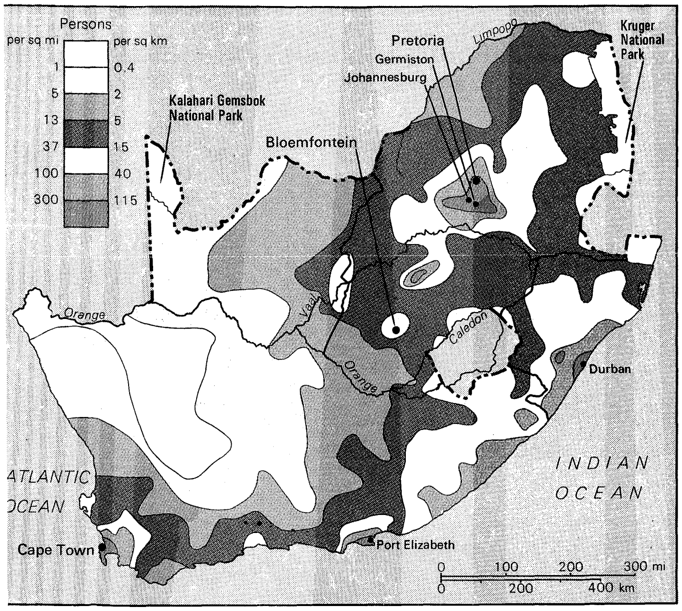
Aberdeen	32.29s	24.04e
Acornhoek	24.37s	31.02e
Adelaide	32.42s	26.20e
Adendorp	32.20s	24.33e
Aggeneis	29.03s	18.51e
Albertinia	34.13s	21.36e
Alexander Bay	28.40s	16.30e
Alexandria	33.39s	26.25e
Alicedale	33.19s	26.05e
Aliwal North	30.45s	26.45e
Allanridge	27.55s	26.44e
Alldays	22.44s	29.04e
Amatikulu	29.06s	31.27e
Amersfoort	27.00s	29.53e
Amsterdam	26.35s	30.45e
Andalusia, see Jan Kempdorf		
Arlington	28.06s	27.54e
Ashton	33.50s	20.05e
Askham	26.59s	20.47e
Aurora	32.42s	18.29e
Avontuur	33.44s	23.11e
Babanango	28.30s	31.00e
Bakerville	26.00s	26.06e
Balfour	26.44s	28.45e
Baltimore	23.15s	28.20e
Barberton	25.48s	31.03e
Barkly East	30.58s	27.33e
Barkly West	28.05s	24.31e
Baroe	33.13s	24.33e
Bathurst	33.30s	26.50e
Beaufort West	32.18s	22.36e
Bedford	32.41s	26.05e
Beestekraal	25.23s	27.38e
Belfast	25.43s	30.03e
Bellville	33.53s	18.36e
Belmont	29.28s	24.22e
Benoni	26.19s	28.27e
Bergville	28.52s	29.18e
Berlin	32.54s	27.35e
Bethal	26.27s	29.28e
Bethelsdorp	33.52s	25.34e
Bethlehem	28.15s	28.15e
Bethulie	30.32s	25.59e
Bitterfontein	31.00s	18.32e
Bizana	30.58s	29.52e
Biadgrond	28.52s	19.57e
Blanco	33.57s	22.24e
Bloemfontein	29.12s	26.07e
Bloemhof	27.38s	25.32e
Blouberg	23.08s	28.56e
Bochum	23.17s	29.07e
Bokhara	27.57s	20.30e
Bonnievale	33.57s	20.06e
Boshoek	25.30s	27.09e
Boshof	28.34s	25.04e
Botha's Hill	29.45s	30.45e
Bothaville	27.27s	26.36e
Bo-Wadrif	32.26s	20.07e
Brandfort	26.47s	28.30e
Brandvlei	30.25s	20.30e
Bredasdorp	34.32s	20.02e
Breyten	26.16s	30.00e
Brits	25.42s	27.45e
Britstown	30.37s	23.30e
Bronkhorstspruit	25.50s	28.45e
Bucklands	29.03s	23.44e
Bultfontein	28.20s	26.05e
Burgersdorp	31.00s	26.20e
Butterworth	32.23s	28.04e
Buxton	27.38s	24.42e
Cala	31.30s	27.37e
Caledon	34.12s	19.23e
Calitzdorp	33.33s	21.42e
Calvinia	31.25s	19.45e
Campbell	28.48s	23.44e
Candover	27.28s	31.57e
Cape Town (Kaapstad)	33.55s	18.22e
Carletonville	26.23s	27.22e
Carnarvon	30.56s	22.08e
Carolina	26.05s	30.06e
Cathcart	32.18s	27.09e
Cedarville	30.23s	29.03e
Ceres	33.21s	19.18e
Charlestown	27.30s	29.55e
Chrissiesmeer	26.16s	30.13e
Christiana	27.52s	25.08e
Citrusdal	32.36s	19.00e
Clanwilliam	32.11s	18.54e
Clocolan	29.00s	27.30e
Colenso	28.50s	29.44e
Colesberg	30.45s	25.05e
Coligny	26.17s	26.15e
Cookhouse	32.44s	25.48e
Cradock	32.08s	25.36e
Cullinan	25.40s	28.32e
Danielskuil	28.11s	23.33e
Dannhauser	28.04s	30.04e
Darling	33.23s	18.23e
Darnall	29.23s	31.18e
Davel	26.24s	29.40e
De Aar	30.39s	24.00e
Dealesville	28.40s	25.37e
De Doorns	33.28s	19.41e
Delareyville	26.44s	25.29e
Delportshoop	28.22s	24.20e
Derdepoort	24.42s	26.20e
De Rust	33.30s	22.32e
Despatch	33.46s	25.30e
Dewetsdorp	29.33s	26.34e
Dibeng	27.35s	22.54e
Donnybrook	30.00s	29.48e
Dordrecht	31.20s	27.03e
Douglas	29.04s	23.46e
Duiwelskloof	23.42s	30.06e
Dullstroom	25.27s	30.07e
Dundee	28.12s	30.16e
Durban	29.55s	30.56e
Dwyka	33.02s	21.30e
Dysselsdorp	33.34s	22.28e
East London	33.00s	27.55e
Edenburg	29.45s	25.56e
Edendale	29.39s	30.18e
Edenville	27.37s	27.34e
Elandsvlei	32.19s	19.33e
Elim	34.35s	19.45e
Elliot	31.18s	27.50e
Elliotdale	31.55s	28.38e
Ellisras	23.40s	27.46e
Empangeni	28.50s	31.48e
Engcobo	31.37s	28.00e
Ermelo	26.34s	29.58e
Eshowe	28.58s	31.29e
Esperanza	30.21s	30.40e
Estcourt	29.01s	29.52e
Evaton	26.31s	27.54e
Excelsior	28.56s	27.06e
Fauresmith	29.42s	25.21e
Felixton	28.50s	31.53e
Flagstaff	31.05s	29.29e
Fochville	26.30s	27.30e
Fort Beaufort	32.46s	26.40e
Fort Hare	32.47s	26.50e
Fouriesburg	28.38s	28.14e
Frankfort	27.17s	28.30e
Frankfort	32.44s	27.27e
Franklin	30.18s	29.30e
Franschhoek	33.55s	19.09e
Fraserburg	31.55s	21.30e
Gamoep	29.55s	18.25e
Gansbaai	34.35s	19.22e
Ganyesa	26.35s	24.10e
Garies	30.30s	18.00e
George	33.58s	22.24e
Germiston	26.15s	28.05e
Glencoe	28.12s	30.07e
Golela	27.20s	31.55e
Good Hope	31.51s	21.55e
Goodhouse	28.57s	18.13e
Graaff-Reinet	32.14s	24.32e
Grahamstown	33.19s	26.31e
Graskop	24.58s	30.49e
Grasmere	26.26s	27.52e
Gravelotte	23.56s	30.34e
Greylingstad	26.44s	28.45e
Greytown	29.07s	30.30e
Griekwastad	28.49s	23.15e
Groblersdal	25.15s	29.25e
Groblershoop	28.55s	20.59e
Groot-Brakrivier	34.01s	21.46e
Groot-Marico	25.37s	26.26e
Haenertsburg	24.00s	29.50e
Hankey	33.50s	24.52e
Hanover	31.04s	24.29e
Harding	30.34s	29.58e
Harrismith	29.18s	29.03e
Heidelberg	26.32s	28.18e
Heidelberg	34.06s	20.59e
Heilbron	27.21s	27.58e
Hendrina	26.11s	29.45e
Hennenman	27.59s	27.01e
Herbertsdale	34.01s	21.45e
Hermanus	34.25s	19.16e
Herschel	30.37s	27.12e
Hertzogville	28.08s	25.33e
Hlobane	27.42s	31.00e
Hluhluwe	28.01s	32.15e
Hofmeyr	31.39s	25.50e
Hondeklipbaai	30.20s	17.18e
Hoopstad	27.54s	25.58e
Hopefield	33.04s	18.22e
Hopetown	29.34s	24.03e
Hotazel	27.15s	23.00e
Houtkraal	30.23s	24.05e
Howick	29.28s	30.14e
Humansdorp	34.02s	24.46e
Hutchinson	31.30s	23.09e
Idutywa	32.02s	28.16e
Illovo	30.05s	30.50e
Indwe	31.27s	27.23e
Ingwavuma	27.09s	32.00e
Irene	25.51s	28.13e
Isipingo	30.00s	30.57e
Iswepe	26.60s	30.31e
Ixopo	30.08s	30.00e
Jacobsdal	29.13s	24.41e
Jagersfontein	29.44s	25.29e
Jamestown	31.06s	26.45e
Jan Kempdorp (Andalusia)	27.55s	24.51e
Jansenville	32.57s	24.40e
Johannesburg	26.12s	28.02e
Joubertina	33.50s	23.51e
Kaapmuiden	25.33s	31.20e
Kaapstad, see Cape Town		
Kakamas	28.45s	20.33e
Kamieskroon	30.09s	17.56e
Keimoes	28.41s	21.00e
Keiskammahoek	32.41s	27.09e
Kendal	26.04s	28.58e
Kendrew	32.31s	24.30e
Kenhardt	29.19s	21.12e
Kestell	28.19s	28.38e
Kimberley	28.43s	24.46e
Kingsley	27.55s	30.33e
King William's Town	32.51s	27.22e
Kirkwood	33.24s	25.26e
Klawer	31.44s	18.36e
Kleinbegin	28.50s	21.36e
Klerksdorp	26.58s	26.39e
Klipdale	34.19s	19.57e
Klippiaat	33.02s	24.21e
Knysna	34.02s	23.02e
Koffiefontein	29.30s	25.00e
Kokstad	30.32s	29.29e
Komatipoort	25.25s	31.55e
Komga	32.35s	27.55e
Komkans	31.16s	18.09e
Kootjieskolk	31.15s	20.21e
Koppies	27.20s	27.30e
Koster	25.57s	26.42e
Kowie, see Port Alfred		
Kranskop	29.00s	30.47e
Kroonstad	27.46s	27.12e
Krugersdorp	26.05s	27.35e
Kruidfontein	32.51s	21.57e
Kruisfontein	34.00s	24.43e
Kuruman	27.28s	23.28e
Ladismith	33.30s	21.16e
Ladybrand	29.19s	27.25e
Lady Grey	30.45s	27.13e
Ladysmith	28.34s	29.45e
Laingsburg	33.11s	20.51e
Lambert's Bay	32.05s	18.17e
Leeudoringstad	27.15s	26.10e
Leeu-Gamka	32.47s	21.59e
Leslie	26.27s	28.55e
Letjiesbos	32.34s	22.16e
Libode	31.33s	29.02e
Lichtenburg	26.08s	26.08e
Lindley	28.00s	27.57e
Loeriesfontein	30.56s	19.26e
Lohatlha	28.02s	23.04e
Lothair	26.26s	30.27e
Louis Trichardt	23.01s	29.43e
Louwsburg	27.37s	31.07e
Loxton	31.30s	22.22e
Luckhoff	29.44s	24.43e
Lusikisiki	31.25s	29.30e
Lutzputs	28.03s	20.40e
Lutzville	31.33s	18.22e
Lydenburg	25.10s	30.29e
Machadodorp	25.40s	30.14e
Maclear	31.02s	28.23e
Madibogo	26.25s	25.10e
Mafeking see Mafikeng		
Mafikeng	25.53s	25.39e
Mahlabatini	28.14s	31.30e
Makwassie	27.26s	26.00e
Malmesbury	33.28s	18.44e
Marble Hall	24.57s	29.13e
Marburg	30.44s	30.26e
Mareetsane	26.09s	25.25e
Margate	30.55s	30.15e
Marikana	25.42s	27.30e
Marguard	28.54s	27.28e
Marydale	29.23s	22.05e
Matatiele	30.24s	28.43e
Melmoth	28.38s	31.24e
Memel	27.43s	29.30e
Merweville	32.40s	21.31e
Messina	22.23s	30.00e
Meyerton	26.33s	28.01e
Mica	24.10s	30.48e
Middelburg	25.47s	29.28e
Middelburg	31.30s	25.00e
Middelpos	31.55s	20.13e
Middelwit	24.58s	27.00e
Mkuze	27.10s	32.00e
Modderrivier	29.02s	24.38e
Molteno	31.22s	26.22e
Montagu	33.45s	20.08e
Mooirivier	29.13s	29.50e
Mooketsi	23.35s	30.05e
Moorreesburg	33.08s	18.40e
Mopane	22.37s	29.52e
Morgenzon	26.45s	29.36e
Morokweng	26.12s	23.45e
Mosselbaai (Mossel Bay)	34.11s	22.08e
Mount Alida	29.09s	30.18e
Mountayliff	30.54s	29.20e
Mount Frere	31.00s	28.58e
Mount Stewart	33.10s	24.26e
Mqanduli	31.48s	28.46e
Mtubatuba	28.30s	32.08e
Murraysburg	31.58s	23.47e
Mynfontein	30.55s	23.57e
Nababiep	29.36s	17.46e
Naboomspruit	24.32s	28.36e
Namies	29.18s	19.13e
Napier	34.29s	19.53e
Nelspoort	32.07s	23.00e
Nelspruit	25.30s	30.58e
Newcastle	27.49s	29.55e
New Hanover	29.28s	30.28e
Niekerkshoop	29.19s	22.51e
Nieu Bethesda	31.51s	24.34e
Nieuwoudtville	31.23s	19.07e
Nigel	26.30s	28.28e
Nkwalini	28.45s	31.33e
Nondweni	28.11s	30.49e
Nongoma	27.58s	31.35e
Northam	25.03s	27.11e
Nottingham Road	29.22s	30.00e
Noupoort	31.10s	24.57e
Nuwerus	31.08s	18.24e
Nylstroom	24.42s	28.20e
Odendaalsrus	27.48s	26.45e
Ohrigstad	24.49s	30.33e
Okiep	29.39s	17.53e
Olifantshoek	27.57s	22.42e
Omdraaisvlei	30.08s	23.08e
Oranjefontein	23.25s	27.41e
Oranjerivier	29.40s	24.12e
Orkney	27.00s	26.39e
Ottosdal	26.58s	26.00e
Ottoshoop	25.45s	25.59e
Oudtshoorn	33.35s	22.14e
Paarl	33.45s	18.56e
Pacaltsdorp	34.00s	22.28e
Pampoenpoort	31.03s	22.40e
Park Rynie	30.25s	30.35e
Parys	27.04s	27.16e
Patensie	33.46s	24.49e
Paterson	33.26s	25.58e
Paulpietersburg	27.30s	30.51e
Paul Roux	28.18s	27.59e
Pearston	32.35s	25.08e
Peddie	33.14s	27.07e
Penge	24.22s	30.13e
Perdeberg	28.59s	25.05e
Perdekop	27.13s	29.38e
Petrusburg	29.08s	25.27e
Petrus Steyn	27.38s	28.08e
Petrusville	30.05s	24.41e
Phalaborwa	23.55s	31.13e
Philippolis	30.19s	25.13e
Philipstown	30.26s	24.29e
Pienaarsrivier	25.15s	28.18e
Pietermaritzburg	29.37s	30.16e
Pietersburg	23.54s	29.25e
Piet Retief	27.01s	30.50e
Piketberg	32.54s	18.46e
Pilgrim's Rest	24.55s	30.44e
Pinetown	29.52s	30.46e
Plettenbergbaai	34.04s	23.22e
Pofadder	29.10s	19.22e
Pomfret	25.50s	23.32e
Port Alfred (Kowie)	33.36s	26.55e
Port Edward	31.02s	30.13e
Port Elizabeth	33.58s	25.40e
Porterville	33.00s	19.00e
Port Nolloth	29.17s	16.51e
Port Saint Johns	31.38s	29.33e
Port Shepstone	30.46s	30.22e
Postmasburg	28.18s	23.05e
Potchefstroom	26.46s	27.01e
Potgietersrus	24.15s	28.55e
Pretoria	25.45s	28.10e
Prieska	29.40s	22.42e
Prince Albert	33.13s	22.02e
Prince Alfred Hamlet	33.18s	19.20e
Protem	34.16s	20.05e
Pudimoe	27.26s	24.44e
Punda Milia	22.40s	31.05e
Putsonderwater	29.09s	21.51e
Queenstown	31.52s	26.52e
Qumbu	31.10s	28.48e
Randfontein	26.11s	27.42e
Reddersburg	29.38s	26.07e
Redlands	29.51s	22.58e
Reitz	27.53s	28.31e
Reivilo	27.36s	24.08e
Richmond	29.54s	30.08e
Richmond	31.23s	23.56e
Rietfontein	26-44s	20.01e
Ritchie	29.02s	24.38e
Riversdale	34.07s	21.15e
Riviersonderend	34.09s	19.55e
Robertson	33.46s	19.50e
Rosmead	31.29s	25.08e
Rouxville	30.29s	26.46e
Rustenburg	25.37s	27.08e
Sabie	25.10s	30.48e
Sakrivier	30.54s	20.28e

MAP INDEX (continued)

Saldanha	33.00s	17.56e
Sannieshof	26.30s	25.47e
Saron	33.11s	19.01e
Sasolburg	26.48s	27.45e
Satara	24.29s	31.47e
Schweizer-Reneke	27.11s	25.18e
Scottburgh	30.19s	30.40e
Senekal	28.30s	27.32e
Setlagodi	26.16s	25.06e
Settlers	25.02s	28.30e
Seymour	32.33s	26.46e
Sezela	30.24s	30.42e
Shannon	29.08s	26.18e
Shingwidzi	23.05s	31.25e
Sibasa	22.53s	30.33e
Silver Streams	28.20s	23.33e
Simonstown (Simonstad)	34.14s	18.26e
Sishen	27.55s	22.59e
Skoenmakerskop	34.02s	25.33e
Skukuza	25.01s	31.38e
Slurry	25.49s	25.52e
Smithfield	30.09s	26.30e
Sodium	30.11s	23.09e
Soekmekaar	23.28s	29.58e
Somerset East	32.42s	25.35e
Somerset West	34.08s	18.50e
Sonskyn	30.47s	26.28e
Sonstraal	27.07s	22.28e
Soutpan	28.43s	26.04e
Springbok	29.43s	17.55e
Springfontein	30.19s	25.36e
Springs	26.13s	28.25e
Standerton	26.58s	29.07e
Stanger	29.27s	31.14e
Steinkopf	29.18s	17.43e
Stella	26.38s	24.48e
Stellenbosch	33.58s	18.50e
Sterkaar	31.05s	23.42e
Sterkstroom	31.32s	26.32e
Steynsburg	31.15s	25.49e
Steynsrus	27.58s	27.33e
Steytlerville	33.21s	24.21e
Stilbaai	34.24s	21.26e
Stilfontein	26.50s	26.50e
Stoffberg	25.29s	29.49e
Stormsrivier	33.59s	23.52e
Strand	34.06s	18.50e
Strydenburg	29.58s	23.40e
Stutterheim	32.33s	27.28e
Sutherland	32.24s	20.40e
Suurbraak	34.00s	20.39e
Swartruggens	25.40s	26.42e
Swellendam	34.02s	20.26e
Talana	28.10s	30.15e
Tarkastad	32.00s	26.16e
Taung	27.33s	24.47e
Thaba Nchu	29.17s	26.52e
Thabazimbi	24.41s	27.21e
Theunissen	28.30s	26.41e
Three Sisters	31.54s	23.06e
Tom Burke	23.05s	28.00e
Tongaat	29.37s	31.03e
Tosca	25.53s	23.58e
Touwsrivier	33.20s	20.00e
Trichardt	26.28s	29.13e
Trompsburg	30.01s	25.46e
Trooilapspan	28.40s	21.25e
Tshidilamolomo	25.50s	24.41e
Tsineng	27.06s	23.04e
Tsolo	31.18s	28.37e
Tulbagh	33.17s	19.09e
Tweeling	27.38s	28.31e
Tweespruit	29.11s	27.01e
Tzaneen	23.50s	30.09e
Ubombo	27.33s	32.00e
Ugie	31.10s	28.13e
Uitenhage	33.40s	25.28e
Ulco	28.21s	24.15e
Umhlanga Rocks	29.43s	31.06e
Umkomaas	30.15s	30.42e
Umtata	31.35s	28.47e
Umtentweni	30.42s	30.28e
Umzimkulu	30.16s	29.56e
Umzinto	30.22s	30.33e
Underberg	29.50s	29.22e
Uniondale	33.40s	23.08e
Upington	28.25s	21.15e
Utrecht	27.38s	30.20e
Uvongo Beach	30.51s	30.23e
Vaalwater	24.20s	28.03e
Vanderbijlpark	26.42s	27.54e
Van Reenen	28.22s	29.24e
Vanrhynsdorp	31.36s	18.44e
Vanstadensrus	29.59s	27.02e
Vanwyksdorp	33.46s	21.28e
Vanwyksvlei	30.18s	21.49e
Vanzylsrus	26.52s	22.04e
Velddrif	32.47s	18.11e
Ventersburg	28.09s	27.08e
Ventersdorp	26.17s	26.48e
Venterstad	30.47s	25.48e
Vereeniging	26.38s	27.57e
Verulam	29.45s	31.02e
Victoria West	31.25s	23.04e
Vierfontein	27.03s	26.46e
Viljoenskroon	27.12s	27.00e
Villa Nora	23.34s	28.01e
Villiers	27.03s	28.35e
Virginia	28.12s	26.49e
Vishoek	34.08s	18.25e
Volksrust	27.24s	29.53e
Vosburg	30.33s	22.52e
Vrede	27.30s	29.06e
Vredefort	27.05s	27.16e
Vredenburg	32.54s	17.59e
Vredendal	31.41s	18.35e
Vryburg	26.55s	24.45e
Wakkerstroom	27.24s	30.10e
Walvisbaai see Walvis Bay		
Walvis Bay	22.59s	14.31e
Warden	27.56s	29.00e
Warmbad	24.55s	28.15e
Warrenton	28.09s	24.47e
Wasbank	28.24s	30.05e
Waterval-Boven	25.40s	30.20e
Weenen	28.57s	30.03e
Welkom	27.59s	26.45e
Wellington	33.38s	18.57e
Wepener	29.46s	27.00e
Wesselsbron	27.50s	25.23e
Westonaria	26.19s	27.41e
Whittlesea	32.10s	26.50e
Williston	31.20s	20.53e
Willowmore	33.17s	23.29e
Willowvale	32.16s	28.30e
Winburg	28.37s	27.00e
Windsorton	28.16s	24.44e
Witbank	25.56s	29.07e
Witdraai	26.58s	20.45e
Witrivier	24.40s	31.00e
Witsand	34.24s	20.50e
Wolmaransstad	27.12s	26.13e
Wolseley	33.26s	19.12e
Worcester	33.39s	19.27e
Wuppertal	32.15s	19.15e
Wynberg	34.02s	18.28e
Zastron	30.18s	27.07e
Zebediela	24.19s	29.21e
Zeerust	25.33s	26.06e
Zoar	33.30s	21.28e

Physical features and points of interest

Addo Elephant National Park	33.29s	25.46e
Agulhas, Cape	34.52s	20.00e
Algoa Bay	33.50s	25.65e
Asbesberge, *mountains*	28.55s	23.15e
Atlantic Ocean	32.00s	15.00e
Augrabies Falls National Park	28.35s	20.19e
Augrabiesvalle, *waterfalls*	28.35s	20.19e
Auob, *river*	26.36s	20.36e
Bamboesberg, *mountains*	31.30s	26.10e
Bashee, *river*	32.15s	28.53e
Baviaanskloofberge, *mountains*	33.35s	24.10e
Bechuanaland, *physical region*	27.10s	22.10e
Blouberg, *mountain*	23.01s	28.59e
Boegoebergdam, *dam*	29.02s	22.12e
Boesmans, *river*	33.42s	26.39e
Brak, *river*	29.35s	22.55e
Breë, *river*	34.24s	20.50e
Buffalo, *river*	28.43s	30.37e
Buffels, *river*	33.45s	21.11e
Bushman Land, *historic region*	29.13s	19.35e
Caledon, *river*	30.31s	26.05e
Champagne Castle, *mountain*	29.06s	29.20e
Columbine, Cape	32.47s	18.52e
Danger Point	34.40s	19.17e
Dasseneiland, *island*	33.26s	18.04e
Die Berg, *mountain*	25.12s	30.09e
Doring, *river*	31.54s	18.39e
Drakensberg, *mountains*	27.00s	30.00e
Elands, *river*	25.10s	29.10e
False Bay (Valsbaai)	34.12s	18.40e
Gamka, *river*	33.18s	21.39e
Gakarosa, *mountain*	27.50s	23.38e
Gamtoos, *river*	33.58s	25.01e
Giants Castle, *mountain*	29.21s	29.27e
Giants Castle Game Reserve, *wildlife refuge*	29.16s	29.30e
Golden Gate Highlands National Park	28.30s	28.40e
Good Hope, Cape of	34.24s	18.30e
Gourits, *river*	34.21s	21.52e
Great Karroo, *plateau*	32.25s	22.40e
Great Swartburg, see Groot-Swartberg		
Griqualand East, *physical region*	30.30s	29.00e
Griqualand West, *physical region*	28.20s	23.30e
Groen, *river*	30.40s	23.17e
Groot, *river*	33.54s	21.39e
Groot-Berg, *river*	32.47s	18.08e
Groot-Kei, *river*	32.41s	29.22e
Groot Letaba, *river*	23.58s	31.50e
Groot Shingwidzi, *river*	23.15s	31.35e
Groot-Swartberge (Great Swartberg) *mountains*	33.22s	22.20e
Groot-Vis, *river*	33.30s	27.08e
Grootvloer, *salt flat*	30.00s	20.40e
Hakskeenpan, *salt lake*	26.48s	20.12e
Hangklip, Kaap, *cape*	34.26s	18.48e
Hantamsberg, *mountains*	31.22s	19.45e
Hartbees, *river*	28.45s	20.32e
Harts, *river*	28.24s	24.17e
Hluhluwe Game Reserve, *wildlife refuge*	28.05s	32.04e
Hout, *river*	23.04s	29.36e
Indian Ocean	32.00s	32.00e
Infanta, Cape	34.29s	20.51e
Injaswti, *mountain*	29.09s	29.23e
Kaap Plato, *plateau*	28.20s	23.57e
Kaffraria, *historic region*	31.30s	28.30e
Kalahari Gemsbok National Park	26.00s	20.30e
Kareeberge, *mountains*	30.53s	21.57e
Kariega, *river*	33.03s	23.28e
Klip, *river*	26.16s	27.49e
Komati, *river*	25.26s	32.00e
Kompasberg, *mountain*	31.45s	24.32e
Komsberg, *mountains*	32.40s	20.50e
Kosi, *lake*	26.55s	32.52e
Kougaberge, *mountains*	33.40s	23.50e
Kraai, *river*	30.40s	26.45e
Krokodil, *river*	25.26s	32.00e
Kruger National Park	24.00s	31.40e
Kuruman, *river*	26.56s	20.39e
Kuruman Heuwels, *hills*	27.40s	23.25e
Langeberg, *mountains*	34.00s	20.40e
Lebombo Mountains	25.30s	32.00e
Limpopo, *river*	22.25s	31.20e
Little Karroo, *plateau*	33.45s	21.30e
Little Namaland, *historic region*	29.00s	17.15e
Loskopdam, *dam*	25.23s	29.20e
Loskop Game Reserve, *wildlife refuge*	25.23s	29.20e
Luvuvhu, *river*	22.40s	30.55e
Macdhui, Ben, *mountain*	30.39s	27.58e
Marico, *river*	23.35s	26.57e
Mentzdam, *dam*	33.10s	25.09e
Mkomazi, *river*	30.12s	30.50e
Mkuze, *river*	27.53s	32.29e
Mkuze Game Reserve, *wildlife refuge*	27.40s	32.15e
Modder, *river*	29.02s	24.37e
Mogalakwena, *river*	23.00s	28.40e
Molopo, *river*	28.30s	20.13e
Mooi, *river*	28.45s	30.34e
Mountain Zebra National Park	32.16s	25.29e
Mount aux Sources, *mountain*	28.46s	28.52e
Mtamvuna, *river*	31.06s	30.12e
Mzimvubu, *river*	31.38s	29.32e
Ndumo Game Reserve, *wildlife refuge*	26.53s	32.15e
Nossob, *riber*	26.55s	20.37e
Nuweveldberge, *mountains*	32.13s	21.40e
Nzheleledam, *dam*	22.44s	30.06e
Olifants, *river*	23.58s	31.50e
Orange, *river*	28.41s	16.28e
Outeniqua Mountains	33.53s	22.35e
Pilanesberg, *mountain*	25.14s	27.04e
Pondoland *physical region*	31.10s	29.30e
Quoin Point	34.46s	19.37e
Richard's Bay	28.50s	32.02e
Riet, *river*	29.00s	23.54e
Roggeveldberge, *mountains*	32.17s	20.08e
Royal Natal National Park	28.45s	28.57e
Sabie, *river*	25.10s	32.02e
Saint Francis, Cape	34.14s	24.49e
Saint Lucia, Cape	28.25s	32.25e
Saint Lucia, Lake	28.05s	32.26e
Saint Lucia Game Reserve, *wildlife refuge*	28.10s	32.28e
Sak, *river*	30.52s	20.25e
Saldanha Bay	33.04s	18.00e
Sederberge, *mountains*	32.23s	19.20e
Sand, *river*	22.25s	30.05e
Sibaya, Lake	27.20s	32.40e
Sneeuberg, *mountains*	31.46s	24.20e
Sondags, *river*	33.44s	25.51e
Sout, *river*	31.35s	18.24e
Soutpansberg, *mountains*	22.55s	29.30e
Stormberg, *mountains*	31.27s	26.55e
Suurberge, *mountains*	33.18s	25.32e
Table Bay	33.53s	18.27e
Tembuland, *physical region*	31.30s	29.20e
Tina, *river*	31.18s	29.14e
Touws, *river*	33.45s	21.11e
Transkei, *historic region*	31.20s	29.00e
Tsomo, *river*	32.25s	27.50e
Tugela, *river*	29.14s	31.30e
Umfolozi, *river*	28.25s	32.26e
Umfolozi Game Reserve, *wildlife refuge*	28.19s	31.50e
Vaal, *river*	27.40s	26.09e
Valley of Desolation National Monument	32.17s	24.30e
Vals, *river*	27.23s	26.30e
Valsbaai, see False Bay		
Verneukpan, *salt lake*	29.58s	21.10e
Verwoerd Reservoir	30.40s	25.40e
Vis, *river*	33.30s	27.08e
Walvis Bay	22.59s	14.31e
Waterberge, *mountains*	20.30s	17.18e
Wild Coast	32.30s	29.00e
Wilge, *river*	27.03s	28.20e
Willem Pretorius Game Reserve, *wildlife refuge*	28.16s	27.13e
Winterberge, mountains	32.28s	26.15e
Witwatersrand, *physical region*	26.00s	27.00e
Zululand, *historic region*	28.10s	32.00e



Population density of South Africa.

members. Among the Africans, the independent African Christian churches have the largest membership, followed by the Methodist, Roman Catholic, and Anglican churches. Most of the Asians are Hindu, but many Muslim Malays live in and around Cape Town. There is also a Jewish minority living in South Africa. (A.Ne./D.F.G.)

THE ECONOMY

In the years since World War II, South Africa has had one of the most consistently high growth rates in the world. Its national income is one of the highest in Africa.

The bases of South Africa's economy consist of agriculture, mining, and manufacturing. Mineral resources are extensive. Apart from gold and diamonds, there are nearly 70 other types of exploitable minerals in South Africa. Whites have a well-developed agriculture that embraces a wide range of products. There is a large supply of cheap power; in addition, there are substantial reserves of cheap labour, there is no shortage of entrepreneurial talent, and there is ready access to supplies of domestic and foreign capital.

iamonds ıd gold

Resources. Diamonds and gold are the best-known and historically the most important minerals, but there are many others of growing importance, including copper, iron ore, manganese, asbestos, chromium, silver, beryllium, antimony, tin, and platinum, and some resources have yet to be exploited. The only important exclusion from the country's list of mineral resources is oil, for which an energetic search was begun in and around South Africa in the 1970s. Meanwhile, oil is made from coal, of which there are substantial deposits in Sasolburg.

Gold remains by far the most important mineral. A spectacular increase in the 1970s in the international price of gold allowed gold revenues in South Africa to grow dramatically, though prices subsequently declined. Gold accounts for more than half of total mineral sales. Gold is produced in three main areas—the Witwatersrand, which includes Johannesburg; the Far West Rand and Klerksdorp fields; and the Orange Free State around Odendaalsrus. The centre of the diamond industry is at Kimberley. There is also diamond production in the Orange Free State and the Transvaal. Known coal deposits are quite extensive, with most of the deposits located in the Transvaal. The biggest use of coal is for the generation of electricity. Many companies produce uranium, which is an increasingly important industry.

Agriculture is of major importance to South Africa, but the supply of good agricultural land is effectively curtailed by the climate. The average rainfall for the whole country is about 18 inches, which is less than the world's average and is very little for a subtropical country with a high evaporation rate. Runoff of water from rainfall is only 8 percent. Much of the country receives an even lower rainfall and is thus only marginally useful as agricultural land. Soil erosion is a problem in some areas, and the country is particularly subject to droughts.

Resources of timber are minimal; only a small fraction of the land area is forested. Conifers occupy a little more than half of the forested area; eucalyptus (used mainly for mining timber and pulpwood) covers almost a quarter; and wattle (acacia), which is used for producing tannin bark, as well as in the mining and pulping industries, accounts for about one-fifth. The forested areas are only found in the wetter regions of the south and east.

As a consequence of the low annual rainfall, natural hydroelectric potential is minimal. The government, however, has emphasized several schemes to develop hydroelectricity, the most important being the Orange River Project and the Cabora Bassa Project on the Zambezi River in Mozambique.

Agriculture, forestry, and fishing. Agricultural production, including forestry and fishing, accounts for a small percent of the gross domestic product. It remains important, however, since variations in its output can seriously affect overall growth in the economy and since this sector contributes a substantial proportion of exports. The agricultural sector is also of major importance in employment. White-, Coloured-, and Indian-owned farms are located in the more fertile and better-watered regions, which Africans are not allowed to farm. Among the major products are corn (maize), wheat, cane sugar, peanuts (groundnuts), millet and sorghum, citrus fruits, grapes, tobacco, wool, and meat. Sheep and goats, cattle, and pigs are the most important livestock raised for food. Dairy (including butter and cheese) and egg production are also important.

Major agricultural products

Compared to the vibrant commercial farming sector, the African subsistence sector—confined principally to the homelands areas, which generally lack fertility and are vastly overpopulated—has stagnated since the 1950s. Output has not kept up with population growth, creating pressures for rural Africans to seek work in white areas as migrant labourers. In no area of South African life is the disparity between white and African as visible as it is in agriculture.

Fishing areas lie chiefly off the south and west coasts and off the coast of Namibia. Both shoal fishing (for pilchard and maasbanker) and offshore trawling for a variety of fish, including kingklip, agulha sole, stockfish (Cape hake), and kabeljou, are employed.

Industry. The mining sector continues to be of major importance, and it figures prominently in government economic policy. Gold is the main product, after which bituminous coal is the most important. Next, in order of production, are diamonds, copper, iron ore, manganese, and asbestos. The dramatic increase in the price of gold in the 1970s gave new life to what had been considered a nearly exhausted industry. Gold seams that were previously not economical to mine were put into productive operation.

Manufacturing has been of increasing importance in the last few decades. Manufacturing has become the prime force in South Africa's rapid and sustained growth. Manufacturing also figures prominently in exports, contributing more than one-fourth to the total. This helps to offset the imported manufactured goods, which constitute the largest group of imports. An increasingly significant percentage of the population is employed in manufacturing.

One of the chief reasons for the rapid growth of the manufacturing sector has been the high proportion of the gross domestic product devoted to investment.

South Africa is a major producer and consumer of energy. Lacking its own oil resources, but with substantial reserves of coal, South Africa has been successful in its policy of energy self-sufficiency through a combination of conservation and effective use of its coal resources, including coal-to-oil conversion. Coal output has increased, much of it exported. The Electricity Supply Commission (Escom) takes most of the coal produced for local consumption. South Africa consumes a fraction of its energy in the form of electricity and produces about half of the electricity generated in Africa.

Oil consumption comprises a significant fraction of South Africa's energy. In 1973 the Organization of Petroleum Exporting Countries (OPEC) declared an oil embargo against South Africa, but South Africa continued to receive 90 percent of its oil imports from Iran until 1979. With its sources of oil gone, South Africa is turning increasingly to alternative energy sources.

Finance. South Africa has a well-developed financial system. The central bank—the South African Reserve Bank—is the sole note-issuing authority. It also exercises all the functions necessary for the implementation of monetary policy. There are nine commercial banks, which have strong ties to foreign banks. The banking system also includes a number of merchant banks and discount houses. There are also two specialist state banks—the National Finance Corporation, which was opened in 1949 to develop a short-term money market, and the Land Bank, which, as its name implies, lends chiefly to agriculture.

Second in financial strength only to the commercial banks are the building societies. Installment buying, savings, and general banking are also important, as are private pension and provident funds and insurance companies. There is an active capital market organized around the Johannesburg Stock Exchange.

Trade. Dependence upon foreign trade is relatively high, and the South African economy is thus particularly sensitive to global economic conditions. Precious metals and base metals have been leading exports. Agricultural goods also play an important role in South Africa's exports. The country's major imports are oil, machinery, electrical equipment, and transportation equipment. A number of imports, most importantly oil, have been classified by the government as "strategic materials." Data for these are not available, making it difficult to assess the exact percentages of various import categories. South Africa's main trading partners are the United States, the United Kingdom, West Germany, Switzerland, and Japan. These five countries account for two-thirds of the country's trade. South Africa has moved to diversify its trading partners, increasing its trade with Israel, Taiwan, South Korea, several Latin-American countries, and other African countries.

Trading partners

Administration of the economy. The economy is essentially based on private enterprise, but the state participates in a number of ways. Through the Industrial Development Corporation of South Africa, it controls several public corporations—including the South African Coal, Oil and Gas Corporation (Sasol), which produces oil from coal; the Phosphate Development Corporation (Foskor), which develops phosphates; the African Metals Corporation Ltd. (Amcor), which makes metals; the Iron and Steel Industrial Corporation (Iscor), the major steel producers; and Escom, the main electricity-generating authority. The government also runs the railways, the national airline, harbours, telecommunications, and much of the road transport system. The government sets out targets for development in each industry in an Economic Development Programme, and, by means of a range of official bodies, encourages development of industry through financing, consultations on tariff protection, export promotion, export credit insurance, research, and a bureau of standards.

Direct taxes on individuals consist chiefly of state and provincial income taxes deducted at the source and taxes on dividends and interest. There is no capital gains tax. The main taxes on businesses are income tax and, in some cases, a tax on undistributed profits.

Trade unions for African workers have become a critical issue in South Africa, and there were two periods (1973–74 and 1979) of large-scale work stoppages by African workers. Although African workers had traditionally been denied the right to form unions and participate through them in collective-bargaining efforts, many African trade unions formed. The most important African trade union federations are the Federation of South African Trade Unions (FOSATU); the Trade Union Council of South Africa (TUCSA), with a multiracial membership; and the South African Congress of Trade Unions (SACTU), many of whose leaders are outside of the country. Most organized white workers are members of the South African Confederation of Labour.

The all-white South African Confederation of Labour has opposed the easing of restrictions on African trade unions, and a major point of contention between white unions and nonwhite workers is the practice of reserving the highest skilled jobs for whites only.

The main aim of economic policy has been to maintain the high rate of economic growth. This growth is built upon a base of cheap African labour in low- and middle-level jobs and is dependent largely on the maintenance of a high level of capital investment, a healthy balance of payments, an adequate supply of skilled labour, and the effective control of inflation. In most of these fields considerable success has been achieved, but the labour situation has become increasingly serious. The problem is aggravated by the government's policy of trying to reverse the flow of Africans to the big cities by locating new industry near the designated African areas and by the white trade unions' unwillingness to admit nonwhites to certain job categories. A range of skilled and semiskilled jobs are reserved for whites, but there is an inadequate supply of skilled white labour to fill vacancies in these job areas, and immigration from Europe has proved insufficient to fill the gap. (E.I.U./D.F.G.)

The labou shortage

Transportation. Because South African rivers are unsuitable for navigation, water transport within the country, except for traffic along the coast, is nonexistent. Railroads and roads have been built from the harbours to points of economic and population concentration in the interior.

The South African railway system is state property. The standard gauge, with a few minor exceptions in the Cape and Natal coastal areas, is three feet six inches, and approximately one-third of the republic's railroads are electrified.

South African Railways is the greatest single employer in the country. Of the total quantity of goods transported, coal alone makes up more than one-fourth, other minerals combined accounting for one-third, and agricultural products accounting for one-eighth.

The railroad system is augmented by a network of more than 142,600 miles of roads, of which about 48,700 miles are paved, including more than 1,000 miles of national roads connecting the country's main urban centres.

All South African harbours are state-owned and, together with railways and airways, are administered by the Ministry of Transport. The shipping of the country is served mainly by South Africa's four major ports—Durban, Cape Town, Port Elizabeth, and East London—as well as by the port of Maputo (formerly Lourenço Marques) in neighbouring Mozambique. Durban, which serves the southern Transvaal hinterland, the economic heartland of South Africa, is the main cargo port. Cape Town, the large port nearest to Europe, is the main passenger port. A large new harbour has been developed at Richard's Bay, on the Indian Ocean in Natal, to accommodate giant tankers.

Air servic

Inland air transport services, both passenger and freight, are operated by state-owned South African Airways and by smaller, local private companies. South African Airways also maintains regional services to such neighbouring countries as Lesotho, Botswana, Zimbabwe, Namibia, Mauritius, and Mozambique and flies international routes to Europe, Australia, and New York City via Rio de Janeiro. The air route used by South African Airways to reach European cities must avoid most of the remainder of the African continent because of the antagonism of black African states toward South Africa. Las Palmas, in the Canary Islands, and Ilha do Sal, in Cape Verde, are used as intermediate stopovers. Major foreign air services also are in operation to and from South Africa.

The three international airports are Jan Smuts Airport near Johannesburg, D.F. Malan Airport at Cape Town, and Louis Botha Airport at Durban. Other major domestic airports are the J.B.M. Hertzog Airport at Bloemfontein and those at Port Elizabeth, Kimberley, and East London. (A.Ne./D.F.G.)

ADMINISTRATIVE AND SOCIAL CONDITIONS

Government. The republic's original constitution, the South Africa Act of 1909, formed a parliamentary system with the British monarch as head of state. The constitution was revised by the Republic of South Africa Constitution Act of 1961, narrowly approved by the all-white electorate, which transformed South Africa from a monarchy within the Commonwealth of Nations to an independent republic with a president as head of state. This post, like that of the British monarch it had replaced, was essentially honorific, and executive power was held by the prime minister. In 1983 white voters approved a new constitution, enacted the following year, that abolished the post of prime minister and transferred its power to the state president. While the new constitution extended the franchise to Coloureds and Asians, it also introduced a distinction between "general affairs," those pertaining to all racial groups, and "own affairs," those particular to a racial group, that guided the formation of separate government bodies for whites, Coloureds, and Asians. In addition to a strengthened state president, the constitution provides for a cabinet, which includes ministers who head departments of general concern, such as defense; a tricameral Parliament, in which whites, Coloureds, and Asians each have their own chamber; three Ministers' Councils, one drawn from each chamber of Parliament, in which members head departments, such as those for social welfare and education, that administer only for their own racial group; and a President's Council, which settles disagreements between the legislative chambers. Africans continued to be excluded from the national government.

The state president is elected for a term of a maximum of five years by secret ballot of an electoral college consisting of 50 white, 25 Coloured, and 13 Asian members of Parliament designated by their respective chambers. The president appoints ministers to the cabinet, which may include persons outside of Parliament, and to the Ministers' Councils, in which ministers must be members of Parliament or become members within a year of appointment. There is a close working relationship between the cabinet and the Ministers' Councils, and ministers may serve on both. The president, the dominant figure in the South African government, generally wields more power than a U.S. president or a British prime minister.

ıree
ambers
ırliament

Originally a bicameral legislature, Parliament was reorganized as a unicameral body in 1980 and expanded in 1984 to three chambers: the existing House of Assembly, with 178 seats, for whites; the House of Representatives, with 85 seats, for Coloureds; and the House of Delegates, with 45 seats, for Asians. Each house is solely responsible for legislation pertaining to its own racial group, but matters that the president determines to be of general concern must be approved by all three chambers. If agreement is not reached between the chambers, the president may direct the President's Council to make a final decision. The President's Council, which is also an advisory body, is composed of 20 whites, 10 Coloureds, and 5 Asians elected by the respective chambers of Parliament and 25 members appointed by the president.

Members of Parliament are directly elected in single-member constituencies by voters of the racial group they represent, though a small number of legislators are appointed by the president or elected by their respective chamber of Parliament. Members must be South African citizens of white, Coloured, or Asian descent, be at least 18 years old, and have resided in the country for at least five years. Seats for the elected members are allotted to the provinces in proportion to the adult population of each racial group; the electoral boundaries for these seats are delimited by a presidentially appointed judicial commission at intervals of no less than five and no more than 10 years. While members are elected for a five-year period, Parliament can be dissolved by the state president at any time. (Ed.)

Regional, state, and local governments. Where the central government's powers are delegated to regional and local governments, such delegation has been effected within the general framework of a separate administration for each racial group (whites, Coloureds, Asians, and Africans).

The provincial governments began a period of transition after the passage of the 1983 constitution. Each province formerly had a provincial council elected by white voters only, an executive committee of the council, and an administrator appointed by the state president, but in 1986 the councils and the executive committees were abolished and their powers were transferred either to the separate administrations of the three chambers of Parliament or to new executive authorities appointed by the state president and headed by the administrator. By the late 1980s a small number of Coloureds and Asians had been appointed to these authorities.

Each province also had its own system of local authorities, elected by white voters only. Coloureds and Asians had been allowed only advisory positions, while the Bantu Affairs Administration Act of 1971 had completed the process of removing all major aspects of government and administration of Africans from local authorities in white areas to the Department of Bantu Administration and Development, which later became the Department of Development Aid. In the early 1990s local governments were also in a period of transition, and some form of representation for Coloureds, Asians, and Africans was being developed.

The Promotion of Bantu Self-Government Act of 1959 recognized eight African national units, designated Bantu Homelands, that were based upon government-designated tribes. Two others later were created, and the official designation was changed to Black States. Except for Transkei, their areas bear little resemblance to the original areas of African settlement. All but two consist of several land blocks, together comprising more than 200 scattered enclaves. In terms of the official policy of apartheid, all 10 are to evolve into independent republics, although there is no intention of consolidating each into a single geographic entity. The Bantu Homelands Citizenship Act of 1970 made every African a "citizen" of one of these national units, even if he or she did not live or had never lived in the area.

The goal of the separate development policy has been to resolve South Africa's racial conflict by eliminating Africans from the body politic. By the end of the 1980s only four of the Black States had been granted independence as republics: Transkei, October 26, 1976; Bophuthatswana, December 6, 1977; Venda, September 13, 1979; and Ciskei, December 4, 1981. Their governments operate under semiparliamentary systems, with cabinets nominally responsible to partially elected and partially appointed legislative assemblies. While the African areas possess most of the abstract criteria for sovereignty—established boundaries, a settled population, and a functioning government—they have not been recognized by any other government or by any international organization. They also lack potential for economic autonomy, due to their overwhelming economic ties to South Africa. In addition, a wide range of technical, defense, and economic agreements give the South African government powerful influence over the internal affairs of the areas. (M.J.L./D.F.G./Ed.)

The political process. No special qualifications are required for the white, Coloured, and Asian franchise. The voting age is 18 years.

The
National
Party

The all-white National Party has been the dominant national and parliamentary party since it came to power in 1948. Its programs have emphasized the concepts of white South African nationalism and the implementation of the separate-development strategy for dealing with South Africa's racial problems. Other major white parties are the Democratic Party (formed in 1989 of three liberal parties, including the former Progressive Federal Party) and, on the right, the Herstigte Nasionale ("Reconstituted National") Party and the Conservative Party (an offshoot of the National Party formed in 1982). (R.B.Ba./D.F.G.)

South Africa's most significant political developments have stemmed from the implementation of the separate-development philosophy and the response of the nonwhite population. Separate development systematically deprives the nonwhite population of political participation in the conduct of national and provincial affairs in exchange for the establishment of their own political institutions. The last African representation in Parliament (provided by three elected whites) was abolished in 1959 under the Promotion of Bantu Self-Government Act. Prior to the establishment of separate parliamentary chambers for whites, Coloureds, and Asians, the last Coloured representation in Parliament had been in the 1960s, but those four seats were abolished by the Separate Representation of Voters Amendment Act of 1968. The Asian population had never been represented. The Prohibition of Political Interference Act of 1968, which, until its repeal in 1985, outlawed multiracial political parties, forced the Progressive Party to limit its membership to whites; the Liberal Party disbanded rather than do so.

In the 1980s successful boycotts of parliamentary elections resulted in low voter turnout among Coloureds and Asians. Nevertheless, within Parliament the Labour Party and the People's Congress Party emerged as major Coloured parties. Significant Asian parties include the National People's Party and the Solidarity Party.

Black Consciousness movement

The two major national African political parties—the African National Congress and the Pan-Africanist Congress—were banned in 1960 under the Unlawful Organisations Act. In 1990, however, they were legalized, as was the multiracial South African Communist Party, one of the largest Communist parties in Africa. In the 1970s two different forms of political participation emerged among the nonwhite population. On the one hand, the implementation of apartheid led to the creation of a number of political parties in the African areas, oriented toward participation in the new institutions of self-government. On the other hand, especially in the urban areas, many Africans responded to enforced segregation by beginning to assert their racial identity. This tendency, which became known as the Black Consciousness movement, was expressed in such organizations as the South African Students Organisation (SASO) and the Black People's Convention (BPC). The widespread influence of Black Consciousness, especially among young urban Africans, has been demonstrated in periodic unrest, one of the most spectacular examples being the Soweto uprising in June 1976.

In the 1970s the church also emerged as a significant focal point for the expression of the political aspirations of nonwhite South Africans. The South African Council of Churches in particular has played a major role in opposing the policy of separate development. In the period following 1976, there were growing numbers of armed attacks on police stations, banks, and other public facilities, carried out predominantly by the underground guerrilla movement Umkhonto (Umkonto) we Sizwe (Spear of the Nation), a wing of the African National Congress.

Justice. The common law of the republic is based on Roman-Dutch law, the uncodified law of Holland retained after the Cape's cession to the United Kingdom in 1814. The judiciary comprises a Supreme Court with local, provincial, and appellate divisions and the lower courts. The judges, headed by a chief justice, are appointed by the state president. The highest court is the appellate division, which has as many judges as the president chooses. The appellate division hears the appeals of local and provincial divisions, and its decisions are binding on all courts.

The lower courts comprise magistrates' courts and the African court systems of chiefs' courts and commissioners' courts. Most civil and criminal litigation is a matter for the magistrates' courts, each responsible for a magisterial district. Regional magistrates are appointed in populous areas with criminal jurisdiction over a number of districts. The African courts do not have jurisdiction over Africans in criminal cases or where whites are involved. In Transkei, Bophuthatswana, Venda, and Ciskei, courts are under the jurisdiction of their governments, but decisions can be appealed to the appellate division of the Supreme Court of South Africa. (D.F.G./R.B.Ba./Ed.)

Nonwhites in the military

Armed forces. Although the active forces are relatively small, there is a large trained reserve force. South Africa's military traditionally has been white, and there is compulsory military service only for white males. Because of the shortage of manpower, the decision was made in the 1970s to use substantially larger numbers of Coloureds and Asians and to begin to include African soldiers in combatant roles. White women were taken into the services in 1970; their training has been confined to exercises in unarmed combat and the use of light firearms.

The navy has a small fleet consisting of destroyers, submarines, antisubmarine frigates, minesweepers, and auxiliary craft. The air force combat aircraft include light bombers; fighter bombers; interceptor fighters; transport, reconnaissance, and training aircraft; and helicopters.

Police activities are coordinated with those of the defense forces. The police, almost half of whom are nonwhites, bear the responsibility of maintaining internal security. A substantial number have received special training in antiguerrilla activity and have been used to counter guerrilla infiltration in several border areas. South African police assisted the former Rhodesian government in its antiguerrilla activity in the 1960s and 1970s.

The military has played an increasingly important political role in South Africa. In the late 1970s military figures were responsible for the concept of a "total strategy" for South Africa to meet the perils of an increasingly hostile outside world and the threat of domestic turmoil. Part of this strategy has concentrated on nuclear development, considered particularly important in light of South Africa's lack of domestic petroleum sources. The South African Atomic Energy Board (superseded in 1982 by the Atomic Energy Corporation) oversaw the development of two nuclear reactors at Pelindaba, near Pretoria. In 1975 a new plant at Pelindaba began enriching uranium. South Africa has been provided with equipment and training by the United States, West Germany, and France. South Africa has about 17 percent of the world's uranium reserves and has the third largest output among Western nations. It has not signed the Nuclear Non-Proliferation Treaty.

Education. General education issues, such as the establishment of educational standards, are the responsibility of the Department of National Education, headed by a member of the cabinet. Each chamber of Parliament has a Department of Education and Culture, headed by a member of the respective Ministers' Council, that is responsible for providing education to its own racial group. African education at all levels is administered by the Department of Education and Training (formerly the Department of Bantu Education). The 10 African territories each have their own departments of education.

The universiti

For whites there are four English-language universities, five Afrikaans-language universities, and one bilingual university; for nonwhites there are four (including the Medical University of Southern Africa) for Africans, one for Coloureds, and one for Asians. Students who want to attend a university designated for a different racial group must obtain government permission. White schoolchildren are taught in a single language—English or Afrikaans according to their home language. Attendance is free and compulsory between seven and 16 years of age, and books and equipment are provided. For Asians and Coloureds education is free and compulsory where the facilities exist.

Schooling for African children is not free, nor is it compulsory for most. Parents are responsible for some of the books and equipment. The number of Africans who receive some schooling has increased dramatically since the early 1950s; by the late 1970s more than three-fourths of school-age African children were receiving some education. The education available to Africans, however, is far inferior to that available to whites, and the dropout rates in African schools are very high. There is a serious shortage of teachers because the government provides very limited resources for African teacher training. In addition, many of those who teach Africans are not well trained. The per capita government expenditure for whites and Africans enrolled in schools varies drastically.

Health and welfare. The administration of health and welfare is also formed in the context of the government's distinction between general concerns and those of a particular racial group. Formulation and overall coordination of the national health and welfare policy, a general concern, is conducted by the Department of National Health and Population Development, though the policies and standards must be approved by all three chambers of Parliament. The services themselves are not considered a general concern and thus are managed separately for whites, Coloureds, and Asians by the departments of Health Services and Welfare of the three chambers. For Africans, national welfare policy is set by the Department of Development Planning, and health and welfare services are provided by the four provincial administrations. For all four racial groups some powers are variously but separately delegated to regional or local bodies.

Public hospitals and clinics are available free or partly free to the indigent. Voluntary organizations that provide welfare services must register with the government, which subsidizes some of these organizations.

Local authorities have been responsible for public housing, which has been segregated by race and controlled by different departments. The Department of Community Development was established in 1961 to oversee a national policy. By 1970, 50 percent of new housing built by local authorities was being used for resettling Coloured and Asian communities removed from one area to another under the Group Areas Act of 1950; the consequence is a severe shortage of housing for these groups. The Department of Co-operation and Development initiated a program of township development in the African areas during the 1960s, mainly in response to the need to provide housing for those Africans who were removed from white areas in the implementation of the separate-development policy. With few exceptions, Africans have not been allowed to own their own housing outside of these areas.

The regular police are organized nationally and comprise regulars as well as reservists. There are about equal numbers of whites and nonwhites. The Security Police are organized separately. A third police force was established in 1969 in terms of the Public Service Amendment Act and the General Law Amendment Act. These acts established a Bureau of State Security, responsible only to the prime minister and not subject to treasury approval for its spending. The bureau was reorganized in 1978 as the National Intelligence Service. It is an offense in South Africa to possess or publish information about military, security, or police matters deemed prejudicial to state security; in addition, any government minister may issue a certificate preventing evidence from being presented at court.

Many people have served sentences under the various security laws. The number detained without trial is not given officially, but it has been estimated to average several hundred at any given time. The prison population in South Africa increased sharply during the 1960s and '70s. In 1977 the daily average prison population was almost 100,000, one of the highest rates in the world. Of these, the majority were imprisonments for statutory offenses against laws (the so-called pass laws) that define the right of Africans to live and work in white areas and that do not apply to other racial groups.

age dif-
entials

The most obvious feature of social conditions is the difference between the wages and cost of living of whites and of nonwhites. In addition, the vast majority of Africans have remained untouched by the wage advances in the mining and industrial sectors of the economy. In the late 1980s more than one-fifth of the African population was unemployed, and those who were employed were generally in the lowest paying and least prestigious positions.

This pattern of wage differentials partially reflects the composition of South Africa's black population, with its many migrants to industrial and urban areas, but it also is a result of the conscious and systematic effort of the government to control black labour and channel it to the benefit of whites. The Industrial Conciliation Act of 1956, for example, allows for a periodic definition of the levels of employment that members of each racial group may occupy and thus maintains the racial hierarchy in employment. The consequent wage differentials result in income levels below the official poverty line. In the African towns of Soweto, surveys in the 1970s showed that 75 percent of the families lived below the poverty line.

White South Africa has a highly sophisticated public health system, and, while racial bias is not explicitly written into health legislation, health care for nonwhites is invariably affected by the economic and political inequalities of the society. The health status of Africans is generally low. Malnutrition is one particularly important example of this, especially among children. Kwashiorkor, a condition resulting from insufficient protein in the diet, is common among Africans in both urban and rural areas. There is an enormous discrepancy between infant mortality rates; the mortality rate is lowest for whites, then substantially higher for urban Africans, with the highest rate of infant mortality among rural Africans. A similar discrepancy exists in the number of doctors available for whites as compared with Africans. In the late 1970s the South African government opened a facility to train African doctors and other medical workers.

The apartheid policy

It is the purpose of apartheid, the separate-development policy, to segregate the races as far as possible in every aspect of their lives. The country is divided into one white area and 10 African areas. The white area is divided by the 1950 Group Areas Act (as amended) into segregated areas in each of which only members of a designated race may live, own land, or occupy premises. In addition to residential segregation, there is political, social, and economic discrimination, though some apartheid laws have been rescinded. Marriages and sexual relations between the races had been outlawed by both the Prohibition of Mixed Marriages Act of 1949, which was repealed in 1985, and the Immorality Amendment Act of 1950, major portions of which were also repealed in 1985. The Reservation of Separate Amenities Act of 1953, repealed in 1990, forbade persons of different races to use the same public amenities or transport facilities and also gave central or local authorities power to disallow racial mixing in places of public entertainment, cultural institutions, and sporting facilities. These changes, however, have not significantly affected the fundamental economic, social, and political realities of separate development. (M.J.L./D.F.G./Ed.)

CULTURAL LIFE

The 10 or more languages spoken by South Africa's people reflect the country's cultural pluralism. The largest group, the Africans, is the most eclectic in culture. Many have adopted a host of European ways, but the solid core of African cultural traditions of language, music, and dance retains its vitality. The various African societies have rich oral traditions, including imaginative narrative, poetic, historical, and epic forms. These oral traditions, while they remain strong in their own right, especially in rural parts of the country, have exerted a strong influence on the written literatures of the African cultures, which also have been influenced by literary traditions of other parts of the African continent, of the Caribbean and the Americas, and of Europe.

Such writers as Guybon Sinxo (Xhosa), B.W. Vilakazi (Zulu), Oliver Kgadime Matsepe (Northern Sotho), and Thomas Mofolo (Southern Sotho) have been more heavily influenced in their written work by the oral traditions of their cultures than by European forms. Works composed in the indigenous languages have been largely ignored or dismissed as works written for the schools or as works with limited audiences, yet these novels and poems have for years been a primary means of expression for African intellectuals. No matter what the size and composition of the audiences, the works are a significant part of the intellectual history in South Africa.

During the 1970s there emerged in the arts powerful themes of national and multiracial, multilingual cultural patterns, as writers and artists from all groups concentrated on exploring and portraying the crises affecting South African society. Reaction to apartheid and various earlier racial policies had engendered a sense of black culture and history that anticipated negritude as it was manifested in West African, Caribbean, and U.S. movements. The themes of Black Consciousness evident in the poetry and prose of urban writers such as Mothobi Mutloatse and Miriam Tlali, and published in such periodicals as *Staffwriter,* are derived from the literary and oral traditions of African languages in South Africa and in literature by Africans in European languages. The authors Adam Small and Alex La Guma have written vividly in Afrikaans and English, respectively, of the effects of racial discrimination and of the complex and frequently violent nature of relations between the races. Because of the growing international concern with South Africa, many of these black writers, and those white authors who deal with racial themes, have received recognition beyond the borders of the country. Newly recognized writers such as John Coetzee, Sipho Sepamla, and Walli Serote have joined such established figures as Ezekiel Mphahlele, Nadine Gordimer, Alan Paton, André Brink, and Étienne Leroux in bringing South African literary life to the wider world.

Censorship of published works

All published works in South Africa are subject to severe censorship, and many works are banned, especially those with strong political themes or descriptive sexual scenes. A large segment of Afrikaner writers have become alienated from the government; and authors such as Breyten Breytenbach, Dan Roodt, Brink, and Leroux have had their works banned. South African playwrights have responded to the new cultural and political milieu with such innovations as bilingual and multilingual plays and the use of ad-libbing to avoid censorship. Plays by Athol Fugard, Fatima Dike, and Pieter Dirkhuys have been performed worldwide.

The major institutional support for culture comes from the four provincial councils for the performing arts. These councils help fund plays, operas, symphonies, and other events. In the late 1970s the activities of the councils, which are subsidized by the government, were open to South Africans of all racial groups.

The press in contemporary South Africa finds itself increasingly under political and legal constraints. Historically, the strongest elements of the press have been the English-language combines (such as the Argus group). The Afrikaans-language press has come to be more influential, sometimes through government financial support. African readership has also expanded greatly. In recent years several African newspapers have been closed, and individual journalists banned.

Television—introduced in the mid-1970s—and radio constitute important forces in South African society. The government tightly controls both media, which air programs for the different linguistic and cultural groups in the country. The government uses television and radio to communicate its own views and to counter cultural tendencies that it perceives to be threatening to the implementation of separate development. For statistical data, see the "Britannica World Data" section in the current *Britannica Book of the Year.*

(R.V./D.F.G.)

History

For the history of the region before the arrival of European settlers, see SOUTHERN AFRICA.

Whether or not the Phoenicians sailed around Africa as claimed by Herodotus, southern Africa was unknown to the Greeks and the Romans and to Europeans before the 15th century. Then, largely as a result of the patronage of Prince Henry the Navigator, Portuguese mariners steadily pushed down the West African coast until in 1488 Bartolomeu Dias de Novais rounded the Cape of Good Hope and in 1498 Vasco da Gama, in a flagship of 120 tons, 80 feet long, reached the objective, India. Thereafter the Portuguese gained maritime supremacy in the Indian Ocean and used it to exploit the Asiatic spice trade and, less effectively, the East African gold trade. But, though they colonized Angola and Mozambique, they made no settlements as far south as what is now the Republic of South Africa, regarding its resources as negligible and its inhabitants as too primitive for commercial or evangelical effort. Throughout the 16th century, therefore, though South African waters were often traversed by Portuguese fleets, the only civilized people who spent much time ashore were the survivors of the many ships that were wrecked on the coast.

DUTCH FOUNDATIONS

Trading companies. By the end of the 16th century Dutch and English seamen had also begun to trade with Asia by the Cape route. They found it desirable to pause for refreshment (the death rate on the long voyage being often as high as 50 percent), and during the first half of the 17th century they usually stopped at St. Helena, the Cape, or Mauritius. In 1615 the English tried in vain to form a settlement at the Cape with 10 felons reprieved from the hangman's noose, and five years later two English captains made the empty gesture of annexing all Africa for James I; but it was St. Helena that eventually became the main port of call for the English company.

Dutch outpost at the Cape

Meanwhile the Dutch East India Company outstripped all rivals and obtained a virtual monopoly of the East Indian spice trade. For several decades Dutch ships made a practice of putting into Table Bay to take in fresh water and to barter with the Khoikhoin for meat; and to find the latest information about the affairs of Europe or Asia in dispatches they left one another under inscribed stones. Then in 1647 Leendert Janssen was shipwrecked in Table Bay with the crew of the "Noord Haarlem," and after his return to the Netherlands he presented a report advising the company to found a permanent refreshment station at the Cape. Largely because the Anglo-Dutch War of 1652–54 was imminent, the directors eventually agreed; three ships were to take out building materials, seed, and implements, and 90 men were to build a fort and develop a vegetable garden.

Jan van Riebeeck. To command the expedition the directors chose Jan van Riebeeck, a doctor by training, who had already had six years of commercial experience with the company. He landed at the Cape on April 7, 1652, and remained 10 years, during which time droughts, floods, and pests, indiscipline among his subordinates, and thefts by the Khoikhoin often brought the settlement to the brink of disaster. But it survived; a timber-and-sod fort was built, the coastline was charted, the interior was explored, and a belt of land extending about eight miles from Table Bay to Wynberg was brought under cultivation.

However, the Cape did not become the simple outpost the directors had had in mind. At that time Khoikhoin, pastoral nomads loosely organized in small tribes, occupied most of the coastal regions of South Africa as far east as the Great Fish River. The Cape Peninsula had been used by three such tribes, who regarded the Dutch as intruders in their domain, especially when it became apparent that the Dutch intended to stay when they began building fortifications and extending their area of cultivation. The Dutch bartered with the Khoikhoin for sheep and cattle, and from time to time the Khoikhoin stole them back again. Van Riebeeck calculated whether he was strong enough to use force, but when the directors cautioned him to do no such thing he fell back on the idea of confining the settlement within a palisade. Yet even at this early stage in South African history it was not possible to keep the races completely apart, because the Dutch continued to barter with the natives.

Importation of slaves

Van Riebeeck also initiated two experiments that had far-reaching consequences. At first all the work of the station was done by the company's servants, but in 1657 a few slaves were imported and more came in 1658 and thereafter. The second experiment started in 1657 when the company, hoping it would save money if grain were grown and stock bred by unpaid freemen instead of by paid employees, freed nine married men from their contracts and gave them 30-acre farms along the Liesbeeck Valley at Rondebosch. These men, who were known as free burghers, were to produce grain and stock and sell them to the company at agreed prices. So van Riebeeck left behind him an embryonic colony of settlement which used slave labour and was confronted with a "native problem."

Immigration and expansion. In 1662 the Cape settlement numbered about 250 white persons, of whom nearly half were company servants. In the next few years the population increased very slowly. But during the governorships of Simon van der Stel (1679–99) and his son W.A. van der Stel (1699–1707) the company made its one serious effort to encourage immigration. In 1707 a census showed a burgher population of 1,779 men, women, and children, owning 1,107 slaves. This was a slow rate of progress when compared with that of the English colonies in North America.

Nevertheless, the modern Afrikaner people are descended for the most part from those who were enumerated in

1707. Most of them were of Dutch and German stock. There was also a small but significant group of fewer than 200 French Huguenots who, after the revocation of the Edict of Nantes in 1685, went to the Netherlands and then were sent to the Cape. As a matter of policy the company scattered the Huguenots among the other settlers, and within little more than a generation they became fully assimilated, discarding the French language and even pronouncing their names in the Dutch manner.

ssimi-
ion of
ench
uguenot
ttlers

The area of settlement expanded with the population—to Stellenbosch beyond the sandy Cape Flats (1679) and then to the Berg River Valley. To meet the needs of local administration a general purposes official (*landdrost*) was appointed to Stellenbosch in 1685; sitting with four local burghers (*heemraden*), he was given jurisdiction in minor civil cases. Nearly all the company servants, however, lived in Cape Town, where the governor and his chief subordinates formed a Council of Policy which issued regulations for the colony, subject to the overriding authority of the governor general and council of the Indies in Batavia and the directors in the Netherlands.

From the first the burghers had fretted at many of the regulations, which imposed restrictions and taxes on internal trade. What was particularly galling to the burghers was the tendency of the officials to farm and trade on their own account, using their authority to make sure that their own produce did not remain unsold. This led to an outcry against the governor, W.A. van der Stel, who embarked on large-scale farming operations and cornered the market so successfully that some malcontents, including the Huguenot diarist Adam Tas, sent a petition to the directors, who recalled van der Stel in 1707. But the underlying cause of the discontent—the commercial policy of the company—was not remedied.

Evolution of the Afrikaner people. The company, probably convinced that white settlement led to political agitation, had no vision of a New Netherlands in South Africa and sponsored no immigration schemes after the time of the van der Stels. Nevertheless, during the 18th century a New Netherlands of a sort was arising from the materials present. The white population grew, mainly by natural increase, to about 15,000 in 1795, by which time they regarded South Africa as their only home and spoke a language, Afrikaans, which deviated from the language of the Netherlands. Those who lived in Cape Town—the only port and the only real town in the colony—were mostly traders and innkeepers. In the southwestern part of the colony, near Cape Town, were the grain and wine farmers, some of them proprietors of large estates run by slave and Khoi labour. Beyond them were the Boers, about half the total white population.

The Boers were seminomadic, many of them completely nomadic, pastoral farmers. Most of the land they occupied was unsuitable for cultivation; moreover, the market for agricultural produce was glutted, and slaves did the skilled work of the colony. There was, therefore, nothing to keep the increasing white population in the southwestern Cape and much to encourage them to disperse and become self-sufficient pastoral farmers.

The way was paved by W.A. van der Stel, who, unlike his father, allowed burghers to concentrate on stock farming. Within a few years it became possible for any burgher to take occupation of 6,000 acres or more of land and treat it as his property, paying the company a small annual license. The result was a continuous expansion of the frontiers of white settlement. In the north the Boers were slowed up by aridity and finally stopped near the Buffels River about 1760. In the east they advanced along either side of the arid Great Karoo until in the 1770s the two streams of pioneers began to mingle beyond it in the vicinity of what became Graaff-Reinet, about 400 miles (640 kilometres) from Cape Town.

pansion
white
tlement

Thus a few thousand white persons scattered over an area about twice as large as England and about the size of Colorado. Their simple needs were met by their stock, by hunting, and by travelling traders who bought their sheep and cattle for the Cape Town market and sold them a few imported goods. The company did little for them beyond establishing courts of *landdrosts* and *heemraden* at Swellendam (1745) and Graaff-Reinet (1785). From time to time officials expressed concern that the isolation of the Boers was causing them to degenerate. But although all the Boers were extreme individualists and some were ruffians, most of them contrived to preserve a sound domestic morality, a smattering of literacy, and a respect for their religion, a rather strict Calvinism.

Slaves, Khoikhoin, and San. During the 18th century the slaves, the Khoikhoin, and some of the San were beginning to interbreed with one another and with Europeans to produce the community now known as the Cape Coloured people. Although few slaves were imported, the slave population outstripped the white to number nearly 17,000 in 1795. Negroid types from the west and east coasts of Africa predominated, with Malays from the East Indies forming an important minority. There was a considerable amount of extramarital miscegenation between white men and slave women. Most of the slaves were the property of the agricultural farmers, for whom they formed the main labour supply. Others were owned by the company and the Cape Town burghers, a few by the Boers. The treatment of the slaves, which had at first been fairly mild, tended to deteriorate, and the 1754 slave code was a stiff one.

The nearest approach to a Khoi war had taken place in the 1670s. But it was smallpox that took the heaviest toll of the Khoikhoin, especially a shattering epidemic in 1713. After that, as the Boers advanced into their territories, the surviving Khoikhoin lost their land and their cattle—their only means of an independent existence. Many were reduced to a status not unlike serfdom, forming the main labour supply of the Boers, whose simple demands they were able to satisfy. Others withdrew beyond the frontiers of white settlement to maintain a more or less independent existence until the next century.

There was a certain amount of miscegenation between white men and Khoi women, although intermarriages had been unlawful since 1685. The offspring were usually disowned by their fathers and became known, bluntly, as Bastaards; some entered into service, others retreated before the advancing Boers, leading Boer lives themselves.

The small San were still more primitive than the Khoikhoin. Before white settlement began they had been driven into the more arid and mountainous parts of South Africa by the Khoikhoin and other African tribes. But, though every man's hand was against them, they contrived to make hit-and-run raids to get sheep and cattle for food, and, if cornered, they were able to kill their enemies with their poisoned arrows. As the Boers in turn began to enter San country and to suffer from their raids they organized Boer commandos to hunt them. Toward the end of the 18th century several San bands made a prolonged stand, operating from mountain fastnesses along the northern fringes of the area of white settlement; but the commandos gradually cleared them out, killing 2,500 San and capturing 650 in the decade ending in 1795. Those captured were children, who were taken into service.

San
resistance
to Boer
incursions

Their experience with slaves, Khoikhoin, and San had important and enduring effects on the racial attitudes of the Afrikaner people. Most of the nonwhites they came into contact with were at so low a cultural level that the Afrikaners tended to think of themselves as innately superior. Their response was not merely hostile, however, because they all—Capetonians, agriculturists, and Boers—came to depend on the use of nonwhite servants.

African opposition and economic discontent. Before the last quarter of the 18th century the Boers had not been seriously impeded by human obstacles. Then, however, they came up against formidable opposition.

Bantu-speaking tribes had occupied the the eastern half of South Africa for several centuries. They were more numerous than the Khoikhoin. They had a more complex social and political structure. They cultivated the soil as well as hunting game and herding cattle. They made iron weapons and tools. South of the Limpopo River were two main groups of Bantu-speaking peoples: the Sotho, who lived on the plateau; and the Nguni, in the coastal belt. The westernmost Nguni tribes slowly expanded westward as population increased and tribes split, but the Boers moved toward them more rapidly. Early in the 18th century Boer hunters began to make contact with the Nguni and bartered with them for cattle and ivory. The company vainly tried to prevent such contacts and delay the convergence of the Boers and the Nguni by issuing proclamations prohibiting barter and defining limits beyond which the Boers were not to go. When the governor, J. van Plettenburg, toured the frontier in 1778 he found that Boers were already intermingled with Africans in the vicinity of the Great Fish River.

A long series of frontier "wars" followed—basically struggles for possession of the land. The first fighting took place in 1779–81 and was sufficient to show the Boers that they faced tough opposition. They therefore appealed to the company for aid; although a *landdrost,* M. Woeke, was appointed to Graaff-Reinet in 1785, no troops were sent. Woeke's life was soon made intolerable by the Boers, and in 1793 he was superseded by H.C.D. Maynier. Further fighting broke out in that year, and again the result was inconclusive. The Boers were disgruntled. For nearly a century they had been left by the company to make their own terms with man and nature, and when at length they had asked for help all they received was a government official who not only failed to defeat the Nguni but also allowed Khoikhoin to come to his court with complaints against them. In 1795, on the eve of the British conquest of the Cape, they drove out Maynier and declared Graaff-Reinet to be a republic. But the name "republic" was a euphemism: in fact there was anarchy at Graaff-Reinet. Similar events took place at Swellendam, where another "republic" came into being.

Meanwhile there was trouble in the southwestern part of the colony also. The company had maintained rigorous commercial restrictions, together with a system of government that gave the colonists practically no voice in the determination of policy. The Boers had managed to escape most of the effects, but the restrictions had always borne heavily on the agricultural farmer who had produce to sell and on the Cape Town man who lived by trade. In the 1770s, which were lean years, the discontent came to a head with the organization of the Cape Patriots movement, which had some connections with the anti-Orange Patriot Party in the Netherlands. A petition was drawn up criticizing many of the officials and asking for full burgher representation on the Council of Policy and the right to trade freely throughout the company's possessions. When it was laid before the directors in Amsterdam, their only response was to dismiss the official who was most roundly criticized—the fiscal, W. Boers.

The Cape Patriots movement

In 1784 a second petition was taken above the heads of the directors to the States-General of the Netherlands. This time the burghers were given equal representation with officials on the Court of Justice (though not on the Council of Policy) and the right to trade with the company's possessions—a right that was nullified by the conditions that were attached. By then, however, the disaffection had subsided because the American Revolutionary War had brought a large French garrison to Cape Town to help the Dutch to defend it against the British, and the garrison provided an abnormally large market for Cape produce and brisk business for the traders and innkeepers. In the early 1790s bad times came again as the company, trying desperately to stave off bankruptcy, made a final effort to make the colony pay its way. The garrison was reduced, public works were suspended, and—the last straw—new taxes were imposed. So, when the British invading force arrived in 1795, it found depression in the west and rebellion in the east. The company had never done much for the white colonists and still less for the nonwhite peoples.

SOUTH AFRICA DIVIDES (1795–1870)

British conquest. In the course of the Napoleonic Wars, the Cape was captured by a British naval and military force in September 1795, nominally on behalf of the Prince of Orange, who had taken refuge in England from the Dutch republicans. In their struggle with France the British considered it essential not to let the Cape fall into enemy hands, and until the Treaty of Amiens a hold was kept on it for strategic reasons. The British tried to conciliate the Afrikaners, and by abolishing the restrictions on internal trade and maintaining a large garrison they caused an economic revival in the southwestern part of the colony. On the eastern frontier, on the other hand, the British were no more successful than their predecessors had been. Although the republics of Swellendam and Graaff-Reinet soon submitted to the new government, anarchy developed in Graaff-Reinet again in 1799, after the arrest of a prominent Boer, A. van Jaarsveld, on a charge of forgery. Not only the Xhosa but also many of the Khoi servants of the Boers made the most of the opportunity created by the divisions among the whites and caused havoc as far west as Knysna. The British general F. Dundas patched up a truce; but H.C.D. Maynier, appointed resident commissioner of the two eastern districts, failed to satisfy the Boers, who rebelled again in 1801, when once again their Khoi servants joined with Xhosa tribesmen to harry the farms over a wide area.

The eastern districts were therefore in turmoil when Britain gave up the Cape under the terms of the Treaty of Amiens (1802). By this time the company was defunct, and the Dutch government—the Batavian Republic organized by the French in 1795 after their conquest of the country—succeeded to its charge. J.A. de Mist, the commissioner-general, and Gen. J.W. Janssens, the governor, tried to restore the morale of the Boers and showed a considerable understanding of the problems of the country; but these men did not stay long enough to consolidate their reforms for in 1806 the British sent out another force to recapture the Cape from Napoleon's Dutch allies. Janssens lacked the means to put up much resistance, and once more the colony passed into the hands of the British, whose title was confirmed in 1814 as part of the general peace settlement.

British policy, 1806–36. Until 1823 British policy was cautious and conservative. The governor had autocratic powers, but the Afrikaners were used to that. Moreover, their Roman-Dutch law was retained, and local administration remained as of old in the hands of *landdrosts* and *heemraden,* most of them Afrikaners. These were comparatively prosperous years. The southwestern farmers profited from their access to British markets, where their wine sold well thanks to a substantial tariff preference. Elsewhere, however, extensive pastoral farming remained the norm. Nevertheless, the British introduced some changes in these years. They appointed more *landdrosts.* In 1809 they promulgated a code for the treatment of Khoi servants, who were to be employed only under written contracts registered at a *landdrost*'s office and were given legal protection against ill treatment; on the other hand, they were not to be allowed to leave their employer's farm without a pass signed by the employer. From 1811 onward high court judges, who had previously sat only in Cape Town, made annual circuits to hear cases in the districts.

British reforms

Mild though these reforms were, they caused trouble. There was much grumbling in 1812 when the judges on circuit in the eastern districts tried a number of Afrikaners on charges of ill-treating their Khoi servants. Three years later a Boer, F.C. Bezuidenhout, opened fire on a party that had come to arrest him for repeatedly ignoring his *landdrost*'s orders to answer a charge of cruelty; in the skirmish Bezuidenhout was killed, whereupon his brother proceeded to organize opposition to the government; but the Slachter's Nek Rebellion soon ended when about 60 rebels were rounded up and tried, and five ringleaders were hanged.

The British also brought superior force to bear against the Xhosa on the eastern frontier and thus started the process of the conquest of the African tribes. In 1811 British regular troops and Boer commandos drove the Xhosa back to the eastern side of the Fish River and in 1819 to the eastern side of the Keiskama.

In 1820, 5,000 British settlers were brought out and placed on 100-acre lots in the Zuurveld, on the western side of the Great Fish River. The 1820 settlers did not, however, form the human barrier that was intended, because the land was not suited to agriculture; within a few years most of them had abandoned their lots and become townsmen in Port Elizabeth and Grahamstown. Others became traders among the African tribes to the east. Those who stayed on the land eventually made good by producing wool for export to England, especially when large farms came onto the market at the time of the Great Trek. This was the first subsidized immigration scheme since the 17th century, and the 1820 settlers were the first white immigrants who were not assimilated by the Afrikaner people.

When they had got their bearings the 1820 settlers were dismayed to find that the legal system of the colony was alien to them, the judges were officials without any particular legal training, and Lord Charles Somerset, the governor from 1814 to 1826, was an autocrat who made it impossible for them to publish a newspaper. Their agitation for British liberties and institutions evoked some response. The Charter of Justice of 1827 created a Supreme Court of qualified judges with security of tenure, abolished the *heemraden,* and introduced English rules of procedure and evidence, including the jury system, while it left untouched the Roman-Dutch substance of the law. In 1834 an Executive Council of officials and a Legislative Council of officials and nominated nonofficials were established. By that time the freedom of the press had been won, English had become the official language of the courts, and British teachers had been imported to start village schools in which English was the medium of instruction.

During these years South Africa was attracting the attention of the British humanitarians or evangelicals who were striving to free the nonwhite populations of the colonies from their legal disabilities. They were supplied with severe criticisms of South African conditions by John Philip, South African superintendent of the London Missionary Society, which had established a number of mission stations up and down the country. Philip reached the conclusion that the Khoi code of 1809, with its stringent pass regulations, was unjust; in 1826 he returned to England to appeal to the government and the British public, and in 1828 the House of Commons passed a resolution for the emancipation of the Khoikhoin. Seeing the trend of affairs, Gen. Richard (later Sir Richard) Bourke, the acting governor, had already passed an ordinance, number 50 of 1828, which abolished the pass system and gave "free persons of colour" complete liberty of movement. Meanwhile, with their eyes focussed primarily on the West Indies, the British government had been insisting on improvements in the treatment of slaves; and in 1833 the reformed Parliament passed the Emancipation Act under which all slaves in the colonies became free after a period of apprenticeship, and £20,000,000 was voted as compensation to their owners. When their period of apprenticeship ended in 1838 the former slaves in the Cape Colony stepped into the position of other "free persons of colour." Thereafter the emancipated slaves and Khoikhoin—the Cape Coloured people—became a rural and urban working class in the Cape Colony.

ıe British
nancipa-
ɔn Act

Meanwhile frontier trouble had flared up again. Unrest had developed among the border Nguni tribes for various reasons, including the frequency of military raids on their villages made in reprisal for the theft of stock from white farmers. In 1834 Gen. Sir Benjamin D'Urban arrived as governor with instructions to devise a more equitable policy, but he did nothing about it until large bodies of Africans invaded the colony and did serious damage. D'Urban then reasserted British military superiority and annexed the land up to the Great Kei River, from which most of the Africans were to be expelled (May 1835). But he found it impossible to enforce the expulsion, and the colonial secretary, Lord Glenelg, disapproved; consequently, in 1836 the territory was disannexed, and an attempt was made to pacify the tribes by making treaties with their chiefs. By that time, however, the frontier was no longer a clear line of division between the races; white traders and missionaries were working among the tribes to the east of it, and Africans were working on white farms to the west.

The Great Trek. For several generations the Boers had been left more or less to their own devices. Now, under an alien government actuated by liberal ideals that they did not comprehend, many of their institutions were being transformed. They were short of labour, and they found it difficult to comply with the regulations and get the compensation money for their slaves; their property was being pilfered by wandering Khoikhoin, and they suffered further losses during the African invasion in 1834. Behind all these things they saw the hand of missionaries, who, they considered, misrepresented them in England. The reversal of D'Urban's frontier policy was the last straw that made many of them decide to throw off British dust and find a new home where they could, in the words of Piet Retief, their ablest leader, "preserve proper relations between master and servant." Later, Anna Steenkamp recorded that the British had placed their slaves "on an equal footing with Christians, contrary to the laws of God and the natural distinctions of race and religion, so that it was intolerable for any decent Christian to bow down beneath such a yoke; wherefore we withdrew in order thus to preserve our doctrines in purity." Imbued with this spirit the Voortrekkers, about 12,000 Afrikaners, men, women, and children, left the colony with their sheep, their cattle, their ox wagons, and their Coloured servants between 1835 and 1843.

The
Voortrekkers

In the previous two decades there had been widespread destruction and dislocation among the African tribes, caused by the Zulu under Shaka, and their offshoot, the Ndebele or Matabele, under Mzilikazi (Mozelekatse). By the time of the Great Trek two areas seemed suitable for white settlement—Natal, south of the Tugela River, and the High Veld, on either side of the Vaal. But appearances were deceptive, for these two areas were kept denuded by Zulu and Ndebele impis (regiments). The first Voortrekkers made for the High Veld, where they were attacked by the Ndebele, whom they routed in 1837. Most of them then decided to settle in Natal. Piet Retief and a pioneer group went to the kraal of Dingane (Dingaan), Shaka's successor, to negotiate for a cession of land; and there, in February 1838, he and 70 Boers and 30 Coloured servants were treacherously massacred, while another 300 Boers, mainly women and children, and 200 Coloured servants, were killed in northern Natal. On December 16, 1838, however, Andries Pretorius gained a decisive victory over the Zulu at Blood River. Most of the Voortrekkers then proceeded to stake out farms in Natal. Border troubles ensued, and, when in 1840 a Boer commando expedition made a sortie to the south against the kraals of a chief named Ncaphayi (Ncapaai), his neighbour Faku, who had a treaty with the British, appealed for aid.

Previously the British had been undecided what to do about the Voortrekkers; the Cape of Good Hope Punishment Act of 1836 had made them liable, as British subjects, to trial in British courts for crimes committed south of latitude 25° S, but little had been done to make it effective. Then, however, anxious lest the activities of Natal commandos lead to disturbances on the Cape frontier, the governor, Sir George Napier, sent a small force to Natal, which, after some fighting, became British territory in 1843. Most of the Voortrekkers who had settled in Natal then inspanned their oxen again and trekked back to the High Veld, north of the Vaal River.

Between the Orange and the Vaal, however, the British were drawn in for a while. There was a racial medley in that region; Griquas (of mixed Khoi–European parentage), who had lived north of the Orange River for several decades; Africans, survivors of the Ndebele devastation of the High Veld, who had been rallied by Mshweshe (Moshesh), a Sotho chief, near the Caledon River; and Afrikaners. The British had made a treaty with the West Griqua chief, Andries Waterboer, in 1834, and they made treaties with the East Griqua chief, Adam Kok III, and with Mshweshe in 1843, hoping to use them to keep the peace along the Orange River. Trouble followed when Kok, in accordance with his treaty, tried to hand an Afrikaner over to the Cape authorities for trial under the Punishment Act. The Afrikaners took up arms and were dispersed by troops from the Cape at Zwartkoppies (1845). Two years later Gen. Sir Harry Smith became governor of the Cape Colony and British high commissioner in South Africa. An optimist, he believed that the South African imbroglio would be pacified if the area of British authority was extended. Accordingly, after overcoming further African resistance on the eastern frontier of the Cape, he annexed the land up to the Great Kei River as the province of British Kaffraria (1847) and the land between the Orange and the Vaal as the Orange River Sovereignty (1848). The colonial secretary, Earl Grey, accepted these annexations most grudgingly.

Humanitarian influence was waning in Britain, and the main concern of the Colonial Office was to cut down expenses. Therefore, when further fighting broke out on the eastern frontier and similar difficulties developed with Mshweshe in the Orange River Sovereignty, the British government decided to reduce its commitments in South Africa by recognizing the independence of the Voortrekkers. This was done in two stages. In 1852 the Transvaalers were granted "the right to manage their own affairs without any interference on the part of the British government" in a convention signed at the Sand River; and in 1854 similar arrangements were made with the Afrikaners of the Vaal–Orange area at Bloemfontein.

Independence for the Voortrekkers

South Africa was thus divided into three camps: Afrikaner republics; British colonies; and still-independent African societies.

Afrikaner republics. At the time of the Bloemfontein convention there were about 15,000 white persons engaged in pastoral farming in what then became the Orange Free State. They soon agreed on a rigid constitution, which vested legislative power in an elected council (Volksraad), executive power in an elected president—the voters being the white male adults—and local administration in *landdrosts* and *heemraden* on the old Cape model. The East Griqua question was settled in 1861 when Adam Kok sold his remaining land rights in the Free State and trekked with his people to found a new Griqualand East nearer the coast. The Sotho question was more difficult. The British had abandoned the sovereignty because they shirked the expense of controlling the rivalry between the Afrikaners and the Sotho for the Caledon River valley, which included some of the best arable land in southern Africa.

When fighting broke out in 1858 the Sotho outmanoeuvred the Free Staters, who were glad to accept the mediation of the high commissioner Sir George Grey (first convention of Aliwal North). The Free Staters, conscious of their weakness, then tried to unite with the Transvaal but were stopped by Grey, who threatened to withdraw the recognition of their independence and advised them to federate with the Cape Colony; yet when they expressed a desire to do so, the British government refused. Grey considered that a federation, under the crown, of all the colonies and republics in South Africa would be in the general interest. The British government, however, was averse to any extension of its area of responsibility. Further fighting started in 1864; this time the Free Staters were more successful and were able to dictate terms, by which Mshweshe lost most of his arable land (Treaty of Thaba Bosiu, 1866). But the peace was not kept and Mshweshe asked Britain to annex Basutoland. This was done in 1868, and the next year the frontier line between the Free State and Basutoland was agreed (second convention of Aliwal North).

Britain's annexation of Basutoland

The intervention of the British preserved for the Sotho people more land than had been left to them in 1866; but it gave serious offense to the Free Staters, depriving them of the fruits of their victory. Thereafter the Free State, its frontiers pacified and absolved from the responsibility of administering a large number of Africans, became what Lord Bryce once described as a model republic, under the wise leadership of J.H. Brand, president from 1864 to 1888. The Cape Colony took over Basutoland in 1871, but later got into difficulties in trying to force the Sotho to forfeit their arms and was glad to hand it back to Great Britain in 1884.

Meanwhile the Voortrekkers north of the Vaal River, numbering perhaps 22,000 at the time of the Sand River Convention, preserved their old Boer mode of life. It was not until 1857 that most of them overcame their factional disputes and united to form the South African Republic under the presidency of M.W. Pretorius, son of the victor of Blood River. Their constitution, which created institutions similar to those in the Free State, frankly stated that "The people desire to permit no equality between coloured people and the white inhabitants, either in church or state." In 1863 there was further factional fighting; and Boer expansion frequently led to disputes with African tribes, the weaker of which were broken up, the remnants becoming squatters and servants.

British colonies after the Great Trek. Not more than a quarter of the Afrikaners had left the Cape Colony during the Great Trek, though others went north in later years. Those who remained gradually became accustomed to the more equal order, and in the course of time a liberal tradition developed in the colony, in contrast with the illiberal tradition of the republics. When the colony received representative government in 1853, the only franchise qualifications were low economic ones, irrespective of race or colour; but most of the Coloured people were unable to qualify and the vast majority of the electorate were white men. By that time the loss of population had been offset by further British immigration; and the decline in viticulture caused by the removal of British tariff preferences was remedied by an increase in the production of wool for export.

In British Kaffraria the high commissioner, Sir George Grey, tried to achieve quick results from a policy of civilizing the Africans by encouraging trade, mission schools, medical services, and public works, by bringing in white settlers, and by using white magistrates to undermine the authority of the chiefs. There was a reaction in 1857, when the Africans destroyed their own stock and crops in the superstitious belief that the white man would then vanish from the country. The result was catastrophic: thousands starved to death, and emaciated survivors poured into Cape Colony in search of work. In 1865 British Kaffraria was incorporated in the colony. Soon afterward the British government urged Cape Colony to assume full responsibility for its own affairs, and in 1872 a responsible government bill was carried and J.C. Molteno became the first prime minister.

The first settlement in Natal was established at Port Natal (now Durban) in 1824 by a small party of British ivory traders from the Cape. The population rose with the arrival of the Voortrekkers in 1838 when for a few years to 1842 Natal became a republic ruled by a Volksraad at Pietermaritzburg. British immigrants began to arrive in numbers after 1847, so that its white population became predominantly British and crept up to 9,000 in 1856, when representative institutions were established, and to 50,000 in 1893, when responsible government was obtained. Mixed farming was practiced inland, and sugar was grown in the subtropical coastal belt.

One result of the defeat of Dingane was an influx of Africans from Zululand into Natal, many of them returning to the region from which they had formerly been evicted by Shaka. Thus the African population rose steeply from about 20,000 in 1840 to nearly 500,000 in 1893. They were handled by Sir Theophilus Shepstone, the son of a missionary who had come to South Africa with the settlers in 1820. He placed most of the Africans in a number of reserves or locations, comprising about one-seventh of the total area of the colony. Their tribal laws and customs were maintained, and they were able to preserve their tribal economy for another generation, when they began to suffer from land shortage. Africans subject to tribal law were not eligible for the franchise, but in 1864–65 laws were passed which enabled Africans to apply to be exempted from tribal law and, seven years later, for the franchise. However, the regulations were such that few became exempted and hardly any became voters.

Shepstone' policy for Africans

The colonists were critical of Shepstone's policy, mainly because it did not make the Africans work for them. The coastal sugar planters, in particular, were short of labour, and it was to help them that Natal began to import Indian labourers in 1860. By 1866, 6,000 had arrived, and when their five-year contracts expired most of them chose to stay in Natal, as they were entitled to do. In 1872 the Indian government prohibited further emigration because some of the Indians had complained of ill treatment; but the migration was resumed on a large scale, with a heavy government subsidy, in 1874, and by the end of the century the Indians outnumbered the whites in Natal.

IMPERIALISTIC POLICIES (1870–1910)

Diamonds and gold. After 1870 the South African economy was rapidly transformed. Previously South Africa had been an economic backwater, with most of its inhabitants, whites and nonwhites, engaged in inefficient near-subsistence farming; and it had lacked the means to attract the capital and the skilled personnel to create the facilities of a modern country. In 1867 diamonds were found along the Orange and the Vaal rivers; in 1870 the dry diamond diggings began to be worked at Kimberley; and in 1886 the gold rush to the Witwatersrand began to eclipse all previous gold rushes. The effects were remarkable. The sum of South Africa's imports and exports rose from £49,000,000 in the decade 1861–70 to £128,000,000 in 1871–80, to £179,000,000 in 1881–90, to £357,000,000 in 1891–1900, and to £700,000,000 in 1901–10. In the last of these decades gold (59 percent) and diamonds (19 percent) accounted for 78 percent of South Africa's exports; and by 1910 £121,000,000 had been invested in the Witwatersrand gold-mining industry and dividends totalling £77,000,000 had been paid by it.

Immigration raised the white population from about 328,000 in 1875 to 1,117,000 in 1904. Railway construction was even more striking: in 1870 there were only 69 miles of railway track in all South Africa; by 1886 there were 1,800 miles and the line from Cape Town had reached Kimberley; by 1895 there were 3,600 miles and Johannesburg was connected with five ports—Cape Town, Port Elizabeth, East London, Durban, and Lourenço Marques (Portuguese East Africa).

Thus South Africa at last entered the world economy and acquired a special place in it as the source of about a third of the world's annual supply of gold and of more than half its diamonds.

e Beers onsol-ated ines id Cecil hodes

The control of these great industries became concentrated in a few hands. The digger phase on the diamond fields was short-lived; companies forced out individuals and then the companies amalgamated until in 1891 De Beers Consolidated Mines controlled the entire South African diamond industry. The digger phase of the gold fields was even shorter, but there the process of concentration never reached quite the same finality; companies became organized into groups or corporations, and the groups became associated in the Chamber of Mines for certain purposes only, notably for the recruitment of African labour. Thus great financial power came to be exercised by a few men, above all by Cecil Rhodes, who went to South Africa for his health in 1870 and became the most successful of all the mining financiers, controlling both De Beers and the Consolidated Gold Fields, one of the strongest gold corporations.

The skilled work in the mining industries was done by whites, mainly immigrants at first, who drew high wages. The unskilled work was done by low-paid Africans who went to the mines from all over southern Africa, especially from Portuguese East Africa, to work for limited periods during which they lived in closed compounds, separated from the ordinary life of the community. In 1898, for example, the Witwatersrand gold-mining industry employed 9,476 whites at an average monthly wage of £26 and 67,797 Africans at an average monthly wage of £2 9*s.* Previously many Africans had been employed as farm labourers in return for squatting rights or low wages. Now the agrarian pattern of race relationships was being adapted to the mining industries.

The discovery of diamonds led to a serious controversy. The Transvaal claimed part of the diamondiferous area and the Orange Free State claimed the rest of it, while David Arnot, a white attorney, claimed it all on behalf of his client Nicholas Waterboer, the chief of the West Griquas. Arnot managed to persuade the British Colonial Office that the Griquas had the best case; therefore when a new high commissioner, Sir Henry Barkly, reached South Africa in January 1871 he had instructions to take a strong line. Barkly persuaded M.W. Pretorius, the Transvaal president, to submit his case to arbitration by R.W. Keate, the governor of Natal. When Keate discovered that Pretorius had unwittingly placed great reliance on a forged document, he decided against him. Although this award did not directly affect the Free State claims, which were the strongest, Barkly, failing to understand their merits, annexed to the crown the entire area claimed for Waterboer in the same month (October 1871). Five years later a land court decision exposed the fallacies in Waterboer's claim to sovereignty over the diamond fields; whereupon J.H. Brand, the Free State president, went to London and exacted £90,000 compensation from the imperial government. In 1880 Griqualand West was incorporated in the Cape Colony.

Annexation and retrocession of the Transvaal. Lord Carnarvon, colonial secretary in Disraeli's 1874 ministry, hoped that the South African states and colonies might unite in a self-governing federation under the British crown. The following year he summoned a conference to discuss the project in London, but it was a failure. The Free State, smarting under the recent annexations, would have nothing to do with the project; nor would the Cape, whose ministers' dignity had been ruffled by the conduct of the historian J.A. Froude, who had toured the colony on behalf of Carnarvon urging federation. Federation from the south having failed, Carnarvon turned his attention to the north and found that the South African Republic was ripe for plucking.

That republic, unlike the Free State, had never acquired stability. Pretorius was bundled out of the presidency after the annexation of Griqualand West, and the Transvaalers chose as his successor Thomas François Burgers, a Cape colonial preacher of the Dutch Reformed Church. But in Burgers they had more than they had reckoned for; he was critical of Transvaal backwardness and proposed to remedy it by social, educational, and religious reforms. He also borrowed money, intending to build a railway to Delagoa Bay to free the Transvaal from its dependence upon British trade routes; but not a single track of the railway was laid, whereas the loan strained the meagre resources of the republic and alarmed the Cape and Natal merchants whose interests it threatened. When in 1876 Burgers personally waged an unsuccessful campaign against the Pedi (one of the Sotho peoples) in the northern Transvaal, he was totally discredited. The Afrikaners, including Paul Kruger, were plotting against him, and the traders were conniving with the Cape and Natal merchants to bring in the British. Carnarvon heard their cries and commissioned Shepstone to annex the Transvaal. Entering the republic in January 1877 with 25 police and a staff of seven, Shepstone annexed it in April in a proclamation that promised "the fullest legislative privileges compatible with the circumstances of the country and the intelligence of the people."

British annexation of the Transvaal

But British rule in the Transvaal was neither efficient nor tactful, and the elected legislature did not materialize. Then the Zulu War revealed the weakness of British arms: the Zulu, their martial spirit revived by Cetshwayo, routed a large British force at Isandhlwana in 1879 before they were vanquished in the following year. By that time the Transvaalers, who had passively acquiesced in the annexation, were moving toward open rebellion under the leadership of Paul Kruger, Piet Joubert, and M.W. Pretorius, and in 1881 they, too, wiped out a British force at Majuba. Gladstone's ministry, which assumed office in Britain in 1880 but failed to reform the Transvaal administration in time to avert the rebellion, then gave the Transvaalers a qualified independence (Pretoria Convention, 1881). Three years later they allowed the republic full internal autonomy, while maintaining British control over its relations with foreign states, other than the Free State (London Convention, 1884). Carnarvon's project had completely miscarried, and his permissive act for the unification of South Africa (1877) was stillborn.

The Rhodes–Hofmeyr alliance. In the Cape, which remained the most civilized of the South African states, the advent of reponsible government and the series of British annexations prompted a group of Afrikaners to form a political organization, the Afrikaner Bond, in 1879. At first the Bond's program was crudely sectional, but three years later Jan Hendrik Hofmeyr gained control and turned it in the direction of Anglo-Afrikaner cooperation, with a British South African federation, which should include the republics, as the ultimate goal. Meanwhile Cecil Rhodes had entered the Cape Parliament and, as his vision of British expansion in Africa was reconcilable with Hofmeyr's, he was able to form a ministry with Bond support in 1890.

Meanwhile, although it had washed its hands of Basutoland in 1884, the Cape Colony by 1894 had annexed all the land below the mountain escarpment as far east as the Natal border at the Umzimkulu River. The Nguni inhabitants of these Transkeian territories remained in occupation of most of their lands after their political autonomy had been extinguished; but steps were taken to prevent them exercising much influence over the colonial Parliament. Laws were passed providing that land occupied on tribal tenure should not satisfy the economic requirement for the franchise (1887) and raising the economic qualification and introducing a simple educational test (1892). A start was also made in encouraging African communities to adopt individual land tenure and to elect local administrative councils (1894). Thus the Cape government drew a line between "tribal" and "civilized" Africans, treating tribal Africans as minors and giving civilized Africans the same rights as white men in theory, though in practice not many Africans got the vote. By 1895, therefore, the Rhodes–Hofmeyr alliance was bidding fair to heal Anglo-Afrikaner tensions in the Cape Colony, without destroying the prospect of a constructive relationship with Africans.

British clash with the Transvaal. During the late 19th century the Bantu-speaking tribes of South Africa, which had previously maintained their autonomy north of the Vaal River, came under white control as a result of Boer expansion and the intervention of Germany and Great Britain. In 1884 Germany took over South West Africa, and Transvaalers began to encroach westward toward it, setting up the republics of Stellaland and Goshen. This expansion, which seemed likely to cut off the British from access to central Africa via the Cape, was checked, largely on the insistence of Rhodes, in 1885, when Great Britain, after a display of force, proclaimed a protectorate over northern Bechuanaland and annexed southern Bechuanaland as a crown colony. Transvaal efforts to keep their north clear of the British were also foiled by Rhodes, who secured a prospecting concession from the Ndebele chief Lobengula in 1888, obtained a royal charter incorporating the British South Africa Company to exploit the concession in 1889, and dispatched a pioneer expedition to occupy what became known as Rhodesia in 1890. Finally, the Transvaalers tried to push toward the east in order to get a port on the Indian Ocean, but, although they incorporated fresh territory, including Swaziland, they did not reach the coast, as Britain annexed Zululand in 1887 and Tongaland, the last gap, in 1895. The Afrikaner republics were thus encircled by the colonies of Great Britain and its oldest ally, Portugal—hemmed in a kraal, as Paul Kruger complained. Southern Bechuanaland was taken in by the Cape Colony in 1895; Zululand and Tongaland by Natal in 1897; but northern Bechuanaland remained a British protectorate.

Paul Kruger as president of the Transvaal

In 1883 Paul Kruger was elected president of the Transvaal. His character had been molded in the hard school of the Great Trek and of commando fighting against Afrikaner factions and African tribes; and recently he had led the opposition to President Burgers and the British. His policy was to regain complete independence for the republic and to preserve Afrikaner control of it. In 1885 he was constrained to suggest a customs union with the Cape, whose government rejected the offer. Soon afterward his financial difficulties were solved by the gold discoveries and he proceeded to use the mining industry as a milch cow, imposing heavy taxation and granting monopolies of essential materials, such as dynamite.

The growth of such an industry in the heart of his republic, however, posed serious problems. The Afrikaners became outnumbered by the newcomers who poured in from all parts of the world, most of them British subjects. To give these uitlanders ("outlanders"), as they were called, an effective vote seemed like suicide; to withhold such a vote was bound to cause trouble with the British. Kruger and the Volksraad never hesitated: they progressively raised the franchise qualifications until by 1894 no uitlander could vote in presidential elections, and only those who were 40 years old and had lived 14 years in the Transvaal, during 12 of which they had been subject to an oath of allegiance, could vote in Volksraad elections. In short, on the one hand the government got most of its revenue from the uitlanders and denied them political rights; on the other hand the uitlanders continued to swarm into the republic where many of them prospered exceedingly. Such a situation required patience. Although Kruger was reelected for a third term in 1893, his majority over the more moderate Piet Joubert was small and a change seemed likely in 1898.

But Rhodes was not patient. Finding that Kruger stood between him and a British South African federation, he planned to overthrow him. Kruger nearly played into Rhodes's hands in 1895 when, contrary to the London Convention, he tried to injure Cape trade with the Transvaal and the Cape railways that were competing with the Delagoa Bay line by making the rates over the Transvaal sector of the Cape line prohibitive and by closing the Vaal drifts to prevent the carriage of goods by wagon. But he gave way when the colonial secretary, Joseph Chamberlain, agreed to support the Cape ministry's protests, by force if necessary.

The Jameson Raid

Rhodes then pushed ahead with his conspiracy: the uitlanders were to rise, an armed force under Leander (afterward Sir Leander) Starr Jameson was to go to their assistance, and the high commissioner was to hurry to Pretoria to "restore order." Chamberlain knew of Rhodes's plans when he provided a suitable jumping-off place for Jameson's force by ceding to the British South Africa Company a strip of land along the Transvaal border of the Bechuanaland Protectorate, but he later withdrew his support when he learned that the uitlander rising would not take place to give the invasion an air of respectability. Therefore, when Jameson, contrary to Rhodes's last-minute instructions, rode into the Transvaal from Pitsani near Mafeking on December 29, 1895, he was disowned by the colonial secretary and the high commissioner. Four days later he surrendered to Transvaal commandos at Doornkop.

Outbreak of the South African War. The raid fiasco cleared the decks for the South African War. Rhodes had to resign the premiership of the Cape Colony, where, in 1898, W.P. Schreiner became head of a ministry that relied upon the support of the Afrikaner Bond, now deeply suspicious of British designs. In the Orange Free State, which had opposed the Transvaal in the drifts crisis, M.T. Steyn was elected president and formed a military alliance with the Transvaal. While Steyn and the Bond were not uncritical of the Kruger regime, Kruger's anglophobia seemed to have been vindicated in the Transvaal, where he won the 1898 presidential election by a large majority. In England, on the other hand, a parliamentary committee failed to probe Chamberlain's dealings with the raiders, a telegram from the German emperor congratulating Kruger turned Jameson and Rhodes into popular heroes, and the press paid much attention to uitlander grievances.

As the tension mounted, Chamberlain, taking the high ground of British paramountcy in South Africa, claimed to be competent not merely to control the foreign relations of the Transvaal (a right expressly granted by the London Convention) but also to intervene on behalf of the uitlanders. If the right to intervene existed, there were grounds enough for exercising it. The tendency of the Volksraad to give sweeping powers to the president was exemplified in the judges' crisis: in 1897 Chief Justice John (afterward Sir John) Gilbert Kotze delivered a judgment that meant the greater part of the laws of the Transvaal had been enacted unconstitutionally and were null and void, whereupon the Volksraad deprived the courts of the testing power and gave Kruger the power to dismiss any judge who disagreed; and the next year Kruger dismissed Kotze. In 1898 an Anglo-German agreement removed the likelihood of German intervention in the event of war in South Africa, and Chamberlain and the high commissioner, Sir Alfred (afterward Viscount) Milner, proceeded to apply the screw. In so doing they brushed aside several attempts by the Cape government and the Afrikaner Bond to avert a catastrophe.

The central issue was the Transvaal franchise. In June 1899, in conference at Bloemfontein, Milner proposed a simple five years' franchise, Kruger refused, and the negotiations collapsed. The last chance of peace was lost in August when Jan Christiaan Smuts, Kruger's young state attorney, offered a five years' franchise and Chamberlain rejected the conditions that were attached to the offer. Milner had long since made up his mind that war was "inevitable," and in September Chamberlain arranged for 10,000 British troops to be sent to augment the meagre British forces in South Africa. The republicans, who had been arming furiously for several years, replied by issuing an ultimatum that expired on October 11, 1899.

(L.M.T./R.J.Da./Ed.)

South African War. The forces were unequal from the beginning. British military strength in South Africa reached nearly 500,000, whereas the Boers were never able to muster more than about 65,000. The British, however, were hampered by their ignorance of the rough terrain. The Boers put their knowledge of the countryside to good effect and attacked the British on two fronts—from the Transvaal into Natal and from the Orange Free State into the northern Cape. The northern districts of the Cape Colony rebelled against the British and joined the Boer forces. In the course of Black Week (December 10–15, 1899) the Boers beat the British in a number of major engagements and besieged the key towns of Ladysmith, Mafeking, and Kimberley. Large numbers of British reinforcements were being landed but before the siege of Ladysmith could be relieved, the British suffered another reverse at Spion Kop (January 1900).

The British began to take the offensive under Maj. Gen. Herbert Kitchener (later Lord Kitchener) and Field Marshal Sir Frederick Sleigh Roberts. The besieged towns were relieved and the Boer armies were beaten in the field. Bloemfontein was occupied in February 1900, and Johannesburg and Pretoria were taken in May and June. Kruger left the Transvaal for Europe. But the war, which until then had largely been confined to military operations, entered upon its most destructive phase. For 15 months Boer guerrillas, under the leadership of generals Christiaan Rudolf de Wet and Jacobus Hercules de la Rey, harried the British army bases and communications; large rural areas of the Transvaal and Free State (which the British annexed as the Orange River Colony) were out of British control.

Kitchener responded with brutal retaliation against the noncombatant civilian populations and initiated a scorched-earth policy. The farms of Boers and Africans alike were destroyed, and the families of Boer guerrillas were rounded up and placed in concentration camps. More than 20,000 Boer women and children died in the carelessly run and unhygienic camps.

he Peace
ereeniging

To contain the guerrillas' actions, Kitchener's troops installed blockhouses and barbed-wire barriers along the railway lines in the Free State and later in the Transvaal. The barriers were extended across the countryside to form zones of fortification. The guerillas continued their attacks, many of them deep into the Cape Colony, Gen. Jan Smuts leading his forces to within 50 miles of Cape Town. But Kitchener's drastic and brutal methods slowly paid off. The Boers unsuccessfully sued for peace in March 1901; finally, they accepted the loss of their independence by the Peace of Vereeniging in May 1902. But the divisions and hatreds exacerbated by this war would affect South African politics for more than half a century.

(Ed.)

Reconstruction and union. After the Peace of Vereeniging, Milner, high commissioner and governor of the crown colonies—the Transvaal and the Orange River Colony—concentrated on their material reconstruction from the ravages of war. The gold mines resumed large-scale operations, and the Afrikaners were returned to their land from the prison and concentration camps and supplied with food, stock, seed, and implements, no less than £10,000,000 being spent by Great Britain on their rehabilitation. Milner also prepared the ground for political union by bringing all the colonies into a South African customs union and by amalgamating the railways of the Transvaal and the Orange River Colony.

All this work was well done by Milner and his band of able young men, who were dubbed the "kindergarten." But the opportunity was not used to emancipate the nonwhite inhabitants of the crown colonies from their legal disabilities, which remained much as they had been during the republican regime. Nor did Milner gain the confidence of the Afrikaners: they disliked him as "the man who made the war," who openly supported an agitation for the suspension of the Cape constitution in 1902, who imposed an English educational system on the crown colonies, and who was responsible for bringing Chinese labourers to the Witwatersrand gold mines in 1904. For the importation of Chinese labourers there were, indeed, strong economic reasons, because the supply of African mine labour had temporarily fallen off; but in Britain the decision played a large part in the fall of A.J. Balfour's Conservative government in December 1905, and the change of government was followed by a radical change in the distribution of power in South Africa.

Sir Henry Campbell-Bannerman's Liberal ministry, anxious to make amends to the Afrikaners for a war that had been brought about by its predecessors, decided to give the new colonies responsible government, with exclusively white electorates. The Liberals defended this decision by referring to the Treaty of Vereeniging, but it did not, in fact, debar them from allowing Coloured and Asian people to vote. Consequently in March 1907 Louis Botha, former commandant general of the forces of the South African Republic, became premier of the Transvaal, with Smuts as his right-hand man, and in December 1907 A. Fischer became premier of the Orange River Colony. The swing of the pendulum was finally completed in February 1908, when J.X. Merriman's South African Party, consisting largely of members of the Afrikaner Bond, ousted Jameson's Progressives from office in the Cape Colony.

The move
toward
union

The high commissioner, Lord Selborne, who succeeded Milner in 1905, argued the case for union in a memorandum published in July 1907, and, once in power, the three Afrikaner parties took it up enthusiastically, with the support of their oppositions. White public opinion in South Africa moved steadily in favour of union, partly out of an idealistic feeling that it was best to bury the hatchet and make a new start on a basis of white equality, and partly because an African rebellion in Natal in 1906 led many to think union essential for white security; some, moreover, saw in union the best way of guarding against any further British interference in South African affairs. There were also compelling economic reasons for union: the four colonies were interdependent and yet, without political union, their material interests were so divergent that the customs union seemed almost certain to collapse.

Accordingly a national convention, comprising 30 members appointed by the four colonial parliaments and three nonvoting members from Rhodesia, met in 1908 and 1909 under the chairmanship of Sir (afterward Baron) John Henry de Villiers, chief justice of the Cape. A constitution was drafted and unanimously approved by the convention and it was carried with scarcely any dissentients in the parliaments of the Cape Colony, the Transvaal, and the Orange River Colony, and by a three-to-one majority of the electorate in a referendum in Natal. It was enacted by the British Parliament in September 1909, substantially as it had been submitted to the British government by delegates from South Africa. The South Africa Act came into force on May 31, 1910, and the four colonies became the provinces of the Union of South Africa.

The convention, except for the Natal members, wanted a close union and complete flexibility, and they were largely successful; the principal feature of the South African constitution was the grant of power to Parliament to legislate on practically every subject by simple majorities in each house. Nevertheless, on two subjects they did not grant

Parliament such extensive powers. First, the Afrikaner members wanted special protection for the section which gave Dutch (to which Afrikaans was added in 1925) equal status with English as an official language of the Union. Second, the convention could not agree on a uniform franchise for the Union, the Cape delegates favouring a colour-blind franchise and the others a rigid colour bar. Eventually a compromise was reached whereby the franchise qualifications were to remain as they were in each province and the established franchise rights of the Cape nonwhites were specially protected. Accordingly, the sections on language equality and the Cape nonwhite franchise were to be amended only with the approval of two-thirds of the members of both houses of Parliament in a joint sitting. The act also made it possible for Rhodesia to join the Union on terms to be approved by the Privy Council and for the High Commission Territories of Basutoland, the Bechuanaland Protectorate, and Swaziland to join on terms laid down in a schedule, but only with the consent of the imperial government, which had special commitments to their African inhabitants.

THE UNION AND THE REPUBLIC

The race issue. The dominating factor in the history of South Africa has been the exceptionally complex character of the population. In 1910 there were nearly 6,000,000 inhabitants, of whom 21.5 percent were whites, 67 percent Africans, 9 percent Coloureds, and 2.5 percent Asians. By 1960 there were 16,000,000 of whom 19.3 percent were white, 68.3 percent Africans, 9.4 percent Coloureds, and 3 percent Asians. In 1910 the whites included an Afrikaner majority, mainly rural, and a British minority, mainly urban. Between them they owned most of the land and the capital, did most of the skilled work, and possessed 93 percent of the votes.

Living conditions of Africans

The Africans were still predominantly tribal rather than modern in culture, but, as a result of their conquest and of white missionary activity and economic enterprise, there was already a distinct trend toward the disintegration of tribalism. Although most Africans still had homes of sorts in the reserves, those were scattered lands that amounted to only one-fourteenth of the area of the country and by no means provided them all with a livelihood; consequently, many Africans went out to earn wages on white farms and in the towns. Moreover about 1,000,000 Africans had long been established as squatters on white farms and a few were already completely urbanized. The Africans were subject to pass laws, which restricted their movements outside the reserves, to Masters and Servants laws, which made breach of contract a criminal offense in many types of employment, to special taxation, and to other special laws. These laws varied from province to province, and in the Cape about 7,000 African voters were exempted from some of them.

The Coloureds were European in culture. Most of them lived in the Cape Province, where their legal status was that of the whites, and about 14,000 were voters, whereas in the other provinces they were subject to discrimination.

The Asians were mainly the product of the indentured Indian immigration scheme to Natal, where they outnumbered the whites and were subject to special laws, including a £3 tax. They were excluded from the Orange Free State, but about 11,000 had settled in the Transvaal, where, too, they came under special laws.

The policies of South African governments differed on many important issues; but all, being responsible to a predominantly white electorate, stood more or less explicitly for "the maintenance of white supremacy," a task that became increasingly difficult as a result of the rapid tempo of economic development in South Africa and the change in the balance of power and the climate of public opinion elsewhere.

The Botha–Smuts regime (1910–24). In 1911 the Afrikaner parties of the former colonies merged to form the South African Party, which, under Botha and Smuts, governed the country until 1924. Botha and Smuts had been zealous fighters for republican independence so long as there was a chance of success, but they were realistic enough to accept the inevitable at Vereeniging and imaginative enough to see great prospects for South Africa when the hot mood of British imperialism had spent itself. By 1910 they hoped for the coalescing of the British and Afrikaner elements into a white South African nation that would freely cooperate with the British Commonwealth in peace and in war. The ideal went too far, however, for the more conservative Afrikaners, who were concerned with preserving their group identity, and not far enough for the more self-conscious nonwhites, for whom it implied a permanently inferior status.

The government wished to keep most of the Africans in the reserves to prevent the whites from being swamped, and to use African manpower as the unskilled base of all forms of economic enterprise—objectives that involved certain contradictions. As the Africans became more accustomed to a money economy, some of them were liable, unless checked by law, to compete successfully with the less efficient whites. This was already happening in the gold-mining industry, and laws were passed in 1911 to preserve the racial hierarchy in that industry. Such competition was also happening in some rural areas, where Africans were pooling their resources to buy more land; a Natives Land Act was passed in 1913 to limit the areas within which such purchases could be made. Neither of these acts was wholly effective. In 1916 a commission reported that if the policy of territorial segregation was to be carried out the reserves should be substantially increased in size, but little was done about it because the whites were not prepared to make the necessary sacrifices.

Gandhi's nonviolent resistance against £3 tax

The Asian question had reached a more crucial phase. Already before union Mohandas K. Gandhi had organized nonviolent resistance against the Transvaal government, and after union he resumed the struggle against the £3 tax in Natal. As a result the Indian government stopped the importation of indentured Indian labourers in 1911 and the £3 tax was removed in 1914. By that time the pattern of later disputes had developed: the Union government regarded the Indians as temporary visitors and tried to persuade them to return to India; and most of the Indians regarded themselves as permanent residents and demanded full rights of citizenship.

The conservative Afrikaners had found their leader in J.B.M. Hertzog. Although he had joined the Botha Cabinet in 1910, Hertzog regarded his colleagues' policy as liable to cause the Anglicization of the Afrikaner people, and he publicly advocated a different, "twin-stream" policy, which led to his exclusion from a reconstituted Cabinet in 1912 and to the foundation of the Nationalist Party. When in 1914 the government unhesitatingly took the part of Great Britain in World War I and Parliament voted funds for the conquest of German South West Africa, a number of former republican generals, some of whom held appointments in the South African defense force, came out in rebellion to avenge Vereeniging. After mastering the revolt, Botha took command of the South West African expedition and forced the Germans to capitulate on July 9, 1915. South African volunteer contingents also fought in East Africa and on the Western Front in Europe and many individuals joined British units. Smuts, after serving as commander in chief in East Africa, did notable work in the British War Cabinet. In July 1919 he and Botha, on behalf of South Africa, signed the Treaty of Versailles and the covenant of the League of Nations, under which South West Africa became a Union mandate. Back in South Africa Botha died before August was out, and Smuts, succeeding to the premiership, faced the discontents caused by thwarted Afrikaner nationalism, a steep rise in the cost of living, and industrial troubles.

Smuts as premier

As early as 1907 Smuts had intervened on the side of the employers in a strike of white workers on the Witwatersrand, and in 1914 he had ended another strike by declaring martial law and illegally deporting nine strike leaders. After the war there was a serious crisis. In December 1921 the Chamber of Mines, faced with rising costs and a fall in the price of gold, decided to organize the industry more rationally by using Africans for semiskilled work. There was a violent reaction by white labour, which Smuts suppressed at a cost of 230 lives. The result was that, although the threat to relax the colour bar was not carried out, the Labour Party, representing the aggrieved white workers, made an electoral pact with the Nationalists. In the hope of redressing the political balance Smuts wooed the Southern Rhodesian electorate to accept incorporation in the Union, but when a referendum was held on the issue in 1922 they preferred to remain separate. Two years later the South African Party was heavily defeated at the polls and Hertzog became premier of a Nationalist–Labour coalition.

Hertzog's nationalist policy (1924–33). Hertzog's main objectives were to complete the emancipation of South Africa from imperial control and to provide greater protection for the whites from the Africans and for the Afrikaners from the British. He played a notable part in the events leading to the Balfour Report (1926) and the Statute of Westminster (1931), which gave statutory definition to the established convention that the British government could not exert authority over a dominion. South Africa's sovereign status was also asserted by the adoption, after a long and bitter controversy, of a distinctive national flag (1927), by the appointment of ambassadors to Italy, the United States, and The Netherlands (1929), and by the separation of the office of governor-general, the head of the South African government, from that of high commissioner, the representative of the British government in South Africa (1931).

Economic nationalism was fostered by the foundation of a state-controlled iron and steel works at Pretoria, by increased protection for agriculture and industry, by a reduction in imperial preferences, and by a commercial treaty with Germany. White supremacy was bolstered by the provision of sheltered employment for "poor whites" in state enterprises; by a Mines and Works Amendment Act (1926), which was more effective than its predecessor in shutting Africans out from skilled mining trades; by a Native Administration Act (1927) and a Riotous Assemblies Act (1930), which gave the executive wide powers over individuals; and by franchise acts (1930–31), which extended the vote to all white men and women, while they left the Cape nonwhite vote as before, restricted to men who possessed property and educational qualifications. Hertzog's legislative program was still incomplete when the world depression undermined the prosperity of the country and the popularity of the government, which continued to cling to the gold standard after Great Britain had abandonded it in 1931. After the government, yielding to pressure from some of its own supporters, left the gold standard in December 1932, there was a fairly rapid economic recovery, but political confusion continued until, in 1933, Hertzog and Smuts formed a coalition government that secured overwhelming support from the electorate.

Hertzog–Smuts regime (1933–39). The coalition was based on a great deal of common ground between Hertzog and Smuts. Consequently, in 1934 the two major parties fused to form the United Party; but Hertzog failed to carry with him a small group of Afrikaner irreconcilables, led by D.F. Malan, who formed the Purified Nationalist Party, while C.F. Stallard at the other extreme dissociated himself from Smuts and formed the Dominion Party. By then the constitutional settlement had been completed by the Status Act and the Seals Act (1934), which secured, so far as words could do, that South Africa was master in its own house. The government then proceeded to complete Hertzog's segregation legislation. A Representation of Natives Act (1936) removed the Cape Province African voters from the common roll and gave them the right to elect three white members to represent them in the lower house, gave the Africans throughout the Union the right to elect four white senators, and created a Natives' Representatives Council with advisory powers. A Native Trust and Land Act (1936) provided for a considerable increase in the size of the reserves. A Native Laws Amendment Act (1937) authorized the executive to prevent more Africans from entering the towns and to compel municipalities to segregate African from white residents. It was Hertzog's hope that the enlarged reserves would become capable of maintaining almost all the African people, so that those who worked for whites could be regarded as temporary visitors from the reserves. But it is notable that he saw the need for consultation with an African council, and that he did not try to discriminate against the Coloureds. Moreover, his government expanded the social services not only for the whites but also for the nonwhites; there was, for example, a considerable increase in government grants for education.

rther
gregation
gislation

Nevertheless, Africans, Asians, and Coloureds were becoming disturbed by the great contrast between their living standards and those of the whites and by an accumulation of laws that caused 500,000 Africans to be convicted of statutory and municipal offenses in a year. As tension began to mount, the Nationalists claimed that "white South Africa" would not be safe unless the restrictions on Africans were increased and the Coloureds were also segregated from the whites. Though Hertzog strongly disagreed, the Nationalists were able to profit from the celebrations held to mark the centenary of the Great Trek in 1938, when they appealed to all Afrikaners to remain true to the principles of the Voortrekkers and of Paul Kruger, placing special emphasis on their race attitudes.

World War II. The outbreak of World War II caused a crisis in South Africa. Although Hertzog and Smuts had cooperated successfully, they differed widely on foreign affairs. Hertzog took an indulgent view of Nazi Germany, believed that its expansion was no concern of South Africa, and moved in the House of Assembly, on September 4, 1939, that South Africa should remain neutral. Smuts, keenly aware of the wider implications of Nazism, maintained that it was the interest and the duty of South Africa to support Great Britain, and he won the crucial division by 80 votes to 67. On September 5 Smuts formed a ministry with the support of the Labour Party, the Dominion Party, and the majority of the United Party, and war was declared on Germany.

South
Africa's
support of
the Allies

The Smuts government concentrated its energies on the war. About 200,000 white men (more than half of them Afrikaners) and 125,000 nonwhites (mainly Africans and Coloured men) joined the forces, and many of them served with distinction in the Ethiopian, Mediterranean, and Madagascan theatres of war. The nonwhites were not allowed to bear arms but were distributed among the combatant units, for whom they performed vital work as stretcher-bearers, labourers, etc. Industry was efficiently switched to the production of munitions and clothing for military purposes. Smuts himself remained in close contact with Winston Churchill and watched carefully the strategy of the war, often leaving his able lieutenant, J.H. Hofmeyr (1894–1948), to carry a lion's share of the burden of administration.

South Africa's achievements were impressive, considering the strength of the opposition to the war. Fortunately for the government, the opposition splintered into fragments during the most critical period: J.F. van Rensburg's Ossewabrandwag (Guard of the Ox Wagon) and Oswald Pirow's New Order, accepting the racial doctrines of the Nazis and their contempt for parliamentary government, patently hoped to profit from a Nazi victory; the Malanites expounded the ideal of an Afrikaner republic in which, as in Kruger's republic, the British would not necessarily have political rights; while the Hertzogites stood firm by their principle of equality among the whites. In the first flush of his disappointment, indeed, Hertzog led his defeated minority from the United Party to join hands with Malan and form the Reunited Nationalist Party, but the reunion was not a happy one and before the end of 1940 Hertzog had been driven out for his tenderness toward the British. His colleague, N.C. Havenga, subsequently founded the Afrikaner Party to maintain his ideals.

The general election of 1943 was a victory for Smuts in that his prowar coalition secured a majority of 67 seats in a lower house of 153; but it was also in a sense a victory for Malan's Nationalists, who won all the 43 opposition seats. Thereafter, as the prospect of a Nazi victory faded, the Nationalists consolidated their position as the political instrument of self-conscious Afrikanerdom, absorbing elements from the Ossewabrandwag, the New Order, and the Afrikaner Party, and drawing support from a variety of cultural and economic organizations, including the Broederbond, a secret group of Afrikaner elite.

As the war drew to a close Smuts drafted the preamble to the United Nations Charter signed at San Francisco in 1945 and returned to mold postwar South Africa along the lines of a generous demobilization scheme, an expansion in the social services for all races, a planned development of agricultural, mineral, and industrial resources, and large-scale white immigration.

Industrial expansion in South Africa. Subsequent events can be understood only in relation to the fact that from about 1938 onward South Africa was experiencing an industrial expansion as intense as the British industrial revolution of the early 19th century, with the added complication of racial differences between the unskilled workers on the one hand and the skilled workers, the capitalists, and the majority of the electorate on the other. The way had already been paved by the rise of the mining industries and by the foundation of many manufacturing industries during World War I and the 1930s; the rate of expansion increased during and after World War II. In the 24 years between 1936 and 1960 the total population increased by 67 percent. In the 28 years between 1935 and 1963 the geographical national income at current prices increased more than eightfold (the value of mining production fivefold, of agricultural production sixfold, and of manufactures thirteenfold). In the same period the price index increased by two and a half times; so that, even allowing for the change in the value of money, the growth of manufacturing industries was remarkable.

The gold-mining industry continued to play an important role in the South African economy, especially in easing the balance of payments problem. Fears that the supply of gold would soon be exhausted were removed by the opening up of new mines on the Far West Rand and in the Orange Free State, where production started in 1951. Nevertheless, gold production did not increase as rapidly as other industries and the relative importance of gold mining decreased as the economy became more diversified. The mining industries continued to be organized on racial lines, and the manufacturing industries adopted a similar basis. Skilled work remained virtually a white monopoly and was well paid; most of the unskilled work continued to be performed by nonwhites for low wages.

Relative importance of gold mining

As a result of this expansion more than four-fifths of the white population were townsmen by 1960, the "poor whites" were absorbed by industry, and the earlier economic distinctions between the Afrikaners and the British inhabitants became blurred. Industry also attracted many nonwhites. More than 3,500,000 Africans (nearly a third of the total African population) were in the towns at the time of the 1960 census. Into some of these towns the influx was so rapid that housing and other amenities were grossly inadequate and there was much crime. But in spite of low wages and squalid living conditions, the urban Africans earned far higher incomes than the rural Africans did.

The 1948 election. As the 1948 election approached, disturbing facts were becoming known in South Africa. Although the reserves had been enlarged under the 1936 act, only 40 percent of the African population was to be found in them at the time of the 1946 census. The Social and Economic Planning Council reported that the quality of the reserve lands was deteriorating; and early in 1948 a commission revealed that many Africans were becoming permanent inhabitants of the towns. These facts did not square with the argument that the Africans had adequate homes in the reserves and could therefore be treated as inferiors elsewhere. Moreover the Natives' Representative Council was frustrated because the government ignored its advice, and the African National Congress, which had been cautious since its foundation in 1912, was now demanding the removal of discriminatory laws in terms of the ideals of the United Nations. This raised the question of whether white supremacy, which had developed in the simple pastoral economy of the 18th century, could be maintained in the industrial economy of the 20th century.

Demands of the African National Congress

The Smuts government continued to profess adherence to the segregation policy but tried to deal with practical grievances in a conciliatory spirit, while Hofmeyr went further than his colleagues on occasion, feeling toward a policy that might be acceptable to all races. But Asians were protesting vigorously against legislation that prevented their buying or occupying new premises in Durban, and the Indian government attacked South Africa in the United Nations General Assembly. Consequently, when Smuts tried to persuade the assembly in 1946 to allow the Union to incorporate South West Africa, he met with a rebuff, and the assembly passed the first of a series of resolutions condemning racial discrimination in South Africa.

The Nationalist leaders wished to reduce South Africa's links with Britain, the Commonwealth, and the United Nations, to advance the power of the Afrikaner people, and, above all, to preserve white supremacy by every possible means. They expressed horror at the "liberalism" of the United Party, declared it would cause the "suicide" of white South Africa, and proclaimed a policy of apartheid, according to which whites, Coloureds, Asians, and Africans would be separated from each other and each race would be able to "develop along its own lines in its own area." Though as far as the Africans were concerned this was little more than Hertzog's segregation policy in a new dress, it was presented to the electorate as something new and dynamic. The Nationalists also exploited the discontents caused by wartime controls and were assisted by the fact that the electoral system favoured the rural voters. The result was that in May 1948 Malan, with the assistance of Havenga's small Afrikaner Party, was able to form a government, with a majority of five in the House of Assembly. This result, which was a surprise to Malan as well as to Smuts, was decisive for South Africa.

Nationalist Party policy from 1948. After the 1948 election the Nationalist Party consolidated its power, absorbing the Afrikaner Party and gaining strength in the House of Assembly in each election—from 73 seats in 1948 to 94 in 1953, to 103 in 1958, to 105 in 1961, and to 126 (in a larger House) in 1966. The process was promoted by legislation and administrative action: Smuts's immigration scheme was scrapped; British immigrants, like others, were required to wait several years before they could vote; the white people of South West Africa were given six seats in the House of Assembly; the voting age was lowered to 18; and the Coloured voters were given separate and limited parliamentary representation and the representatives of the Africans were removed from Parliament. Nationalist power was also promoted by the strong leadership of the prime ministers Malan (1948–54), J.G. Strijdom (1954–58), and H.F. Verwoerd (1958–66). The United Party, led by J.G. Strauss after the death of Smuts in 1950 and by Sir D.P. de Villiers Graaff from 1956, remained the main opposition party, but it lost seats at each election and suffered from secessions on both flanks. Several new parties were founded, including a Liberal Party and a Progressive Party, but they had little success at the polls and the former party has since been disbanded. The first indication of possible change came in 1970 when the Nationalist Party lost nine seats in a general election. The Nationalist Party, however, still retained more than 70 percent of all seats in the assembly (117).

The Nationalist governments enacted a mass of racial legislation to preserve white supremacy. They created a population register to fix the racial category of every South African. They made marriages and unions out of wedlock between whites and nonwhites unlawful. They systematically divided the towns as well as the rural areas into zones in which members of only one race could own or occupy property or conduct business. They assumed control over African and Coloured school education, eliminating the mission schools. They excluded nonwhites from the established universities and founded separate colleges—one each for Coloureds, for Asians, for Xhosa, for Zulu, and for Sotho. They intervened in the labour-union movement to separate white from nonwhite. They gave officials sweeping powers to remove "undesirable" Africans from towns. They extended the practice of reserving particular types of jobs for whites. They enforced segregation where it did not previously apply, as in buses, trains, post offices, libraries, motion-picture houses, and theatres in the Cape Peninsula. White and nonwhite South Africans could thus rarely meet, except as masters and servants or rulers and subjects.

xclusion
f
onwhites
·om the
uthorita-
ve
olitical
/stem

Nonwhites were almost totally excluded from the authoritative political system. After a long struggle the Cape Coloured voters were removed from the common voters' roll and given the right to elect four whites to represent them in the House of Assembly. At first the government lacked the two-thirds majority laid down in the South Africa Act for such legislation and, therefore, in 1951 attempted to pass it through Parliament by the ordinary simple-majority procedure. It was declared invalid by the Supreme Court. Under a further act the Nationalist members of Parliament then sat as a high court and reversed the judgment, but the Supreme Court held that the "high court of Parliament" was itself invalid. Eventually the Senate was enlarged and its composition changed to give the Nationalists 77 seats out of 89 in 1955. Coloured voters were then removed from the common roll by the necessary two-thirds majority in a joint sitting of both houses.

In 1959 the representatives of the Africans were removed from both houses of Parliament and the Cape Provincial Council. The House of Assembly then contained 156 whites elected by the white voters of the Union and South West Africa and four whites elected by the Cape Coloured voters. In 1960 the Senate contained 43 whites elected by white electoral colleges and 11 nominated by the government. All provincial councillors were whites elected by white voters, except two in Cape Province who were elected by Coloureds. Thus the compromise embodied in Hertzog's 1936 legislation was destroyed.

The only nonwhites who had any say in the composition of the Parliament or of the Provincial Councils were the Coloured people of Cape Province, and their say was meagre. Instead, nonwhites were given various types of segregated subordinate institutions. For the Africans there was a system of "Bantu authorities" in the Bantu areas, consisting mainly of chiefs, councillors, and their nominees. In the Transkei, the largest Bantu area, a Legislative Assembly, with a majority of chiefs ex officio and a minority of members directly elected by Africans, and a cabinet and prime minister responsible to the assembly were established in 1963. Such bodies had powers in defined fields, subject to the overriding authority of the government of South Africa. In the towns some Africans could elect members of Bantu urban councils. An advisory Coloured Persons Representative Council with an elected majority was established in 1964.

The Nationalists were able to transform South Africa into a republic after a national referendum in 1960 narrowly approved a new constitution that reflected the change. In March 1961 Verwoerd asked a Commonwealth prime ministers' conference whether South Africa might remain a member of the Commonwealth when the change took place, but he withdrew his request when other members of the conference criticized his government's racial policies. Consequently when South Africa became a republic on May 31, 1961, it left the Commonwealth. The republican constitution was substantially the same as the constitution of the Union, except that an indirectly elected state president replaced the British monarch and the governor-general as head of state.

epublican
›nstitu-
›n and
ithdrawal
›m the
›mmon-
ealth

Internal opposition to Nationalist rule. In spite of all the separatist laws the peoples of South Africa were becoming more and more interdependent as a result of the economy's continuous growth, and many South Africans, white as well as nonwhite, opposed the government's racial policies.

African opposition was more serious. The leaders of the African National Congress (ANC), with the cooperation of Coloured and Asian congresses and small white groups, organized a series of campaigns to elicit the support of the African masses and to intimidate the government into making concessions. In 1952–53 several thousand people, mainly Africans, deliberately courted arrest by "passive resistance to unjust laws." For 10 weeks in 1957, 45,000 Africans walked many miles each day to work in Johannesburg and Pretoria rather than pay increased bus fares. There were also intermittent riots and clashes with the police in many parts of the country. These reached a high point in 1960 when the Pan-Africanist Congress (PAC), a radical offshoot of the ANC, organized demonstrations against the pass laws; at Sharpeville, near Johannesburg, the police opened fire against demonstrators, killing 69 Africans and wounding 178. Blood was also shed in Cape Town and elsewhere. Between 1959 and 1961 there was a revolt against the Bantu authorities system in the Transkei.

Parliament enacted many laws to help the executive break the revolutionary opposition. The nonviolent resistance campaign of 1952–53 was called off in the face of harsh new penalties for protest actions, and in 1960 order was restored by the declaration of a state of emergency. The Transkei revolt was handled under proclamations giving extraordinary powers to chiefs and white officials. General Law Amendment acts allowed the executive to outlaw any organization and to ban persons from attending meetings or publishing statements and to arrest and hold them incommunicado without trial. The ANC and the PAC were banned in 1960, and by the end of 1964 all known revolutionary leaders were in prison or exile, and all known revolutionary organizations had been outlawed.

Foreign opposition. The governments of the newly independent African states wished to complete their revolution by destroying the last strongholds of white supremacy in Africa. Although they had the diplomatic support of the Asian and Communist countries, they did not have the resources to coerce South Africa without the cooperation of South Africa's principal trading partners—Britain, western Europe, and the United States. They therefore tried to use the United Nations as an instrument of coercion. The General Assembly had condemned South Africa's racial policies since 1948; the Security Council since 1960. In 1962 the assembly, by majority vote, called on member states to apply economic sanctions against South Africa and to break off diplomatic relations; in 1964 the Security Council set up an expert committee to consider the feasibility of sanctions. Britain, France, and the United States, although subscribing to many resolutions criticizing apartheid, did not commit themselves to coercion. In the mid-1960s, when the UN was financially weakened and the major powers were distracted by the conflict in Vietnam, the prospect of a UN blockade of South Africa diminished. By then South Africa had well-equipped and well-trained military forces, consisting of about 125,000 white men, who could be rapidly mobilized and who were capable of overcoming any but a very large-scale attack.

UN call for
economic
sanctions
against
South
Africa

In October 1966 the UN General Assembly voted to revoke South Africa's mandate over South West Africa. The UN Council for South West Africa voted in 1968 to rename the territory Namibia. Renewed pressure on South Africa to relinquish its administration of South West Africa came with a ruling of the International Court of Justice against South Africa in 1971.

(L.M.T./R.J.Da./Ed.)

South Africa since the mid-1960s. Verwoerd, the prime minister, was assassinated in September 1966 and was succeeded by B.J. Vorster, a member of the Nationalist Party's extreme right wing. Under Vorster, restrictions against the nonwhite populations were tightened, and segregation was even more strictly enforced. This led to increasing unrest among the Asians and Coloureds as well as Africans, who saw even the small rights they possessed being stripped away. The UN moved repeatedly to isolate South Africa from the world community. Vorster further alienated world opinion by his deployment of South African troops to fight the Soviet-backed Popular Movement for the Liberation of Angola (MPLA) in the Angolan civil war.

In mid-June 1976 the worst race riots in South African history broke out in Soweto (South-Western Townships), the residential area reserved for Africans near Johannesburg. The riots began when African students marched in protest against a new rule that made Afrikaans the compulsory language of instruction in African schools. The police responded with tear gas and made mass arrests, and fighting erupted. After nearly three weeks of constant battle, massive destruction of property, and increasingly repressive police action, the riots were finally suppressed.

Along with increasing internal dissent, South Africa faced further conflicts on the question of independence for

Namibia. A main point of contention with Namibia was control over Walvis Bay, which was finally conceded to South Africa on September 1, 1977. But as independence talks continued without any tangible results, the South West Africa People's Organization (SWAPO), representing Namibia, conducted guerrilla raids into South Africa. In March 1990 Namibia gained its independence. (For further treatment of Namibia, see SOUTHERN AFRICA.)

P.W. Botha, a prominent Nationalist, became prime minister in 1978 upon Vorster's retirement. The decision to draft a new constitution that would extend the franchise to the Coloured and Asian population caused a split in the National Party, resulting in the formation of the new Conservative Party, which objected to alterations in the current system. The new constitution was enacted in 1983 and implemented in September 1984. In addition to extending the franchise, it created a tricameral parliament, each house being elected separately by the white, Coloured, and Asian voters. Criticism of the constitution came from both the right and the left; the right objected to the relinquishing of complete political control, while the left asserted that the constitution further entrenched apartheid and failed to provide representation for the majority black population. The United Democratic Front, a coalition of some 30 multiracial organizations, organized boycotts against the elections of the Coloured and Asian parliaments, and voter turnout was less than 31 percent for both elections. The implementation of the constitution sparked riots and killings in the black townships. Conflict intensified until a state of emergency was declared in July 1985. Violence continued, however, and except for a short period in 1986, the state of emergency was maintained throughout the rest of the 1980s. Many countries enacted economic sanctions against South Africa to condemn the government's actions, and the lack of confidence abroad caused the rand to fall and investment to slacken.

The constitution of 1983

In 1989 Botha, after suffering a stroke, resigned and was replaced by Frederik W. de Klerk, a former minister of finance. Leading the National Party in the September 1989 elections, de Klerk campaigned for negotiations on a new constitution with leaders of all racial groups, including those of the African community. While the Nationalists held onto a majority in the all-white House of Assembly, they lost significant support to the Conservative Party and the newly formed liberal Democratic Party. In early 1990 de Klerk announced that the Separate Amenities Act, as well as prohibitions and restrictions on more than 60 political organizations—including the ANC, PAC, and the South African Communist Party—would be rescinded. Nelson Mandela, the leader of the ANC who had been in jail since 1962, was released and began negotiations on political change with the government.

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

For coverage of related topics in the *Macropædia* and *Micropædia,* see the *Propædia,* sections 945 and 978, and the *Index.* (Ed.)

Cape of Good Hope

PHYSICAL AND HUMAN GEOGRAPHY

Cape of Good Hope, or Cape Province—officially Kaapprovinsie in the Afrikaans language—is situated on the southern extremity of the African continent and is the largest of South Africa's four provinces. The name Cape of Good Hope, while specifically referring to the promontory about 30 miles (48 kilometres) south of Cape Town, officially applies to the entire province. Occupying the southern and western region of the republic, the province has an area of 278,381 square miles (721,003 square kilometres), including the territory of the three republics, Transkei, Bophuthatswana, and Ciskei, within its borders and the exclave of 434 square miles around Walvis Bay on the coast of Namibia. In addition, 1,476 square miles of territory were transferred from Griqualand East to Natal Province in 1978. Cape Town, the provincial capital and the legislative capital of the republic, is the largest city.

The land. *Relief.* Most of the province lies at an altitude of between 3,000 and 5,000 feet (900 and 1,500 metres) above sea level, occupying part of the great interior plateau of southern Africa. The edges of the plateau form conspicuous escarpments that trend southward from the Namaqua Highlands in the northwest to the Roggeveldberg in the southwest and from there eastward as the Nuweveldberge, Sneeuberg, and Stormberg mountains until the southern end of the Drakensberg escarpment is reached. These escarpments separate the plateau from a coastal zone with a varied relief in the west and from the broad basins of the Doorn Karoo and the Great Karoo plateaus to the southwest and south. Across the southern part of the province a series of sandstone ranges of equal height enclose the elongated basin of the Little Karoo, running approximately parallel to the south coast. At the western end of these ranges, a complex of similar ranges runs northward from Cape Hangklip, southeast of Cape Town, to the Bokkeveld, about 200 miles to the north. Lowlands, between the mountains and the coast, are narrow and discontinuous, apart from a strip 10 to 50 miles wide between Mossel Bay and Cape Agulhas (the southernmost tip of the African continent) and an area 20 to 60 miles wide between False Bay (immediately south of Cape Town) and St. Helena Bay (about 100 miles to the north).

Climate. Maritime influence on the climate is restricted to the coastal fringe. Over most of the province the mean annual precipitation is less than 15 inches (380 millimetres), and large areas in the northwest receive less than 5 inches. Years of drought are common. Only the southwestern and southern margins receive more than 25 inches of precipitation, mostly in the form of rain resulting from winter cyclones in the southwest, and summer rain resulting from easterly winds beyond Mossel Bay on the south coast. In the interior most of the rain is from summer thunderstorms; the occasional winter rains and the rarer snowfalls are associated with the passage of cyclones. Solar radiation is intense at all seasons, because of the latitude, the elevation, and the usually clear skies; inland mean daily maximum temperatures usually range from 65° F (18° C) in winter to 90° F (32° C) in summer. Frost occurs on more than 60 nights a year; the frostless season lasts for less than 240 days on the southern margin of the Upper Karoo. Snow occasionally mantles all the higher mountain peaks, and only the coastal zone is frost-free.

Droughts

Plant and animal life. The vegetation is predominantly xerophytic (adapted for growth under dry conditions). Thornbush savanna (parklike grassland) occurs in the north, while low-growing Karoo bush vegetation, typified by many species of plants adapted to semidesert conditions, occurs south of the Orange River. Desert shrub and succulents (juicy plants) are found in Namaqualand in the northwest. In the southwestern region, where winter rainfall occurs, the sclerophyll bush (characterized by plants with thickened hardened foliage) is extremely rich in endemic species. Indigenous forest, predominantly evergreen, is found only on the south coast between Mossel Bay and Algoa Bay and on the southern slopes of the Amatole Range, to the northeast of Algoa Bay, where rain occurs at all seasons; farther east it passes into grassland and, in the coastal zone, into xerophytic low forest and bush.

Immense herds of springbok and other antelope, wildebeest (gnu), and zebra formerly ranged the interior, while hippopotamuses were found in all the perennial rivers, and elephant and rhinoceroses roamed the coastal zone from Natal in the east to the Olifants River in the west; lion, leopard, and other carnivores were ubiquitous. Though many local place-names recall this rich and varied animal life, today, apart from the smaller species, it has mostly disappeared except in a few scattered game reserves, in some sparsely settled areas where a few springbok still roam, and in some remote mountain areas where the rare leopard preys on the few surviving baboons. Birdlife, however, remains numerous and varied. Along the coast, particularly in the west, cormorants and other seabirds abound, while inland each vegetation zone is associated with its own characteristic birds, including, in summer, great numbers of migrants from the Northern Hemisphere.

The people. The government recognizes four population groups: Cape Coloureds (of mixed descent, principally

Khoi, white, and Indonesian), Africans, whites, and Asians (mostly Pakistanis and Indians). About three-fourths of the population lives in the urban areas of the four largest cities: Cape Town, Port Elizabeth, Kimberley, and East London. Most of the province is sparsely populated. The highest rural population densities are in a coastal zone from St. Helena Bay to East London and in a few irrigated areas inland, notably along the Vaal and Orange rivers.

frikaans, nglish, hosa, and otho

Official languages are Afrikaans and English. Afrikaans, a derivative of Nederlands (the official name of the Dutch language), is the first language of more than half of the whites and provides a powerful cultural bond for Afrikaners. It is also the first language of nearly 90 percent of the Coloureds; the remaining whites and Coloureds are English-speaking. Xhosa is the language of 80 percent of the Africans, Tswana (mainly in the northern interior) and Sotho of most of the remainder.

More than half of the whites and a third of the Coloureds belong to the Nederduits Gereformeerde and allied churches (Dutch Reformed churches); about a fourth of these groups belong to Anglican or Methodist churches. Malays, constituting about 10 percent of the Coloureds, are adherents of Islām. Two-thirds of the Africans are Christian, more than a fifth being Methodists and lesser numbers members of independent churches.

The economy. *Resources.* Mineral resources are extensive, but mining employs relatively few people. Copper has been mined in Namaqualand since 1852, and diamonds in Griqualand West since 1870. Since 1928 diamonds also have been recovered in Namaqualand from deposits on coastal terraces and, more recently, from submarine deposits offshore. Blue asbestos is mined in a zone extending northward from Prieska to beyond Kuruman. Enormous reserves of manganese and high-grade iron ores also are found in Griqualand West; in 1976 a 540-mile railway was completed from Sishen to a new iron-ore-loading terminal accommodating 250,000-ton vessels at Saldanha Bay.

Agriculture and fishing. Although the economy is fairly diversified, farming remains of paramount importance. In a normal year, of South Africa's total farm production, two-thirds of the wheat and wool, 70 percent of the alfalfa (lucerne) hay, most of the mohair, the deciduous fruit, and the grapes, and all of the wine are produced in the Cape. Sheep are raised in the Karoo and cattle in the thornveld north of the Orange River. Alfalfa, sultanas, and cotton are grown under irrigation along the Orange River near Upington; alfalfa and peanuts (groundnuts) are grown in the Harts Valley. Ostriches still graze in the Little Karoo, but alfalfa seed, hay, fruit, tobacco, wool, and mohair have long surpassed ostrich feathers in importance. In the winter rainfall region the lowlands are given over to grain cultivation, sheep raising, and dairy farming, the foothills to vineyards and orchards, and the upland basins to deciduous fruits. These activities support a lively export trade in fresh fruit, table grapes, wine, brandy, and canned goods. In the higher rainfall areas, particularly between Mossel Bay and Humansdorp on the south coast, extensive timber plantations have been established; farther east, citrus fruit is grown under irrigation in the valleys of the Gamtoos, Sundays, Great Fish, and Kat rivers. Chicory and pineapples are produced between Algoa Bay and East London.

Rich fishing grounds lying offshore from the Agulhas Bank to Namibia are actively exploited by South African and foreign vessels, supplying an important fish meal and canning industry.

Industry. Industries include food processing, marine engineering, ship repairing, and the manufacture of textiles, clothing, footwear, motor vehicles, tires, agricultural fertilizers, pesticides, mining explosives, and pharmaceuticals.

Transportation. Harbours and public rail, road, and air transportation services are operated mainly by the South African Railways and Harbours Administration and its subsidiary, South African Airways. Ships drawing up to 40 feet can berth at Cape Town and Port Elizabeth, and ships drawing up to 35 feet at East London; there are two graving docks at Cape Town (one more than 1,000 feet long) and another at East London. Port Elizabeth has ore-loading facilities for iron and manganese, and there are grain elevators at Cape Town and East London. There is an extensive railroad and highway system. South African Airways operates air services linking the port cities and Kimberley to Bloemfontein, Johannesburg, Pretoria, and Windhoek and also international services to Europe, Australia, South America, and the United States.

Administrative and social conditions. *Government.* Administrative authority is shared by the province and the republic, though most functions are discharged by the republic. Until the late 1980s an administrator appointed by the state president and a council of 55 members chosen by the provincial white electorate directed the administration of primary and secondary schools and teacher-training institutions, as well as of hospital, public health, library, and conservation services. The provincial council, however, was abolished in 1986, and its powers were transferred to the three chambers of Parliament and to a new executive authority appointed by the state president and headed by the administrator. In the early 1990s the provincial government was still in a process of transition.

Tribal territories

The area east of the Great Kei River, formerly consisting of tribal territories that were given a considerable degree of local autonomy in 1894, was formed into a semiautonomous territory for Xhosa-speaking peoples within the republic in 1963. This region, Transkei, is now administered as a republic with a president and a legislative assembly of paramount chiefs, tribal chiefs, and members chosen by the African electorate, with executive authority invested in a cabinet of ministers. Bophuthatswana, consisting of an area in the north of the Cape of Good Hope, together with areas in the provinces of Transvaal and Orange Free State, was similarly constituted as a republic for Tswana-speaking peoples in 1977. It is administered by a president and an elected legislature. Ciskei adopted a republican form of government in 1981.

Health and welfare. Both birth and death rates are considerably higher among Coloureds than whites. The higher Coloured death rate is attributed to high infant mortality, as well as greater incidence of diseases such as gastroenteritis, pneumonia, and tuberculosis that are associated with lower standards of living and hygiene. Meaningful data for the African peoples, among whom registration of births and deaths is not required, are unavailable.

Free hospital services are provided for lower income groups by the administration and by charitable health services. Smallpox is no longer a hazard, but immunization against poliomyelitis is compulsory and free to all.

School attendance is compulsory for all white and Coloured children from 7 to 16 years of age. Among most Africans attendance is optional, and educational facilities for them still lag behind population growth. As late as 1970 more than half the economically active Africans had received no formal schooling, although most urban and many rural Africans over age 15 were literate in at least one language. By 1980 an estimated three-fourths of the Africans of school age, about one-fifth of the total African population, were attending school. Universities have been established for all population groups: for whites at Cape Town, Grahamstown, Port Elizabeth, and Stellenbosch; and for Coloureds and Asians at Bellville. A university for Bantu-speaking peoples is located at Fort Hare in Ciskei.

Colonial artifacts

Cultural life. The South African Library (founded in 1818), Museum (1825), and National Gallery (1871) are located in Cape Town. There are noteworthy museums in all the larger urban centres and numerous smaller museums in towns such as Graaff-Reinet, Swellendam, and Tulbagh that foster interest in cultural history, including the work of Cape silversmiths and other colonial artifacts. The colonial heritage is also manifest in traditional country dances and folk songs and the largely indigenous Cape Dutch architecture of many farmhouses in the western part of the province. The provincial library operates branches in most towns and serves rural areas with mobile libraries.

There are several performing arts organizations, especially in Cape Town, where the city's orchestra has presented regular concerts since 1914 and the university has operated a theatre since 1931 and a ballet school since 1934. The Eoan Group, a Cape Coloured organization, fosters operatic talent through amateur productions of professional standard. Since the establishment of the Cape Performing

Arts Board in 1963, government subsidies have enabled it to send touring companies to perform outside Cape Town. Since 1971 the board has staged professional ballet, operatic, and theatrical productions throughout the year in its new complex in Cape Town.

HISTORY

Dutch settlement

The first European settlement in southern Africa was located in what is now Cape Province. This was the refreshment station established in 1652 by the Dutch East India Company at Table Bay, 30 miles (50 kilometres) north of the promontory that the Portuguese navigator Bartolomeu Dias de Novais had named Cabo da Boa Esperança (Cape of Good Hope) in 1488. In 1814 the Cape settlement was ceded to the British and became the Colony of the Cape of Good Hope. In 1910 the colony was constituted a province of the Union of South Africa (since 1961 the Republic of South Africa).

As the early agricultural ventures of the Dutch East India Company proved inadequate to supply the garrison and ships in transit, lands were assigned to independent settlers in 1657 and, to supplement their labour, from 1658 slaves were imported, at first from West Africa and later in larger numbers from Madagascar, Ceylon, and the East Indies. As the amount of livestock obtainable from the nomadic Khoikhoin (called Hottentots by the Dutch) to supply ships also proved insufficient, the company itself began to raise stock and later to encourage settlers to do so. At first the paucity of grazing during the summer dry season necessitated the establishment of pastoral outposts in well-watered localities in the immediate hinterland; later, however, colonial pastoralists moved inland seasonally in search of good grazing, as had the Khoikhoin.

By 1700 grain, wine, and fruit farms were scattered along the eastern foot of Table Mountain (which overlooks Table Bay) as well as along the western foot of the Hottentots-Hollandsberge and Drakenstein ranges 30 to 40 miles to the east, while pastoral farmers ranged as far as 100 miles to the north and east of Table Bay. To this encroachment the Khoikhoin offered no sustained resistance, withdrawing before the colonists or accepting service with them as shepherds, guides, and interpreters; in the smallpox epidemic of 1713 thousands died. The survivors fled, and by 1730 no tribal remnants remained within 250 miles of Table Bay. Concurrently, the growing volume of shipping and the consequent demand for livestock provided the incentive for the spread of colonial sheep farmers northeastward beyond the Roggeveldberg and Hantamsberg and eastward through the Karoo (a plateau region) until, by 1779, they were in contact and conflict with Bantu-speaking peoples along the Great Fish River.

Stock thefts and reprisals subsequently resulted in a series of frontier wars; the introduction of some 3,500 settlers by the British in 1820 between the Dutch settlers and the Africans failed to establish an effective buffer settlement, and pacification was achieved only after a series of annexations had by 1894 advanced the frontier to the Mtamvuna River, the southwestern border of the colony of Natal. All the tribal territories east of the Great Kei River were thus reduced to the status of tribal reserves under colonial administration.

The gradual northward diffusion of pastoralists across the Upper Karoo meanwhile had led to the discovery of diamonds in 1867–68 where the trail to Bechuanaland—followed for half a century by traders and missionaries—crossed the Orange River.

The diamond rush

The discovery of diamond-bearing intrusions north of the river precipitated a rush of immigrants. It also led to the annexation of Griqualand West by Cape Colony in 1871, despite conflicting claims by the Orange Free State; to the extension (between 1873 and 1884) of the small Cape Town-to-Wellington railway more than 600 miles inland to Kimberley; and to the construction of railways inland from Port Elizabeth and from the port of East London. The fortuitous discovery of diamonds was particularly timely, since the opening of the Suez Canal in 1869 threatened to bankrupt the colony. The era that followed was one of unprecedented population growth and economic activity that attracted capital from overseas. The attention paid to the potential, although then still unknown, mineral wealth of the hinterland gave impetus to expansion. Political manoeuvring culminated in 1885 in the proclamation of a British protectorate over the area extending north to 22° south latitude, and in 1895 the section south of the Molopo River, the present boundary with Botswana, was incorporated within the Cape Colony. (W.J.T./Ed.)

Natal

Natal, the smallest of the four provinces of the Republic of South Africa, is bounded on the east by the Indian Ocean, on the south by Cape Province, on the west by Lesotho, the Orange Free State, and Transvaal, and on the north by Swaziland and Mozambique. Its area, which includes Zululand and Tongaland, is 33,578 square miles (86,967 square kilometres). It was given its name by Vasco da Gama, who, on Christmas Day 1497, sighted the entrance to the harbour of Port Natal, today called Durban. Its capital is Pietermaritzburg. The province attained its present extent by 1902.

PHYSICAL AND HUMAN GEOGRAPHY

The land. *Relief.* Natal is generally hilly or mountainous, especially along its western border, and rises from the coast to over 11,000 feet (3,353 metres) at the Drakensberg Escarpment (the edge of the interior plateau of South Africa and of Lesotho). The slope, however, is not gradual and is converted by various outcrops into steps of undulating land separated by escarpments. The lowest step is the coastal plain, less than 500 feet above sea level, widening to the north; then follow surfaces at 2,000 to 2,500 feet and at 3,500 to 4,500 feet (parts of which are referred to as the Midlands). Beyond the province, to the west, lies the Highveld, or high plateau. The continuity of these steps is often interrupted by spurs extending from the Drakensberg and by rivers—notably the Tugela—that cross the province in troughs hundreds of feet below the surrounding countryside.

Relief and drainage

Climate. The climate is generally related to major relief features. Rainfall decreases from more than 50 inches (1,270 millimetres) annually along parts of the coast to 30 to 40 inches inland, and then increases to well over 60 inches in the Drakensberg. About 60 percent of the coastal rainfall and 85 percent in the west falls in the summer months, October to March, often as heavy afternoon showers associated with thunderstorms. Temperatures decrease with altitude from the frost-free coast (January 70° to 75° F [21° to 24° C], July 60° to 65° F [16° to 18° C]) to the west (January 60° to 65° F [16° to 18° C], July 40° to 45° F [4° to 7° C]). There are significant day-to-day changes in weather and innumerable local climates due to minor differences in relief. In sheltered valleys or basins annual rainfall is often below 25 inches, and temperature inversions occur in winter (April to September); sea-facing slopes receive higher rainfall, while in the Midlands, mist belts often form in summer along escarpments and hot mountain winds blow from the interior occasionally. In general, summers are hot with occasional rain, and the humidity is high near the coast; hence, the Drakensberg mountain resorts are well patronized at this season. The warm, dry, and sunny winters have made the coast the principal holiday playground of southern Africa.

Plant and animal life. The natural vegetation, where it has not been modified or destroyed, falls into several belts according to altitude and climate. Tropical forest or bush, mangrove-fringed in the north, occurs in patches on the more open land along the coast. Farther inland in the north is bushveld, or savanna (grassy parkland), while farther south, in the river basins and valleys, cactus-like euphorbias (spurges) are characteristic. The Midlands and southwest sustain temperate forests, but these now exist only in relict patches. Open country, studded with thorn trees, covers the western half of the province; with true grassland occurring only at the high altitudes.

Game and nature reserves

The province, especially its coastal margin, is regarded as a transitional corridor along which tropical and more temperate animal life and vegetation have spread and intermingled. Measures for conservation are taken. The

Natal Parks Board controls 34 parks or game and nature reserves, with a total extent of 527,000 acres (213,277 hectares), which provide facilities for wilderness trail hiking. The larger reserves are situated in two different environments—along and behind the Zululand coast and in the Drakensberg. In Zululand, rhinoceros, buffalo, hippopotamus, and crocodile are characteristic varieties of animals. Fishing owls and fish eagles are also found, while vegetation includes the fever tree and the crane flower. In the Drakensberg are found the eland (a large African antelope) and the baboon, the martial eagle, and the lammergeier (bearded vulture). Vegetation includes tree ferns and ground orchids. Trout are found in the rivers of the interior, while along the coast, fish catches range from shad to sharks; in winter, sardine or pilchard shoals attract many game fish.

The people. The people of Natal form a mosaic of different ethnic groups living together within a complex social structure. As elsewhere in South Africa, the main distinction is between two composite groups, white and black, who, in accordance with the government's policy of apartheid (separate development), live in different but coexistent worlds. The whites are mainly of British or Dutch descent, while the nonwhites (according to South Africa's classification system) include an African majority, a number of Asians (Indians), and a group of Coloureds who are the product of interbreeding. Most of Natal's people live along and behind the coast (except in the north) or in the centre of the province; the extreme west and northeast are lightly populated. The Africans are especially concentrated in many fragmented areas (formerly referred to as Bantu Homelands, later called Black States), which, in 1970, were designated as the Zulu Territorial Authority and granted some autonomy. These homelands cover about a third of the province, and—except in the northeast—consist of broken rugged country. Whites own about two-thirds of the land in Natal although they compose only one-fourth of the population. Most whites live in or near Durban, in the towns along the route running west to the Witwatersrand ridge on South Africa's central plateau, or along the coast. The vast majority of South Africa's Asians live in Natal, and the Asians and Coloureds have a distribution similar to that of the whites.

Urbanization is proceeding apace. The towns are a European innovation, superimposed upon a previous rural settlement pattern, but they have proved increasingly attractive to the other population groups who live in separate residential areas within them, and few towns now have a majority of whites. Durban and to a lesser extent Pietermaritzburg dominate the urban scene. The other towns are few in number and are considered small in size.

Each population group has a cultural identity reflected in its religion, language, and customs. Most of the whites have English as their basic language, most of the remainder speak Afrikaans, and many whites speak both languages. The whites are mainly Christians; the Coloureds have a close affinity to the whites and have a similar culture. The Asians, on the other hand, have adopted many elements of the Western culture, but they retain an identity and a diversity by maintaining much of their own cultures. Some 70 percent profess Hinduism, 20 percent are Muslims, and the remainder are largely Christians. Most are fluent in English, but most speak various Indian languages at home. The Africans have not been culturally assimilated to the same extent as the other non-whites. They have retained their identity through their use of Zulu, a Bantu language, which is the language of the overwhelming majority of the Africans, and through a rich heritage of folklore, ceremony, and customs that reflect the diversity of tribal allegiances. The rest of the Africans speak cognate languages.

The economy. Natal has two coexistent economies. One is the subsistence economy of the Black States based upon pastoralism (largely cattle raising) and cultivation (mainly of corn [maize] grown in small plots) and supplemented by migrant labour. The Black States have poor resources, although considerable effort is being directed toward their development; their economy is supplemented by the earnings of Africans who work elsewhere in South Africa. The other is the advanced commercial economy pursued by the whites, with the assistance of nonwhites, elsewhere in the province. It is through this economy that most of the considerable and varied resources of the province are being developed.

Resources. The mineral wealth of Natal consists essentially of coal, which is mined in the Newcastle–Vryheid–Dundee–Utrecht area of the northwest; 85 percent of the output is high-grade bituminous coal, much of it of coking grade. This coalfield is South Africa's chief source of coking and semi-anthracite coal.

Agriculture and forestry. The most important agricultural area lies along the coast, where sugarcane is the major crop. Sugar growing is now of increasing importance in the Midlands. Refining for domestic and export requirements is mainly carried out at Durban. Plantations of pine and eucalyptus provide raw materials for sawmills and for paper, hard and soft board, and rayon pulp mills. Subtropical fruits, especially pineapples and bananas, are also grown. Durban is also the base for an important whaling industry. In the Midlands, plantations of pine, eucalyptus, poplar, and especially black wattle cover large areas. A tanning agent, of which South Africa is the largest world producer, is made from the wattle bark.

Dairying is important. Milk is produced near Durban and Pietermaritzburg; butter and cheese are produced inland. The northwest is largely devoted to cattle raising and maize cultivation, although the Drakensberg Escarpment cannot be used for agriculture because of its rough terrain. The northeast coastal plain is also unproductive, but projected irrigation schemes may provide the impetus for a final movement of the economic frontier of the province to the Mozambique border.

Industry. The products of mining, farming, forestry, and fishing, together with imported materials, support important manufacturing industries. Most are established in the towns, with the number of factories and output values roughly proportional to urban populations. Rural industries are generally restricted to those where processing takes place on the spot, although since 1960 decentralization in the form of the establishment of border industry (industry located just outside the Black States) has been in progress, notably at Elangeni (Hammarsdale), between Durban and Pietermaritzburg.

Rural and border industries

Durban, with nearby Pinetown, is the economic heart of the province and its industrial centre. It possesses most of Natal's factories, and—with more than 10 percent of those of the republic—is South Africa's second or third most important industrial region. Its factories are associated primarily with shipping, food processing, chemicals, sugar, and oil refining. An oil pipeline connects Durban with the southern Transvaal. Pietermaritzburg has a number of industries, including an aluminum plant, several footwear factories, and food-processing establishments. The northern coalfield has not yet developed industrially, but plans for the development of the Tugela Basin and a recent decision to establish a large steel plant at Newcastle may engender considerable industrial growth.

Transportation. A railway network consists of a main electrified axis from Durban, northwest through Pietermaritzburg to the Witwatersrand, a line along the coast, and various branch lines. The roads parallel the railways and also serve more remote areas; main roads are tarred, country roads have gravel or earth surfaces. Aircraft fly from Durban's airport to South African cities, and to Mozambique and Zimbabwe, and—via Jan Smuts Airport (Johannesburg/Pretoria)—to world centres; other towns have smaller airfields.

A completely new port and industrial growth point is being developed at Richard's Bay, 100 miles (160 kilometres) north of Durban; dredging began in 1972, and the port was expected to open in 1976. An aluminum smelter was already in production in the early 1970s. In the meantime, however, Durban remains Natal's only port. It serves much of the interior of southern Africa and is South Africa's main cargo port. It has shipping links with the world, and its importance was considerably increased by the closure of the Suez Canal.

Administrative and social conditions. The constitution of the republic formerly provided for provincial administration from Pietermaritzburg, through an administrator (chief executive officer), a provincial council, and its executive committee. The province, under the aegis of the central government, imposed duties and granted certain rights to local government authorities. In 1986 the council was abolished and was replaced by a presidentially appointed executive authority, headed by the administrator. Many of its powers were also transferred to Parliament. In the early 1990s the provincial government and local administrative units were in a period of transition.

The provincial administration, the administrations of the three chambers of Parliament, local authorities, voluntary bodies, and a range of welfare authorities cooperate in caring for health and welfare. Medical facilities are largely concentrated in the towns but are supplemented by many outlying clinics and health centres. Medical training for doctors is available for nonwhites at Natal University, Durban. Diseases that especially affect the Africans include tuberculosis, bilharzia (a parasitic disease of the blood), and kwashiorkor (malnutrition resulting from severe protein deficiency).

Education below the university level is provided for whites at some 360 schools, two technical colleges, and three teacher-training colleges; and for nonwhites at about 1,200 schools, nine technical colleges, and 10 teacher-training colleges. The University of Natal, which was founded in 1909 as a constituent college of the University of South Africa and assumed university status in 1949, is the chief institution of higher education. It is located partly in Pietermaritzburg and partly in Durban. Higher education for Asians is available at the University of Durban-Westville—established in 1960—and for Africans at the University of Zululand, located at Kwa Dlangezwa, near Empangeni, established in 1957.

Universities

Cultural life. The people of Natal have a rich cultural heritage drawn from their differing backgrounds. Generally speaking, that of the whites and Coloureds is of the West, the Asians of the East, and the Africans of Africa, although non-Europeans are becoming increasingly Westernized. Cultural institutions include botanic gardens, parks, museums, libraries, art galleries, cinemas, theatres, music societies, and agricultural and industrial exhibitions in many centres, notably Durban and Pietermaritzburg. Of particular importance is the recently established Natal Performing Arts Council.

HISTORY

The area was occupied centuries ago by the Nguni branch of the Bantu-speaking peoples, who entered in strength from the north. The first European settlers established the trading post called Port Natal in 1824, but little attempt was made to develop the interior, whose inhabitants had been decimated by the Zulu chief Shaka by 1820. This apparently empty land was entered by the Voortrekkers, Afrikaners who left Cape Colony, via the Drakensberg passes in 1837. After the defeat of the Zulus by the Afrikaners at Blood River in 1838, the Afrikaners established the Republic of Natal. When Natal was annexed by the British in 1843, many of the Afrikaner inhabitants left for the interior and were replaced by immigrants, mainly from Britain. From 1860 onward, increasing numbers of Indians entered to work as labourers in the sugar plantations on the coast. The colony was extended by successive acquisitions—notably that of Zululand (the area north of the Tugela River), which was added after the Zulu War of 1879, fought between the Zulu and the British. The lands north of the Buffalo River were added in 1902. During the South African War (1899 to 1902), when the British fought the Boers, the Afrikaner descendants of Dutch settlers, Natal was invaded by the Boer forces who were checked by the British defense at Ladysmith. In 1910 the colony became a province of the Union of South Africa and in 1961 of the republic. (O.W./Ed.)

European settlement

Orange Free State

The Orange Free State—Oranje Vrystaat in the Afrikaans language—is the second smallest of the four provinces of the Republic of South Africa. It has an area of 49,866 square miles (129,152 square kilometres) and a population of which three-fourths is black and less than one-fourth is white. Landlocked, the province is bordered to the north by Transvaal Province, to the east by Natal Province and by the independent state of Lesotho, and to the south and west by Cape Province. The administrative capital is Bloemfontein, which is also the judicial capital of the Republic of South Africa.

More than three-quarters of the white population is Afrikaner, and the state is a stronghold of Afrikaner culture. While its undulating plains produce up to 40 percent of South Africa's corn, a quarter of its wheat, and a quarter of its wool, the Orange Free State also contains the most productive of South Africa's seven principal goldfields, from which uranium is also obtained. As South Africa's central province, the state is the focus of its transport network.

PHYSICAL AND HUMAN GEOGRAPHY

The land. *Relief.* The Orange Free State is situated on the Highveld, a high plateau that rises to 6,000 feet (1,829 metres) on its eastern boundary, sloping down to the west to almost 4,000 feet. Its surface is formed by beds of sandstones and shales (laminated rock form of clays) and is extremely even, except where it is broken by the occurrence of dolerite (coarse basalt) intrusions (inflows of formerly molten rock) or by the occurrence of mesas. This general evenness results in low gradients for river courses; coupled with the semiarid climate, it also produces a peculiar type of wind-eroded pan (depression) in the west that is filled with water during the rainy season.

Drainage. The province is entirely drained by two rivers—the upper Orange River, which, as mentioned, forms the southern boundary of the province, and the Vaal River, which forms part of the northern boundary, together with their tributaries. The largest of the tributaries is the Caledon River, which flows into the Orange from the Drakensberg mountains in neighbouring Lesotho. The Wilge River, which drains the northeast, is the principal tributary of the Vaal. An interesting geologic feature is the Vredefort Dome, a series of incomplete circular ranges enclosing a mass of old granite partly eroded by the Vaal River.

Soils. The eastern part of the province is covered by claylike, heavy, and somewhat acidic soils. Farther to the west are alkaline soils that can be impervious to water. The northwestern region forms the so-called Sandveld, on which good crops are obtained only in years of high rainfall. The west is covered by sandy soil resembling that of the Kalahari, where successful cultivation is possible only under irrigation.

Climate. The climate varies from a warm and temperate type with an annual rainfall of 40 inches (1,016 millimetres) in the east to a semiarid type with a rainfall of 15 inches in the far west. Mean annual surface temperatures gradually increase from about 58° F (14° C) in the east to 62° F (17° C) in the west. Frost is common over the entire province from May to September, while dust storms normally occur during drought-stricken summer periods. Whirlwinds are common during warmer days; and hailstorms, sometimes destructive, occur on an average about seven times a year in the east but only three times a year in the west. Because rainfall is unreliable, long periods of drought are frequent each year. Thunderstorms, almost exclusively a summer phenomenon, are the normal sources of precipitation, occurring between 60 and 100 days a year. Sunshine is abundant; there are only about six days a year on which Bloemfontein has no sunshine. The rate of evaporation is consequently high. Because rainfall is sporadic, much reliance is placed on underground supplies of water and on storage dams.

Rainfall and temperatures

Plant and animal life. From west to east the vegetation consists of four successive zones—desert shrub, sweet grassveld, mixed grassveld, and sour grassveld. In most of the province, animal life, like the vegetation, has given way to crop cultivation and human settlement. Many animal and plant species are now protected by legislation. At Willem Pretorius Game Reserve, Highveld game animals, including herds of such species of antelope as blesbok and springbok, are plentiful; zebra and giraffe are also to be found. Golden Gate Highlands National Park was developed to restock the area with antelope species—such as black wildebeest and blesbok—and other animals and to establish a habitat for a wide variety of birds. Several of the smaller reptiles and snakes, such as adders, ringhals (venomous spitting snakes), and tree snakes, are still to be found over the entire province. The rarely observed freshwater fishes of the Orange and Vaal river systems include yellowfish, carp, and barbel.

The people. *Ethnic composition.* The ethnic pattern is revealed by the home language spoken. Among the white population, the vast majority are Afrikaans-speaking; a small percentage speak English; and the remainder speak German, Dutch, or other European languages. Historically, the Afrikaners, whose ancestors came from the preindustrial Netherlands, tended to be farmers; while the English, who came from an industrialized Britain, tended to be city dwellers. Today, most whites live in towns and cities. In general, the rural white population increases from south to north, with Bloemfontein and the northwestern goldfields being the most densely populated regions. The African population was originally largely confined to the northern and eastern districts, adjoining Lesotho, but today their numbers are rapidly increasing in the goldfield area. More than half belong to the South Sotho (also known as Sotho proper or Basuto) tribe. Various other tribal groups include the Zulu to the northeast, the city-dwelling Xhosa (Xosa), and the Tswana. African densities also increase in a northerly direction. The Coloureds, almost without exception, speak Afrikaans; slightly more than half of them live in towns and cities; the remainder increase in number toward the southwest.

frican
bal
oups

Religion. Throughout the province, church buildings are dominant in towns. The Provincial Council begins the day's work with solemn prayer, and, in the distant farmhouse, divine worship is part of the family's evening activities. The three main Afrikaans churches (Nederduits Gereformeerde, Nederduits Hervormde, and Gereformeerde) account for about three-fourths of white church membership, and the three main English churches (Anglican, Methodist, and Presbyterian) for a smaller fraction. Many white Christians engage in missionary activity. The Coloureds are almost entirely Christian. The Africans of the Orange Free State are more Christianized than those of the neighbouring Transvaal.

The white population is dispersed over the country on individually owned farms and single homesteads. New discoveries of gold since World War II in the northwest have resulted in a maize and cattle country being turned into a mining and industrial belt. New towns, including Welkom, Virginia, and Odendaalsrus, have sprung up. Welkom, the second largest town, has grown rapidly since 1950 and provides an example of modern planning, in which residential suburbs radiate from the town centre with green wedges of open spaces between them. Bloemfontein, the largest city, was the province's administrative, transport, educational, and cultural centre. It is also developing industrially. Kroonstad is the third largest town, while Sasolburg has a burgeoning chemical industry. Most of the remaining towns are small.

The economy. *Resources.* The Orange Free State is not as rich in mineral resources as the neighbouring Transvaal. The gold resources of the Welkom district are the most important. Diamond production totalled 220,000 metric carats in 1968 alone. Estimated coal reserves of 3,400,000,000 tons are worked in the north and are mainly used in the coal-to-oil project at Sasolburg that supplies about 10 percent of the nation's gasoline requirements. Salt is derived from the pans of the semiarid region, and bentonite (a moisture-absorbing clay) is mined from a large deposit near Parys on the Vaal River.

Agriculture. The province forms part of the country's grain belt. The undulating plains afford excellent grazing, and stock farming is important. The province produces almost one-quarter of South Africa's wool output, and its cattle are mainly located in the wetter northeastern districts of Harrismith, Vrede, and Bethlehem.

Water is the province's most essential resource. Apart from several storage dams—like the large Allenmanskraal, the Erfenis, and the Kalkfontein—the region will benefit enormously from the vast Hendrik Verwoerd and Van der Kloof dams, part of the great Orange River Project.

Transportation. The state-owned railways and highways provide internal provincial transportation. The J.B.M. Hertzog Airport at Bloemfontein is the largest in the province and is of growing importance. Post-office and telecommunication services expand from year to year to meet an increasing demand. The South African Broadcasting Corporation broadcasts to the province in English, Afrikaans, and several Bantu languages. There is a radio station at Bloemfontein.

Administrative and social conditions. Until the late 1980s the Orange Free State was administered by a provincial council, in which members were elected every five years, and an administrator appointed by the state president. The council had jurisdiction on such matters as taxation to provide revenue for white education, hospitals, roads, and nature conservation. The council, however, was replaced by a presidentially appointed executive authority, and many of its powers were transferred to Parliament. In the early 1990s provincial government, as well as the local governments, was in a period of transition. Forty-nine magisterial districts control the administration of justice.

There are several provincial hospitals. Apart from the leper, mental, and tuberculosis institutions, there are provincial and private hospitals and clinics. The province and the administrations of the three chambers of Parliament provide a wide range of health services. Increased attention to health has led to a decline in infant mortality over the last decades.

Education

Schooling, when available, is free for all and is compulsory for whites between seven and 16 years of age. Where demand justifies the establishment of facilities, Coloureds between seven and 16 years are also compelled to attend schools. The University of the Orange Free State, located in Bloemfontein, is the institution of higher learning for whites. Administration of primary and secondary education and the training of teachers, as well as the administration of the university and higher technical training, are under the central government's Department of National Education and the separate Departments of Education and Culture of the three chambers of Parliament.

The Parliamentary administrations and the provincial authorities, in cooperation with private organizations, undertake a wide variety of welfare services, including care for the aged, workshops and housing for the blind, and the provision of housing for the Africans in larger urban centres.

Cultural life. The cultural life of the whites resembles that in Western countries. As the Orange Free State is predominantly a rural province, the more austere and moralistic aspects of the Afrikaner character are probably more in evidence here than elsewhere. The well-known hospitality, a general trait of white South African life, is a result of the stimulus of a mutual dependence that developed in an earlier era when life on the trekker's frontier was hard and sometimes dangerous. African culture is still strongly influenced by tribal life; the supremacy of the chiefs is recognized, and the heritage of traditional animist religion is still in evidence, despite the influence of Christianity. The sunny climate influences all aspects of life, from clothing to food and drink and from literature to architecture. Cultural institutions such as museums and theatres are to be found in the principal towns.

HISTORY

Before the arrival of Europeans, the region was the home of seminomadic Bantu-speaking tribes. In the 18th century, Europeans first crossed the Orange River, which forms part of the state's southern border, to enter the area. In the early 19th century, Boers began to settle the area. After 1836 came the Great Trek, a migratory movement in which Boer farmers seeking freedom from British rule also moved north across the Orange River. From 1848 to 1854 the British administered the territory as the Orange River Sovereignty, after which the British withdrew, and the Boer settlers formed the independent Orange Free State. The constitution of the new state combined traditional Boer institutions with Dutch and United States constitutional theory. During the first few years of the state's existence, it was much harassed by raids from Basuto (Sotho) tribesmen from the east; the Basutos were, however, at length conquered, and part of their territory was annexed. During the South African War (1899–1902), the Orange Free State fought against the United Kingdom by the side of its sister state the South African Republic (now the Transvaal), with which it had a defensive alliance. The two Boer republics won some victories against the British army but could not finally prevail. In 1900 the Orange Free State was annexed by the United Kingdom as the Orange River Colony. Self-government was restored in 1907, after which, in 1910, the colony became the Orange Free State Province of the Union of South Africa. In 1961, when the Union of South Africa became the Republic of South Africa, the province remained unchanged in form and administration. (J.N.S./Ed.)

Transvaal

The Transvaal is the northernmost and second largest province of the Republic of South Africa. With an area of 109,622 square miles (283,917 square kilometres), it is bordered by Botswana and Zimbabwe to the north, Mozambique and Swaziland to the east, the provinces of Natal and the Orange Free State to the south, and Cape Province to the west. About three-fourths of the population is black and less than one-fourth is white. Many of the whites are descendants of the Afrikaners who in the earlier 19th century migrated north over the Vaal River, which now forms part of the province's southern boundary, thus giving the province its present name—Transvaal, "across the Vaal."

Rich in mineral resources, including gold and uranium, as well as in power resources, the Transvaal is South Africa's industrial heartland and also an important agricultural region in its own right. Pretoria, the provincial capital, is also the administrative centre of the republic. The province also contains Johannesburg, the largest city in South Africa.

PHYSICAL AND HUMAN GEOGRAPHY

The six regions

The land. *Relief.* There are six distinct physiographic regions in the province—the Highveld, the Bushveld Basin, the Waterberg Plateau, the Pietersburg Plain, the Limpopo Valley, and the Lowveld. In the south, the altitude of the Highveld ranges between 4,000 and 6,000 feet (1,219 and 1,829 metres), declining slightly from east to west. The eastern area is covered by horizontal sandstones and shales (laminated clays) that produce gray, claylike acidic prairie soils with a leached upper layer that merges into lateritic (leached, reddish, iron-bearing) soils farther to the north and west. The relief is undulating but changes into a rugged landscape in the vicinity of the Witwatersrand Ridge, where older rocks have been bared. To the west of Johannesburg, the Highveld is monotonously even, being composed of water-bearing limestone and lavas. The region is drained by the Vaal River and its tributaries.

The central part of the province is occupied by the Bushveld Basin, formed by a large substratum of rock that has been solidified from the molten state. The peripheral ridge-and-valley landscape is composed of quartzites and intervening shales that dip toward the centre. These ridges are known as the Western Bankeveld around Pretoria and the Eastern Bankeveld near the Drakensberg Mountains. North of Pilgrimsrest, water erosion on the Eastern Bankeveld has created the spectacular canyon of the Blyde River. The black clays of the basin are associated with norite (a granular, mineral-bearing rock solidified from the molten state) and basalt and are of exceptionally high fertility, as they are at the Springbok Flats.

The Waterberg Plateau rises from 3,500 to nearly 7,000 feet in altitude to the northwest of the basin. Few of its grass- and bush-covered sandstone areas are suitable for cultivation. To the northeast the Pietersburg Plain is almost entirely composed of a granite surface. It is bounded by the Strydpoort Mountains in the south and the Soutpansberg Mountains in the north.

Soils. The Limpopo Valley forms the northern border with Zimbabwe. It cuts mainly into old granite in a bushy area covered by the sandy soil of the nearby Kalahari, and a red-brown unleached subtropical sandy soil to the east. Farther east, this sandy soil extends into the Lowveld, a bush-clad plain, generally below 2,000 feet in altitude, that gently slopes toward the Lebombo Mountains on the Mozambique boundary. On the slopes of the Great Escarpment—the ridge that separates the plateau from the Lowveld—thick, red, iron-bearing earths support dense plantation growth. The whole of the northern, and most of the eastern, Transvaal is drained by the Limpopo River and its tributaries, which include the Crocodile, Mogalakwena, Letaba, and Olifants rivers.

Climate. Temperatures are directly related to the altitudes of the region. Rainfall is unreliable and varies from year to year. Almost exclusively a summer phenomenon, it normally arrives as thunderstorms and showers. Hailstorms that frequently are destructive may be expected about five times a year, while the great amount of sunshine results in extensive evaporation.

The Highveld is warm and temperate, with mean annual surface temperatures of 57° F (14° C) in the east and 65° F (18° C) in the southwest. Rainfall varies from 40 inches (1,016 millimetres) in the east to 18 inches in the west. Frost is common from mid-May to mid-September, but its duration decreases to the north and east. North of Pretoria, the climate becomes subtropical and semiarid, with an annual average rainfall ranging from 28 inches on the Waterberg Plateau to 16 inches in the Sand River Valley, which is the driest part of the Transvaal. The subtropical Lowveld is warm and oppressive except during the winter. The mean annual surface temperature is 74° F (23° C), and rainfall increases from 19 inches in the northeast to more than 72 inches on the Drakensberg slopes, where mist and drizzle are frequent.

Plant and animal life. The natural vegetation of the Highveld and part of the Bankeveld consists of successive sweet, mixed, and sour grassland toward the east, the extent of each zone depending upon the climate and type of soil. The whole of the Lowveld and the greater part of the country to the north of the Bankeveld carries a parkland type of vegetation, with acacia types predominating and tall grasses and scrubs forming the undergrowth. Where rainfall is high in the mist belt, true timber forests with dense undergrowth are found, but they have suffered severely from the hand of man.

Kruger National Park

Animal life has been severely decimated, except in conservation areas. Kruger National Park—covering some 8,000 square miles in the eastern Lowveld—supports a wide variety of primates such as monkeys and baboons; insect-eaters such as hedgehogs and ant bears; carnivores such as lions, cheetahs, jackals, and hyenas; and hoofed animals such as rhinoceroses, hippopotamuses, giraffes, buffalo, kudus, blue wildebeests, and impalas. There are many species of birds. Crocodiles and several species of snakes are found, as well as freshwater fish.

Settlement patterns. Because of large, individually owned farms, rural settlement among whites is dispersed, with long distances separating neighbours. The pattern in the African areas differs totally because African communal ownership and communal land use result in the appearance of rural villages separated by open stretches of cultivated and grazing lands.

South Africa's urban heartland

Towns are generally small and few in number, having evolved as mining, marketing, or administrative centres. Although the towns were founded by whites, they are numerically dominated by the Africans. The highly built-up Witwatersrand region covers about 200 square miles. Johannesburg, Pretoria, and the Vereeniging–Vanderbijlpark complex—in fact, the whole of the southern Transvaal—forms the mining, industrial, commercial, and financial heart of South Africa. In the west the gold-mining complex of Klerksdorp–Stilfontein–Orkney, and in the east the Witbank–Middelburg coalfields and industrial area, are important. Other important centres are Pietersburg, Nelspruit, Ermelo, and Lichtenburg.

The people. *Ethnic composition.* There are four main population groups in the Transvaal, the Africans, whites, Coloureds, and Asians.

More than half of the white population is Afrikaans-speaking and one-third is English-speaking. The balance are of German, Dutch, and other European origin. Most of the whites are urban-centred, being concentrated in the Pretoria–Witwatersrand–Vereeniging metropolitan complex. The Afrikaners, once overwhelmingly engaged in farming, live increasingly in the cities; now the vast majority is urbanized.

The African population, in spite of a broad cultural solidarity, is diversely grouped in numerous tribes and clans. The main groups are the Sotho, Nguni, Tsonga, and Venda. Since few members of the main groups can understand each other's language, a lingua franca called Fanagalo has developed, especially among the African gold miners. The majority of the Africans live in rural areas.

The Coloureds are urbanized and are concentrated mainly in the Witwatersrand area and in the larger towns. They speak both Afrikaans and English. The Asians are also predominantly urban; they consist mainly of Indians, although there is also a Chinese minority.

Religion. More than half of the white population are members of the three Afrikaans churches—the Nederduits Gereformeerde and the Nederduits Hervormde (both Dutch Reformed churches), and the Gereformeerde (Reformed Church of South Africa)—and about one-fourth are members of the three main English churches—Anglican, Methodist, and Presbyterian. There is also a small Jewish community. Because of strong missionary activity, the vast majority of the Coloureds are Christian. Many of the Africans have accepted Christianity, but they tend to form independent sects; a significant fraction of the Africans belong to more than 1,000 separatist churches. The Asian community has retained its affiliation with Hinduism and Islām.

ld,
anium,
d
monds

The economy. *Resources.* The Transvaal contains one of the greatest known concentrations of mineral deposits, both in variety and quantity. Producing large quantities of gold and uranium, the mines of the Witwatersrand, Klerksdorp, and eastern Transvaal have stimulated the growth of a large complex of industries and economic activities. Gold purchased for industrial purposes had by 1968 reached the level of total world output of newly mined gold. Diamond output amounted to 2,550,000 metric carats in 1968. Enormous reserves of platinum, as well as of chromite, tin, and nickel, are associated with the Bushveld. Reserves of coal are estimated at 75,000,000,000 tons. Other mineral deposits include iron ore at Thabazimbi, copper at Messina and Phalaborwa, asbestos, vermiculite (a micaceous mineral), antimony, corundum (crystals used industrially as abrasives), and limestone. An ever-growing need for explosives for mining has resulted in the development of a large chemical industry.

Agriculture. Agriculture is productive, having adopted modern methods of conservation, irrigation, and fertilization. Products include corn (maize), wheat, peanuts (groundnuts), sunflower seeds, cotton, sugar, tobacco, and potatoes and other vegetables, as well as a wide variety of fruits. Beef and dairy cattle are important, and there is some sheep ranching. Huge plantations of eucalyptus and coniferous trees in the mist belt comprise a large fraction of South Africa's total acreage of plantations.

The demand for water for mining, industry, power stations, and irrigation has exceeded the supply available from the Vaal River, virtually all of which is already in use. Planners are seeking supplementary supplies from the Tugela River of Natal and from the desalinization of seawater.

Transportation. Railways, as well as a vast highway network, cover the province. An underground pipeline from the port of Durban in Natal Province conveys petroleum products to the Witwatersrand over a distance of 450 miles. South Africa's main international airport, Jan Smuts, is situated near Johannesburg. Post-office services and the telecommunication system have expanded steadily to cope with increasing demands. There are national radio broadcasting stations at Johannesburg, Pretoria, and Pietersburg. The Voice of South Africa transmits to other countries from Bloemendal in 10 languages. Radio Highveld broadcasts continuously in Afrikaans and English.

Administrative and social conditions. *Government.* Until the late 1980s provincial authority was delegated by the Parliament of South Africa to a provincial council, in which members were elected every five years, and an administrator appointed by the state president. Taxation for provincial purposes was under the council's control. The council, however, was abolished in 1986 and replaced by a new executive authority appointed by the state president and headed by the administrator, and many of its powers were transferred to the three chambers of Parliament. In the early 1990s the provincial government and the local governments were in a period of transition. There are 62 magisterial districts with responsibility for the administration of justice.

Health and welfare. There are provincial and private hospitals. There are also mine hospitals and hospitals for leprosy, mental illness, and tuberculosis. The government maintains field services to combat malaria, plague, typhus, tuberculosis, trachoma, and other diseases. Indigent whites and most nonwhites receive free medical attention from the provincial hospitals. A major hospital specializing in the treatment of a wide range of diseases is Baragwanath, a hospital for nonwhites near Johannesburg.

Education

Elementary and secondary education and the training of teachers are administered by the central government. Schooling is compulsory for whites between the ages of seven and 16 and, where the demand justifies the establishment of facilities, for Coloureds between seven and 16 and Asians between seven and 15. Six universities—the universities of South Africa, Pretoria, the Witwatersrand, Potchefstroom, Rand Afrikaans University, and the University of the North for the Africans—are subsidized by the federal government but remain autonomous.

Welfare activities are undertaken by voluntary organizations, the departments of Health Services and Welfare of the three chambers of Parliament, and the provincial administration. From 1960 onward, in order to clear the so-called squatters' camps that resulted from a rapid industrialization that was accompanied by an unprecedented influx of Africans into urban areas in the southern Transvaal, more than 100,000 houses were built; provision was also made for schools, parks, community halls, and other facilities.

Cultural life. The Transvaal's population diversity has resulted in a variety of different cultures. "White" Transvaal is culturally comparable to any Western country. Its well-known hospitality can be traced back to the rough environment and hard life of past decades, which stimulated mutual dependence and a close family life. There is a strong desire to preserve objects of historical significance; Afrikaner attachment to national tradition is symbolized by the Voortrekker Monument near Pretoria, which honours the first pioneers.

The Transvaal has attained international recognition for its contributions to art, music, and ballet. Opera is presented regularly, and the Aula at Pretoria and Johannesburg Civic Theatre are well-known cultural centres. A wealth of literature and folk song has accumulated, but there is little if any traditional architecture.

Among the Africans, tribes are normally composed of a number of clans whose members are related. A strong recognition of the power of the tribe's chief prevails, and traditional medicine has considerable influence. The Africans are renowned for their music and dancing, of which the "Domba," or snake dance, of the Venda is an example.

The climate lends itself to outdoor living, and sports are extremely popular. Motoring vacations over large distances are common, and an exodus to the beaches of Natal is a holiday phenomenon at all seasons.

HISTORY

In the early 19th century the land between the Vaal River in the south and the Limpopo River, which today forms the northern border, was inhabited by tribes of agricultural Bantu-speaking peoples who were skilled in metalwork-

Settlement by the Voortrekkers

ing. They became unsettled in the 1820s and 1830s by invasions of refugee Bantu tribes fleeing south and west from the warring Zulu. Between 1837 and 1838 seminomadic pastoral Afrikaner farmers, the Voortrekkers, moving northward to avoid British rule, crossed the Vaal River and entered the area, where they settled in isolated farms. More white settlers arrived when the United Kingdom annexed both Natal and the Orange River Sovereignty in the 1840s. In 1852 the independence of the Afrikaners in the Transvaal was recognized by the United Kingdom with the conclusion of the Sand River Convention. In 1857 the new Transvaal state was formally proclaimed as the South African Republic; its authority, however, was virtually limited to the southwest of the present Transvaal. In 1877 the republic was annexed by the United Kingdom, but the Afrikaners resorted to arms to maintain their independence, which they regained—subject to certain provisos—in 1881 after the British forces had been overwhelmed at the Battle of Majuba Hill. The discovery of gold in the Witwatersrand area in 1886 resulted in a tremendous influx of fortune seekers, primarily English and German, which caused new problems for the Afrikaner republic. Tension with the United Kingdom increased when an English adventurer, Leander Starr Jameson, led an abortive raid across the frontier of the South African Republic in an attempt to provoke an internal uprising. War between the republic and the United Kingdom subsequently broke out in 1899. Defeated by superior arms and numbers, the republic, together with its ally the Orange Free State, lost its independence when peace was concluded in 1902, after which the Transvaal became a British crown colony. In 1907 it regained self-government and in 1910 became a province of the Union of South Africa, a status that was maintained when the Union became the Republic of South Africa in 1961. (J.N.S./Ed.)

BIBLIOGRAPHY

Physical and human geography. General (including the physical landscape): MONICA COLE, *South Africa,* 2nd ed. (1966); RICHARD ELPHICK and HERMANN GILIOMEE (eds.), *The Shaping of South African Society, 1652–1820* (1979); N.C. POLLOCK, *Africa* (1968); JOHN H. WELLINGTON, *Southern Africa,* vol. 1 (1955).

People and population: LEO MARQUARD, *The Peoples and Policies of South Africa,* 4th ed. (1969); A. NEL, *Geographical Aspects of Apartheid in South Africa* (1962); LEONARD THOMPSON (ed.), *African Societies in Southern Africa* (1969); SOUTH AFRICA, DEPARTMENT OF STATISTICS, *South African Statistics* (biennial); *The Standard Bank Annual Economic Survey.*

Economic and social conditions: HERIBERT ADAM, *Modernizing Racial Domination: South Africa's Political Dynamics* (1971); D. HOBART HOUGHTON, *The South African Economy* (1964; 4th ed., 1976); FRANCIS WILSON, *Migrant Labour: Report to the South African Council of Churches* (1972); Study Commission on U.S. Policy Toward Southern Africa, *South Africa: Time Running Out* (1981); SOUTH AFRICA, DEPARTMENT OF FOREIGN AFFAIRS AND INFORMATION, *Official Yearbook* (annual).

Race relations: MURIEL HORRELL (ed.), *Survey of Race Relations in South Africa* (annual), published by the South African Institute of Race Relations, is the most comprehensive record of affairs in the country. The Institute also publishes a series of more specific studies. Books on this theme include: JOHN DE ST. JORRE, *A House Divided: South Africa's Uncertain Future* (1977); BERNARD MAGUBANE, *The Political Economy of Race and Class in South Africa* (1978); and ROBERT M. PRICE and CARL G. ROSBERG (eds.), *The Apartheid Regime: Political Power and Racial Domination* (1980). The United Nations Centre Against Apartheid publishes the "Notes and Documents Series," which includes a number of studies critical of the government policy of separate development.

History. J.D. FAGE and R.A. OLIVER, *A Short History of Africa* (1963), helpful for placing Southern Africa in the context of African history; *The Cambridge History of the British Empire,* vol. 8, *South Africa, Rhodesia and the High Commission Territories,* 2nd ed. (1963); P.L. VAN DEN BERGHE, *South Africa: A Study in Confict* (1965), contains some useful insights, written by a sociologist; E.A. WALKER, *A History of Southern Africa,* 3rd ed. rev. (1962), extremely useful for basic factual information; MONICA WILSON and L.M. THOMPSON (eds.), *The Oxford History of South Africa,* 2 vol. (1969–71), the only general reference work to make a serious attempt to record the history of all the peoples of South Africa; C.R. BOXER, *The Dutch Seaborne Empire, 1600–1800* (1965); M.W. SPILHAUS, *South Africa in the Making, 1652–1806* (1966) and *The First South Africans and the Laws Which Governed Them* (1950); S.D. NEUMARK, *Economic Influences on the South African Frontier* (1957). J.S. GALBRAITH, *Reluctant Empire: British Policy on the South African Frontier, 1834–1854* (1963); W.M. MACMILLIAN, *Bantu, Boer, and Briton,* 2nd rev. ed. (1963), and *The Cape Colour Question: A Historical Survey* (1927); I.E. EDWARDS *Towards Emancipation: A Study in South African Slavery* (1942); E.A. WALKER, *The Great Trek* (1934), a readable and generally reliable account; J.A.I. AGAR-HAMILTON, *The Native Policy of the Voortrekkers: An Essay on the History of the Interior of South Africa, 1836–1858* (1928), and *The Road to the North: South Africa, 1852–1886* (1937); C.W. DE KIEWIET, *British Colonial Policy and the South African Republics, 1848–1872* (1929); EDGAR H. BROOKES AND COLIN DE B. WEBB, *A History of Natal* (1965); SHULA MARKS, *Reluctant Rebellion: The 1906–08 Disturbances in Natal* (1970). C.W. DE KIEWIET, *The Imperial Factor in South Africa* (1937, reprinted 1965); C.F. GOODFELLOW, *Great Britain and South African Confederation, 1870–1881* (1966); F.A. VAN JAARSVELD, *Die ontwaking van die Afrikaanse nasionale bewussyn, 1868–1881,* 2nd ed. (1959; Eng. trans., *The Awakening of Afrikaner Nationalism 1868–1881,* 1961). JOHN W. CELL, *The Highest Stage of White Supremacy: The Origins of Segregation in South Africa and the American South* (1982), a comparative study emphasizing the role of the economy in encouraging segregation; BARBARA VILLET, *Blood River* (1982), a sympathetic history of South Africa's white Afrikaners. D.W. KRUGER, *Paul Kruger,* 2 vol. (1961–63); D.M. SCHREUDER, *Gladstone and Kruger: Liberal Government and Colonial "Home Rule," 1880–85* (1969); J.S. MARAIS, *The Fall of Kruger's Republic* (1961); CECIL HEADLAM (ed.), *The Milner Papers,* 2 vol. (1931–33); J.A. HOBSON, *The War in South Africa: Its Causes and Effects* (1900); L.S. AMERY (ed.), *The Times History of the War in South Africa, 1899–1902,* 7 vol. (1900–09); G.H.L. LE MAY, *British Supremacy in South Africa, 1899–1907* (1965). GWENDOLYN M. CARTER, *The Politics of Inequality: South Africa Since 1948,* 2nd rev. ed. (1959); L.M. THOMPSON, *Politics in the Republic of South Africa* (1966); D.W. KRUGER, *The Making of a Nation: A History of the Union of South Africa, 1910–1961* (1969), a short, general political account from the white side; D.J.N. DENOON, *The Grand Illusion: The Failure of Milner's Reconstruction Policy in the Transvaal, 1900–1905* (1973), a major scholarly reassessment; NICHOLAS MANSERGH, *South Africa 1906–1961: The Price of Magnanimity* (1962); L.M. THOMPSON, *The Unification of South Africa 1902–1910* (1960); W.H. VATCHER, *White Laager: The Rise of Afrikaner Nationalism* (1965); F.A. JOHNSTONE, "White Prosperity and White Supremacy in South Africa Today," *African Affairs,* 69:124–140 (1970), an important article, challenging the assumption that increased prosperity will end the colour bar; PETER CALVOCORESSI, *South Africa and World Opinion* (1961); DENNIS AUSTIN, *Britain and South Africa* (1966); EDWARD ROUX, *Time Longer Than Rope: A History of the Black Man's Struggle for Freedom in South Africa,* 2nd ed. (1964); MARY BENSON, *The African Patriots: The Story of the African National Congress of South Africa* (1963) and *South Africa: The Struggle for a Brithright* (1966); EDWARD FEIT, *South Africa: The Dynamics of the African National Congress* (1962) and *African Opposition in South Africa: The Failure of Passive Resistance* (1967); PETER WALSHE, *The Rise of African Nationalism in South Africa: The African National Congress, 1912–1952* (1970).

Cape of Good Hope. SOUTH AFRICA, OFFICE OF CENSUS AND STATISTICS, *Official Year Book of the Union of South Africa* (annual 1918–60), detailed accounts of the history, physical geography, geology, climate, and vegetation appear in the earlier issues and of all phases of administrative and economic development in all issues; SOUTH AFRICA, DEPARTMENT OF INFORMATION, *Official Yearbook of the Republic of South Africa* (annual 1974–), a replacement for the earlier, discontinued series; SOUTH AFRICA, BUREAU OF STATISTICS, *Statistical Year Book* (annual 1964–66); and SOUTH AFRICA, DEPARTMENT OF STATISTICS (before 1970 BUREAU OF STATISTICS), *South African Statistics* (biennial 1968–), useful sources of statistical data; *Standard Encyclopaedia of Southern Africa,* 12 vol. (1970–76), a comprehensive encyclopaedia; A.M. and W.J. TALBOT, *Atlas of the Union of South Africa* (1960), a graphic survey of the climate, vegetation, resources, economic development, and external trade of the Union (1910–50), presented in more than 600 maps and charts; UNIVERSITY OF STELLENBOSCH, INSTITUTE FOR CARTOGRAPHIC ANALYSIS, *Economic Atlas of South Africa* (1981), 132 statistical maps.

Natal. Authoritative studies include: *Natal Regional Survey,* vol. 1–15 (1951–); *Reports* of the Department of Economics, University of Natal (1951–); and *Reports* of the Natal Town and Regional Planning Commission, Pietermaritzburg (1952–). See also E.H. BROOKES and C.B. WEBB, *A History of Natal* (1965); E.J. KRIGE, *The Social System of the Zulus,* 2nd ed. (1951); F. MEER, *Portrait of Indian South*

Africans (1969); and various issues of the *South African Geographical Journal.*

Orange Free State. W. ALBERTYN (ed.), *Official South African Municipal Yearbook* (annual); BUREAU OF STATISTICS, REPUBLIC OF SOUTH AFRICA, *Population Census 1960;* M.M. COLE, *South Africa,* 2nd ed. (1966), physical and human geography; DEPARTMENT OF PLANNING, REPUBLIC OF SOUTH AFRICA, *Development Atlas* (1966), an extensive work containing maps and detailed descriptive information; A.K. HAAGNER, *South African Mammals* (1920); and with R.H. IVY, *Sketches of South African Bird Life,* 2nd ed. (1914), two standard works, illustrated; A.C. HARRISON *et al., Fresh-Water Fish and Fishing in South Africa,* ch. 4 (1963), on species in the Vaal and Orange river systems; D.H. HOUGHTON, *The South African Economy,* 2nd ed. (1967), a general survey; L.C. KING, *South African Scenery,* 3rd ed. rev. (1963), geomorphology and topography; E. PALMER and N. PITMAN, *Trees of South Africa* (1961); N.C. POLLOCK and S. AGNEW, *An Historical Geography of South Africa* (1963); *State of South Africa Yearbook,* a general descriptive and statistical annual; J. VISSER, *Poisonous Snakes of Southern Africa* (1966), a descriptive classification with colour photographs.

Transvaal. W. ALBERTYN (ed.), *Official South African Municipal Yearbook* (annual); BUREAU OF STATISTICS, REPUBLIC OF SOUTH AFRICA, *Population Census 1960;* T. CAMPBELL, R. FINDLAY, and J. VAN DER MERWE, *Birds of the Kruger and Other National Parks,* 4 vol. (1957–65), illustrated; M.M. COLE, *South Africa,* 2nd ed. (1966), physical and human geographical coverage; DEPARTMENT OF PLANNING, REPUBLIC OF SOUTH AFRICA, *Development Atlas* (1966), contains maps and descriptions of the physical background, social aspects, water resources, minerals and mines, agriculture, communications, and economic aspects; A.K. HAAGNER, *South African Mammals* (1920), still a descriptive classic, illustrated; A.C. HARRISON *et al., Fresh-Water Fish and Fishing in South Africa,* ch. 3 (1963), deals with species in the Vaal and Limpopo river systems; D.H. HOUGHTON, *The South African Economy,* 2nd ed. (1967), a general survey; L.C. KING, *South African Scenery,* 3rd ed. rev. (1963), geomorphology, including chapters on topography; E. PALMER and N. PITMAN, *Trees of South Africa* (1961), a description of 51 families and their distribution, illustrated; J.N. SCHEEPERS, *A Cartographic Analysis of the Man-Land Ratio: An Adventure into the Population Geography of the Transvaal* (1967), contains analysis of the man-land ratio and chapters on distributional patterns, illustrated; J.H. WELLINGTON, *Southern Africa,* vol. 1 (1955), physical geography.

South America

South America is situated between longitudes 35° and 80° W and latitudes 12° N and 55° S. It is the fourth-largest of the continents, having an approximate area of 6,874,200 square miles (17,814,400 square kilometres), or about one-eighth of the land surface of the Earth.

Politically, the continent and its adjacent islands are divided into 12 sovereign republics and two dependencies. The republics are Argentina, Bolivia, Brazil, Chile, Colombia, Ecuador, Guyana, Paraguay, Peru, Suriname, Uruguay, and Venezuela; the dependencies are the Falkland Islands (owned by Britain but also claimed by Argentina as the Islas Malvinas) and French Guiana (Guyane Française).

South America is compact and roughly triangular in shape, being broad in the north and tapering to a point in the south at Tierra del Fuego. It is bounded by the Caribbean Sea to the northwest, the Atlantic Ocean to the northeast, east, and southeast, by Drake Passage to the south, and by the Pacific Ocean to the west. In the northwest it is joined to North America by the Isthmus of Panama, which forms a land bridge narrowing to about 50 miles (80 kilometres) at one point; to the north, the islands of the West Indies roughly form a chain running from Trinidad to Florida. At the continent's southernmost extent are the small Islas Diego Ramírez, about 60 miles (100 kilometres) southwest of Cape Horn.

In relation to the area of the continent, the coastline, which is 15,803 miles in length, is exceptionally short—one mile of coast for every 435 square miles of area.

An important feature of the continent is its dissymmetry. Because the Andes border the Pacific, the gigantic backbone that they create divides the continent into two parts differing greatly in both size and character. This unequal division has had profound effects on climate, wildlife, and even human settlement.

No other continent has a greater latitudinal extent or penetrates so far to the south. Though its northern part is crossed by the Equator and 80 percent of its landmass is located within the tropical zone, South America also extends into the subantarctic. Because of the high altitudes and wide extent of the Andes within the tropics, extensive zones of temperate or cold climate exist in the vicinity of the Equator—a circumstance that is unique. These conditions produce a great range of climates; the unrivalled diversity of ecological zones is probably the most prominent characteristic of South American geography.

Some parts of the continent are now industrialized, but most regions still follow an agricultural way of life. The possibilities of development, offered by the wealth of mineral products and renewable resources, are considerable. The human devastation of nature and of natural resources has nevertheless reached alarming proportions. Ecologists find this development particularly disconcerting because the human population is increasing at a rapid rate throughout the whole continent. Despite the wealth of the continent's resources, the trend could prove to be an obstacle to optimum development of most South American countries.

This article treats the physical and human geography and the history of South America, followed by discussion of geographical features of special interest. For discussion of individual countries of the continent, see specific articles by name, *e.g.,* ARGENTINA, BRAZIL, and VENEZUELA. Other South American countries are treated in the article CENTRAL AMERICA. For discussion of major cities of the continent, see the articles BUENOS AIRES, CARACAS, LIMA, RIO DE JANEIRO, and SÃO PAULO. For discussion of the indigenous peoples of the continent, see the articles AMERICAN INDIANS and PRE-COLUMBIAN CIVILIZATIONS. Related topics are discussed in the articles LATIN AMERICA, THE HISTORY OF; and LATIN-AMERICAN LITERATURE. For further references, see also the entries for these topics in the *Index.*

The article is divided into the following sections:

PHYSICAL AND HUMAN GEOGRAPHY

Geological history

The geological history of South America is complex and to some extent still unknown. The present geological structure is the result of the movement that uplifted the Andean mountain ranges.

BASIC STRUCTURAL UNITS

The five structural units

The geological structure of South America is dominated by five geotectonic units (structural rock masses of the Earth's crust), the history of which can be traced from the end of the Precambrian Era (570,000,000 years ago). The five units are the continental shields (stable masses of Precambrian rock), the epicontinental basins (*i.e.,* those situated on the continental plateau), the insular shields, the belt of pericratonic basins (*i.e.,* the basins surrounding large relatively immobile parts of the Earth's crust), and the Andean Geosyncline (*i.e.,* the gigantic linear trough in which sediments accumulated and which runs from north to south on the western side of the continent).

The continental shields consist of three old landmasses that together form the vast Guiana-Brazilian Shield that constitutes the greater part of northeastern South America. The complex, as a whole, is of crystalline rock, and—unlike the Andes—has long been quiescent. Its three parts are, firstly, the Guiana Shield, situated between the Orinoco and Amazon rivers; secondly, the Central Brazilian Shield, between the Amazon, the Parnaíba–São Francisco Basin, and the Paraná Basin, together with a long spur extending to central Paraguay; and, thirdly, the Eastern Brazilian Shield, covering eastern Brazil and Uruguay.

The epicontinental basins consist of three huge basins—the Amazon, Parnaíba–São Francisco, and Paraná basins—situated in weaker areas between large and relatively immobile masses. They have great unity of structure and have undergone uniform sedimentation, having been at first invaded by seas of the Paleozoic Era (from 570,000,000 to 225,000,000 years ago), then filled with continental sediments and volcanic deposits. Only tectonic movements resulting from volcanism have changed their structure.

The Amazon Basin is about 540,000 square miles in area and consists mostly of marine sediments from 280,000,000 to 570,000,000 years old, which underlie continental deposits. The Parnaíba–São Francisco Basin, in northeastern Brazil, is filled with sediments almost 10,000 feet deep; the oldest of these are marine limestone deposits, covered by sandstone, conglomerate deposits from rivers, and continental sediments which accumulated during the Mesozoic Era (from 225,000,000 to 65,000,000 years ago), together with some marine deposits in the north. The Paraná Basin, more than 500,000 square miles in area, is filled with sediments about 6,500 feet thick and with up to 5,000 feet of basaltic lava. Continental sediments of sand and silt, from 190,000,000 to 295,000,000 years old, are covered with lava deposits which extend over an area of more than 460,000 square miles.

The insular shields, which lie to the east of the Andes, are Precambrian massifs (mountainous masses) situated in Argentina; they include the Pampine Sierra and the Patagonian and Deseado massifs and may extend along the continental shelf as far as the Falkland Islands.

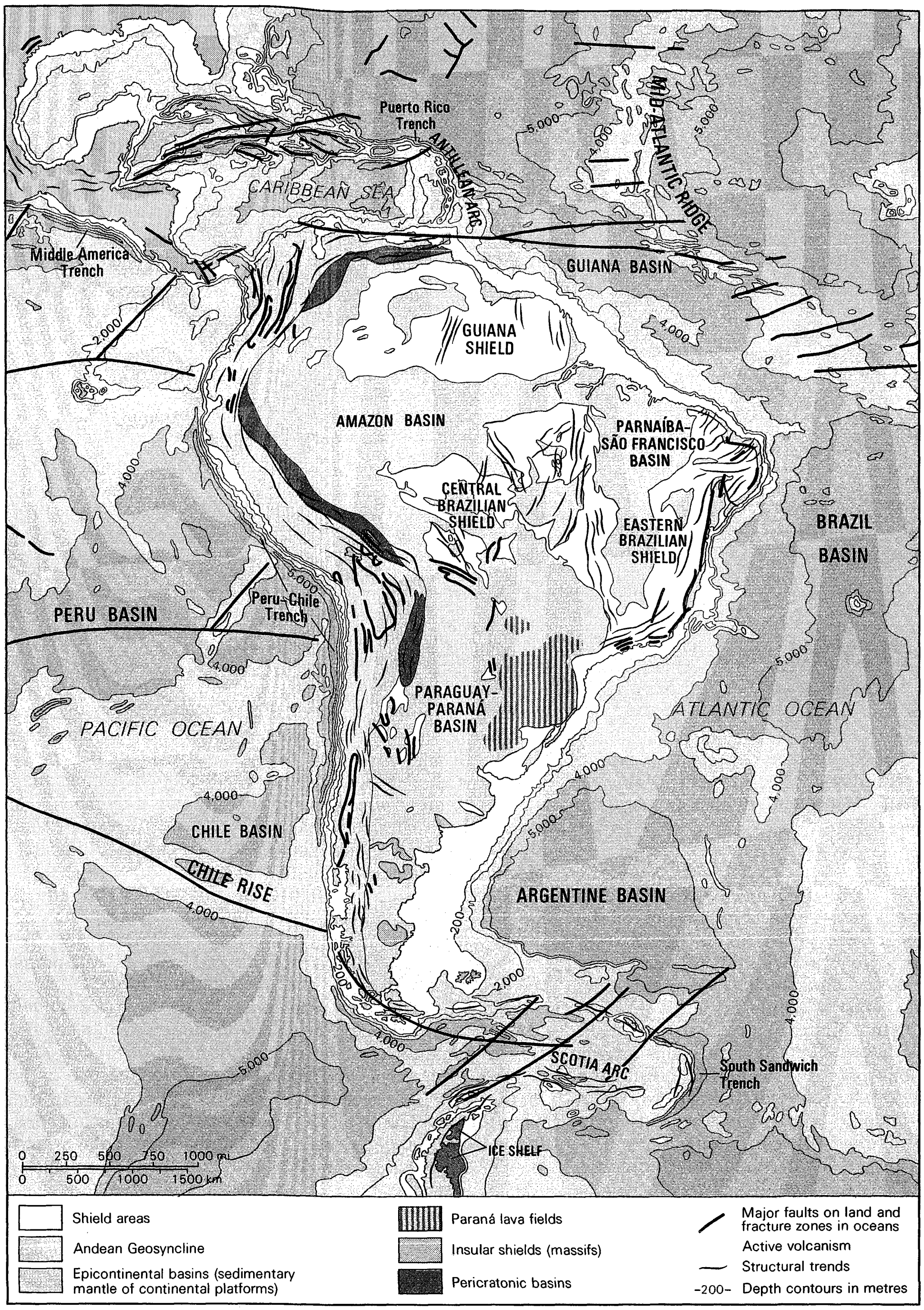
The belt of pericratonic basins lies along the western edge of the Brazilian shields and has a tendency to subsidence. It is long and narrow and runs from north of the Orinoco through Acre, Beni, and Chaco to south of the Río de la Plata. From the Precambrian Era to the beginning of the Mesozoic Era (about 225,000,000 years ago), seas invaded this slowly sinking region, depositing sediments, which slowly merge with continental strata on the east.

The Andean Geosyncline

The Andean Geosyncline occupies a zone stretching from the island of Trinidad on the northern coast to Tierra del Fuego in the extreme south. The original geosyncline contained sediments from the western highlands and from a continental mass that has now subsided under the Pacific, as well as of marine deposits that accumulated from the Late Precambrian Era onwards. Rapid and persistent subsidence was counterbalanced by the occurrence of extensive sedimentation. Movements of the Earth's crust began in the Paleozoic Era and were accompanied by intense volcanic activity, but the main activity began in the Early Mesozoic Era, reaching a maximum during the Middle and Upper Tertiary Period (from about 26,000,000 to 2,500,000 years ago). Movements continued up to the Pleistocene Epoch (from 2,500,000 to 10,000 years ago). Intense volcanic activity occurred from the Jurassic Period (from 190,000,000 to 136,000,000 years ago) to the Tertiary Period (from 65,000,000 to 2,500,000 years ago). Plutons (bodies of rock formed beneath the Earth's crust by solidification from a molten or partly molten state), running from Colombia in the north to Cape Horn in the south, form the bases of many cordilleras (groups of mountain ranges) and the summits of important massifs.

GEOLOGICAL DEVELOPMENT

The Precambrian Era is divided into early and late periods. The Serra do Mar and the "sugar loaf" formations



Structural features of South America.

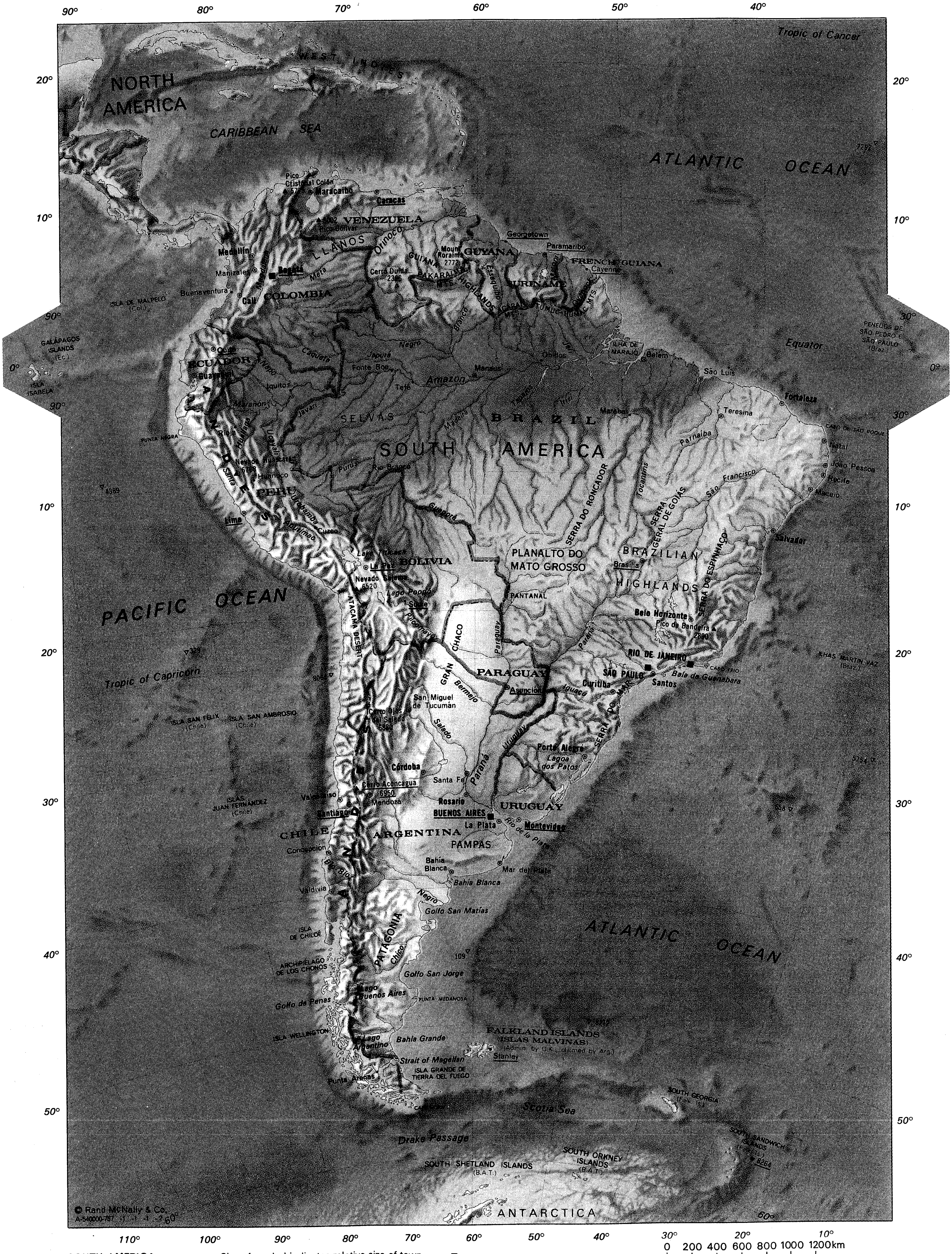
in the Baía de Guanabara (Bay of Guanabara), both in southern Brazil, belong to the early period; the so-called Minas series (a long belt of metamorphosed sediments located in the State of Minas Gerais, Brazil), on the other hand, which contains valuable minerals, belongs to the late period.

During the Cambrian Period (from 570,000,000 to 500,000,000 years ago) marine deposits accumulated in the Andean syncline and in the Amazon region of Brazil. Continental deposits of glacial or fluvial origin were laid down in various other districts, including northern Paraguay and the Brazilian Mato Grosso.

During the Ordovician Period (from 500,000,000 to 430,000,000 years ago) only the Brazilian shields were not beneath the sea, which covered the Gran Chaco region, the Paraguay Basin, and the whole of Patagonia. Thick beds of shale (laminated sediments consisting mostly of clay) and sandstone were consequently scattered from Huánuco, Peru, to La Rioja, Argentina.

Those parts of the geosyncline that extended from Tierra del Fuego to central Peru and from Quito in Ecuador to the Orinoco Basin were uplifted during the Silurian Period (from 430,000,000 to 395,000,000 years ago). Between these two ranges the sea invaded the Amazon Basin, and even the Parnaíba–São Francisco Basin. A geosyncline extended from Bolivia to the mouth of the Río de la Plata and to northeastern Argentina. During the late Silurian Period the sea regressed from almost all the land, except the Andean Geosyncline.

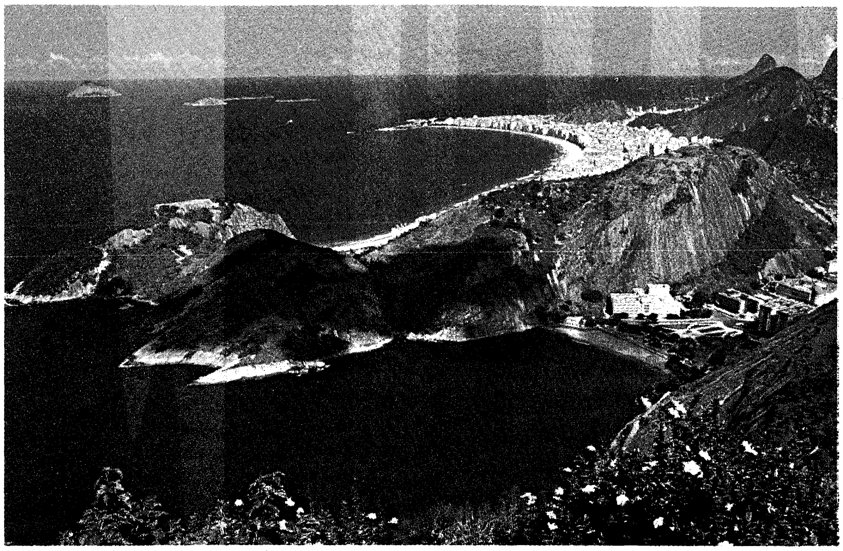
During the Devonian Period (from 395,000,000 to 345,000,000 years ago) the sea again invaded the whole Andean Geosyncline and later joined Amazonia and the Parnaíba Basin to Patagonia. During the later part of the period it regressed once more when mountain-building movements occurred in the Andean region. Much sedimentation occurred during this period, with sands being deposited in the Andean Geosyncline, sandstones in Amazonia, and conglomerates and sandstones in the Parnaíba Basin.



SOUTH AMERICA

Size of symbol indicates relative size of town • ◉ ■

Elevations and depths in metres



Exposed Precambrian rock along the coast at Rio de Janeiro, Brazil. The "sugar-loaves" in Baía de Guanabara are among the oldest rock formations of the continent.
Carl Purcell

The Carboniferous Period

During the Carboniferous Period (from 345,000,000 to 280,000,000 years ago) the sea was at first restricted to Chile, western Argentina, central Colombia, and part of the Parnaíba Basin but later invaded a large area from west of Lake Titicaca to Venezuela, depositing deep layers of sandstone in the Amazon Basin and limestone and shales in Peru and northern Colombia. Glaciation occurred south of a line from Lake Titicaca to São Paulo, as a result of which glacial sediments and moraines (deposits of boulders, gravels, sands, and clays left by glaciers) were laid down in the Paraná Basin.

During the earlier part of the Permian Period (from 280,000,000 to 225,000,000 years ago) a shallow sea invaded Bolivia; later in this period, however, the sea withdrew from the entire continent. Red beds (red coloured sedimentary rocks) and continental sandstones were deposited on the border of a geosyncline which extended from Peru to northern Argentina. In Peru, volcanic activity resulted in the depositing of volcanic detritus that had been explosively ejected from volcanic vents. The Andean region of the continent as a whole was subjected to movements of the Earth's crust that permitted basaltic or andesitic magma (mobile rock material of volcanic origin) to flow out.

South America was free from invasion by seas during the Triassic Period (from 225,000,000 to 190,000,000 years ago); red beds were deposited from Colombia in the north to Patagonia in the south, sometimes attaining thicknesses of almost 10,000 feet. In the later part of the period, the Paraná Basin became a huge desert; sandstones deposited at this time are at least partly of windborne origin and now cover more than 500,000 square miles.

During the Jurassic Period (from 190,000,000 to 136,000,000 years ago) the continent was generally not covered by the sea, although it periodically invaded some districts in the Andean region, depositing limestone beds containing ammonites (one of a large extinct group of mollusks) in many places from Colombia to Patagonia; the Chilean coastal plain (associated with the Andean Geosyncline because of the downward warping of the western section of the Chilean Andes) also remained maritime. There was much volcanic activity on the east side of the geosyncline; in Peru and western Chile it resulted in the depositing of lava and tuffs (rocks formed of compacted volcanic fragments) from about 30,000 to 60,000 feet thick, forming one of the largest volcanic masses on Earth. Volcanic activity continued to the end of the Mesozoic Era, when it reached its greatest extent, notably in the Paraná Basin where lava flows about 5,000 feet thick occurred, as well as along a lengthy belt from Venezuela to the lower Amazon and to the Parnaíba Basin. At the end of the Jurassic Period the rise of the Andes began.

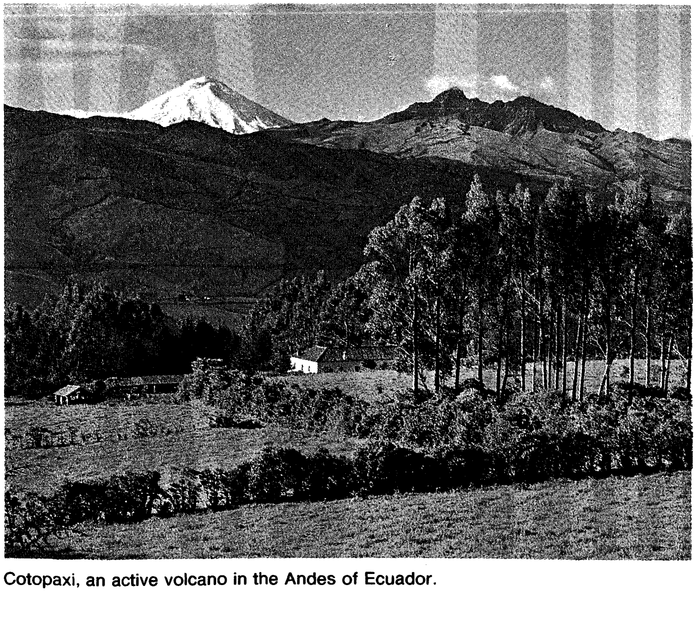
Beginning of the Andes

During the Cretaceous Period (from 136,000,000 to 65,000,000 years ago) seas invaded the larger part of the continent, with the Precambrian shields remaining as islands. The northern part of the Andean Geosyncline was still sinking rapidly, as indicated by the deposits of limestone about 46,000 feet thick that are to be found in Colombia. Continental sedimentation was active in the Amazonian, Paranaiba, and São Francisco basins, while layers of calcareous sand were deposited over the Brazilian Shield.

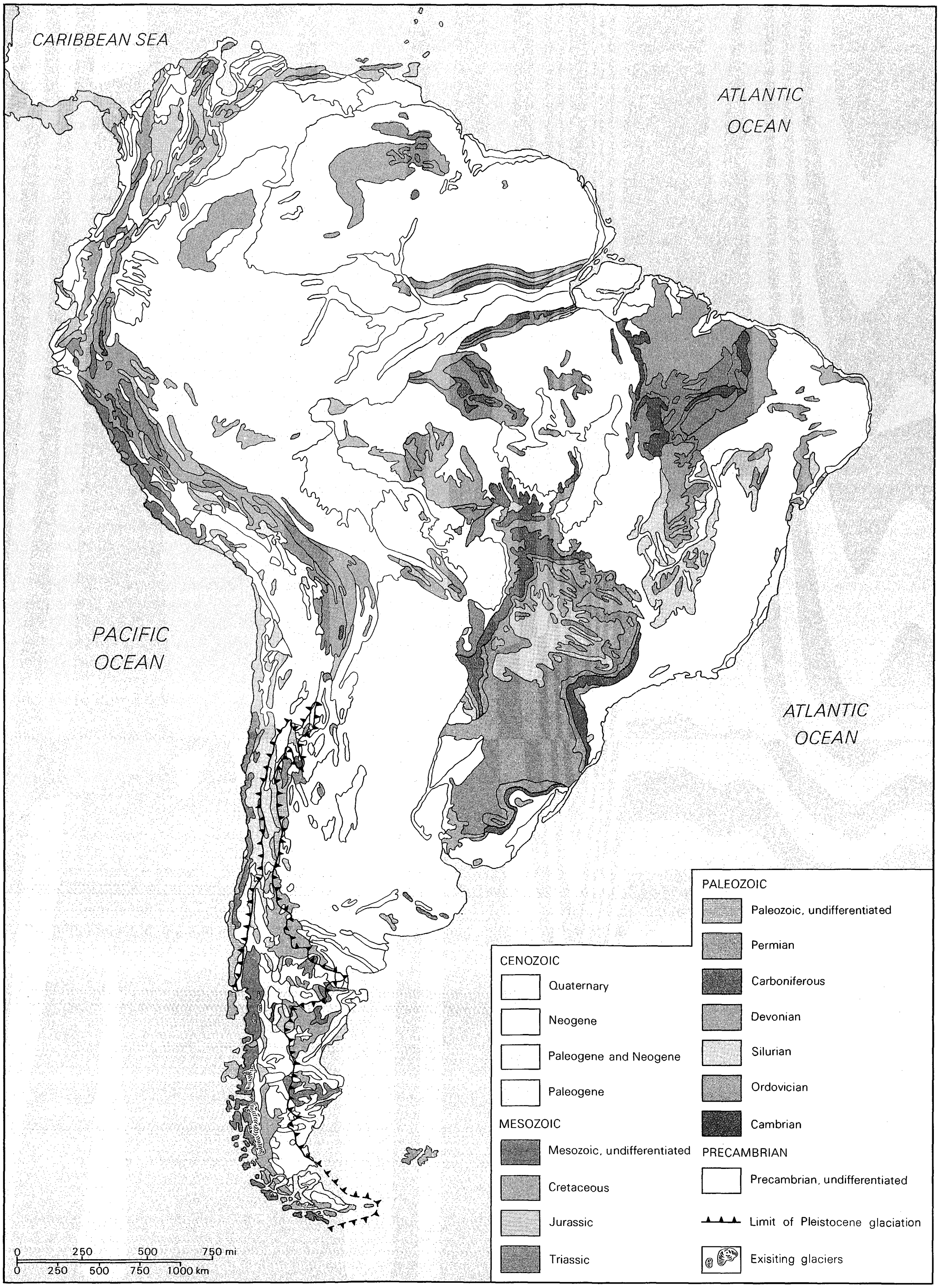
During the final part of the Cretaceous Period, the rise of the Andes reached its first maximum of activity with the uplifting of the Cordillera Central of Colombia, the Cordillera of Ecuador, the Cordillera Occidental of Peru, and the main ranges of Chile and Argentina. Mountain-building activity was strong along the length of the western Andes, and many intrusive rocks coalesced to form large and complex batholiths (very large masses of intrusive igneous rock), which are visible from Peru to Chile. Volcanic activity continued, and erosion resulted in continental sedimentation occurring in Peru, Bolivia, and Argentina (including the depositing of sandstones with abundant dinosaur remains in Patagonia). The sea regressed from the Amazon Basin at this time.

During the Eocene Epoch (from 54,000,000 to 38,000,000 years ago) of the Tertiary Period (from 65,000,000 to 2,500,000 years ago), great changes occurred. The sea

Shostal—EB Inc.



Cotopaxi, an active volcano in the Andes of Ecuador.



Geological structure of South America.

regressed almost everywhere, except in the northwest and in the centre of Patagonia between the Deseado massif and the cordilleras.

During the Oligocene Epoch (from 38,000,000 to 26,000,000 years ago) the sea returned, invading eastern Peru. The main event during the Tertiary Period was, however, the persistent uplifting and folding of the Andes, which assumed their present form in the Miocene Epoch (from 26,000,000 to 7,000,000 years ago). From this time onward the range formed a huge barrier and the gate between the Amazon region to the Pacific via Ecuador was finally closed. The Amazon Basin at first formed a huge freshwater lake but later drained into the Atlantic. At the same time the uplifting of the Andes determined mountain-building movements on their eastern border; nappes (laterally displaced, faulted, and overturned folds), sometimes more than 3,000 feet thick and several hundreds of miles long, were projected eastward against the Brazilian plateaus. Basins became filled with sediments eroded from the surrounding mountains, and much volcanic activity occurred.

evelopments n the leistocene poch

During the Pleistocene Epoch (from 2,500,000 to 10,000 years ago) sedimentation continued and the general uplifting of Patagonia occurred. The Pleistocene Epoch was mainly marked by important glaciation. Ice caps covered the Andes from Tierra del Fuego to Ecuador as many as four times, reaching the Atlantic in the south where moraines are to be found today in Patagonia; glacial terraces covered the plateaus at these times. Much loess (a loamy deposit, formed by the wind) deposition took place, particularly in the Argentine Pampas (grassy plains), which are almost entirely covered by a thin layer of continental deposits of the Quaternary Period (the last 2,500,000 years). Deposits from rivers or lakes covered more than 38,000 square miles in the Pantanal (the marshy plains in the Paraguay River Basin). Volcanic activity continued, especially in the Andes. Thick deposits of volcanic ash were laid down in the inter-Andean trough extending from Chile to Ecuador. Climatic changes following the last glaciation produced laterization (the formation of reddish residual soil, leached of silica and containing concentrations of iron and aluminum hydroxide) and large-scale soil modifications in Brazil and the Guianas (Guyana, French Guiana, and Suriname). Gigantic lakes formed in the Andean region as the ice melted included Lake Ballivián, which covered the area now occupied by Lake Titicaca, as well as others in Argentina and Paraguay.

The land

RELIEF

South America has two major mountain systems of unequal mass. The young cordilleras of the Andes in the west extend down the entire continent from north to south, bordering the Pacific. On the eastern side are the ancient Guiana and Brazilian highlands, which are much less high and slope gently to the west; further south are the Patagonian plateaus. Lowlands—the Orinoco, the Amazon, and the Paraguay basins and the Pampas—divide the highlands from one another. The relief of the whole continent shows an imbalance; a drop of rain that falls only 100 miles east of the Pacific will flow down to the Atlantic, 2,500 miles away.

The Andes. The Andes mountain ranges, about 5,500 miles long and second only to the Himalayas in average height, constitute a formidable and continuous wall with many summits more than 22,000 feet high. Near the southwestern frontier of Colombia and through Ecuador they form a single mountain mass, from which three distinct massifs—the Cordilleras Real (Oriental), Central, and Occidental—branch out. The valley of the Río Magdalena, between the Oriental and the Central ranges, and the valley of the Río Cauca, between the Central and the Occidental ranges, are in fact huge rifts. An overall view of the Andes in Colombia shows, within relatively short distances, a succession of hot lowlands and high ranges with a cold climate.

he olombian ndes

In Peru the Andes form two parallel cordilleras, one along the Pacific and the other descending eastward toward the valleys of the Ucayali and Urubamba rivers. The western range is the highest, particularly in central Peru where it splits into two chains; its impressive snowclad peaks include Huascarán (22,204 feet). Between the two cordilleras extends the Altiplano—a vast complex of high plateaus between 12,000 and 15,000 feet high and sometimes as much as 125 miles wide. This Altiplano forms a maze of valleys, hills, and vast plains without equivalent except in Tibet. Water accumulates in closed basins to form marshes and lakes, one of which is Lake Titicaca. This huge massif has been dissected by rivers such as the Marañón, Huallaga, Apurimac, and Urubamba, all of which have cut spectacular gorges down the eastern slopes.

In the south, the Andes form a single but very complex chain with many peaks; one of which—Aconcagua—at 22,834 feet (6,960 metres), is the highest mountain in the Western Hemisphere. In Patagonia, a part of the cordillera runs beneath the sea, forming innumerable islands with steep slopes. The Andes have also been deeply carved by glaciers, which still occupy about 1,900 square miles, constituting an ice cap with long terminal tongues running into lakes or into the sea.

The Andes are studded with volcanoes that form part of the so-called "ring of fire"—the vast circle of volcanoes that border the Pacific. Earthquakes are frequent; almost every city has at one time been totally demolished, even along the coastal plains, where clear signs of recent vertical movement are visible.

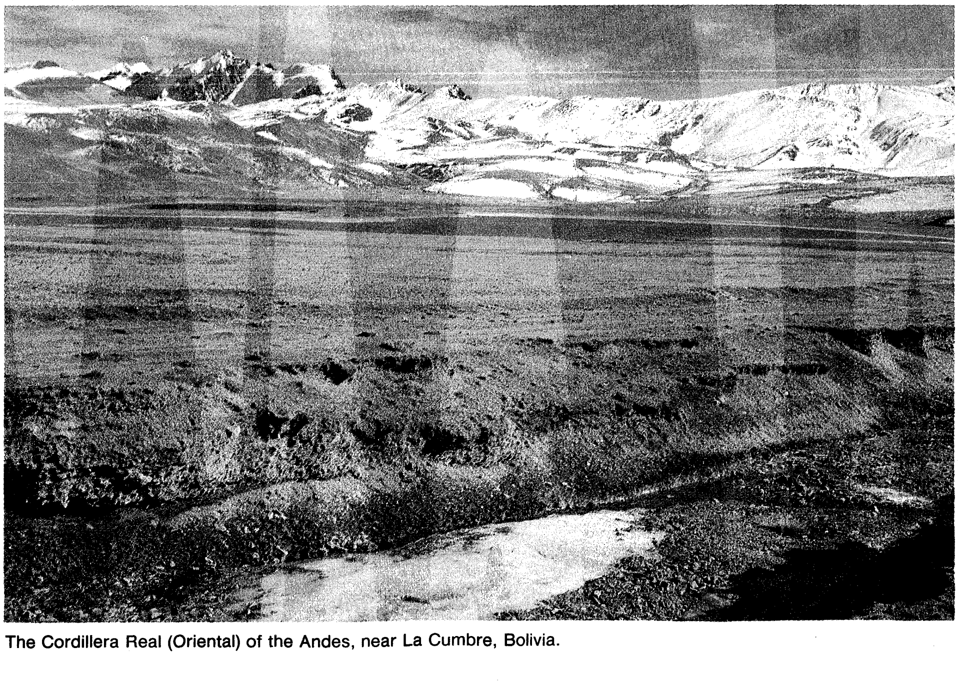
Detailed discussion of the Andes can be found at the end of this article.

The eastern plateaus. To the east, the Guiana Highlands and Brazilian plateaus consist of very old crystalline rocks, most of them greatly worn down. The Guiana Highlands are a monotonous expanse, about 800 feet high, forming a series of small hills which are separated by marshy depressions and are surmounted by dome-shaped granitic inselbergs (steep-sided residual hills), some 2,000 feet high. The southern edge rises abruptly to a mountain chain, the Tumucumaque, and to some high tablelands, such as Mt. Roraima (9,094 feet) and Mt. Duida (7,859 feet).

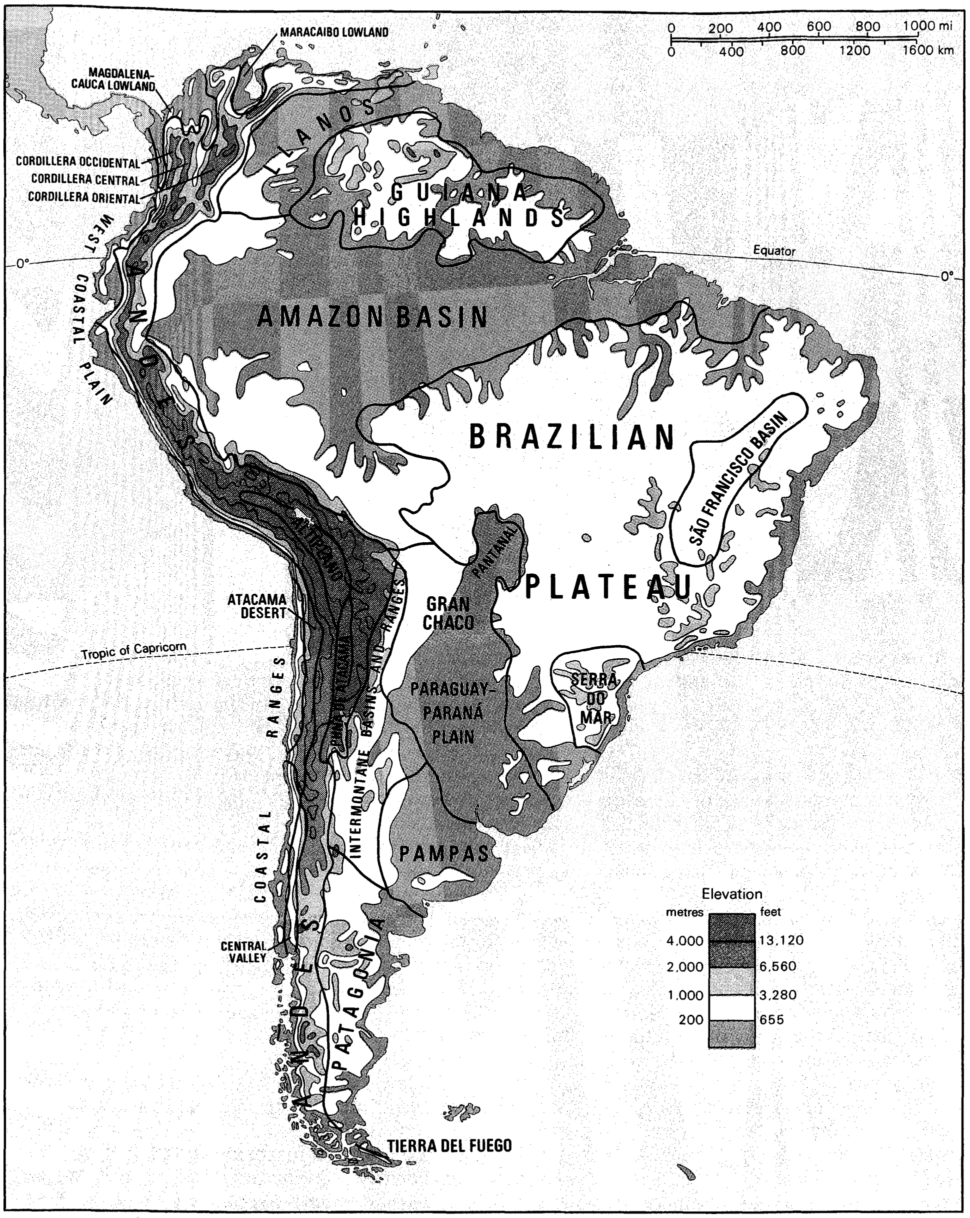
The Brazilian Plateau

Covering an area of about 580,000 square miles, the Brazilian Plateau (also called the Brazilian Highlands)

Shostal—EB Inc.



The Cordillera Real (Oriental) of the Andes, near La Cumbre, Bolivia.



Physiographic regions of South America.
Adapted from *Odyssey World Atlas*, © General Drafting Co. Inc.

rises to about 3,000 feet and is crowned by short ranges of hills; the São Francisco River has cut deeply into it. In the north the plateau slopes gently down to the sea, but in the east it drops abruptly down, sometimes as much as 2,600 feet. Skirting its southern edge, the Serra do Mar has summits of more than 7,000 feet—such as the Pedro do Sino (7,365 feet). The sea has partly invaded the original coastal ranges; there rocky masses, such as the Pão de Açucar, or Sugar Loaf (1,279 feet), and the Pico do Corcovado (2,309 feet), have been isolated by the sinking of surrounding land.

Eastern Patagonia constitutes a series of gigantic tablelands covered with rounded pebbles and crumbling sandstones. Volcanic eruptions have also spread layers of basaltic lava flows and have dotted the sedimentary plateaus with volcanic cones.

The lowlands. The Orinoco Basin, which is also called the Llanos, extends between the coastal ranges of the Venezuelan Andes and the Guiana Highlands and is covered with alluvia which has been brought down by the Andean torrents.

The Amazonian depression, the largest river basin in the world, forms an enormous drainage area of 2,722,000 square miles, bounded by the Andes to the west, the Guiana Highlands to the north, and the Brazilian plateaus to the south. The ancient platform of primary rock that underlies the Amazonian depression is covered with layers of alluvial sand and clay, so that it forms an immense plain of low undulations, the general incline being extremely slight. Iquitos, in Peru, is at an altitude of only 384 feet, while Manaus, in the heart of the basin, has an altitude of 144 feet.

The basin of the Paraguay River, between the Bolivian highlands in the west and the Brazilian plateaus in the east, consists of a series of vast alluvial plains drained by a complex network of rivers interspersed with marshes. These plains, called the Pantanal, are only a few hundred feet above sea level and are subject to annual floods; they form an immense swamp during the rainy season.

The Pampas, covering almost 300,000 square miles, consist of an immense accumulation of loose sediments brought down from the Andes. These deposits, 1,000 feet deep at Buenos Aires and much deeper at some other places, have completely buried the ancient features of the land. The landscape is monotonous and seems perfectly level, though in fact it rises imperceptibly toward the west—rising from an elevation of 66 feet near Buenos Aires to 1,640 feet at Mendoza. Some hills, such as the Córdoba and San Luis ranges, stand out from the plains like islands.

Tidewater glacier at the end of a fjord in the Tierra del Fuego region of southern Chile.
Reflejo—EB Inc.

Detailed discussion of the Gran Chaco, the Patagonian Desert, and the Llanos can be found at the end of this article.

CLIMATE

As South America extends over a vast range of latitudes, it is hardly surprising that it has a great variety of climates. Although it is the only southern continent to reach such a high latitude, it nevertheless has its broadest extent in the equatorial zone, so that tropical conditions prevail over half of the continent.

ir
rculation
atterns

Factors influencing climate. The circulation of air over the continent is influenced by winds related to the great anticyclones (circulation of winds around central regions of high atmospheric pressure) of the Atlantic and Pacific oceans. Trade winds blow from the northeast in the Northern Hemisphere and from the southeast in the Southern Hemisphere. Another southern high-pressure region is located off the west coast, so that strong westerly winds blow toward the continent. Two low-pressure centres—one situated over the south polar region, the other extending from northeastern Brazil to the Gran Chaco region of Argentina—affect wind patterns inland.

Landforms largely influence specific circulation patterns. Above all, the Andes prevent winds from the South Pacific anticyclone from reaching the interior of the continent.

Ocean currents have a strong influence, particularly the so-called Peru (Humboldt) Current, which in effect is a series of upwellings along the steep Pacific coast, which result in cold water being carried to the latitude of southern Ecuador, where the flow turns westward. Average surface temperatures vary from about 58° to 64° F (about 14° to 18° C)—about 18° F (10° C) lower than is usual in such latitudes. The east coast is washed by the cold Falkland Current in the south and by the warm Brazil Current in the north.

Finally, altitude plays an important part. Wide areas located within the tropical zone were uplifted by the rise of the Andes.

Climatic regions. South America can be divided into several well-marked climatic regions, as determined by the factors already mentioned. Four major types of climate occur—tropical, temperate, arid, and cold.

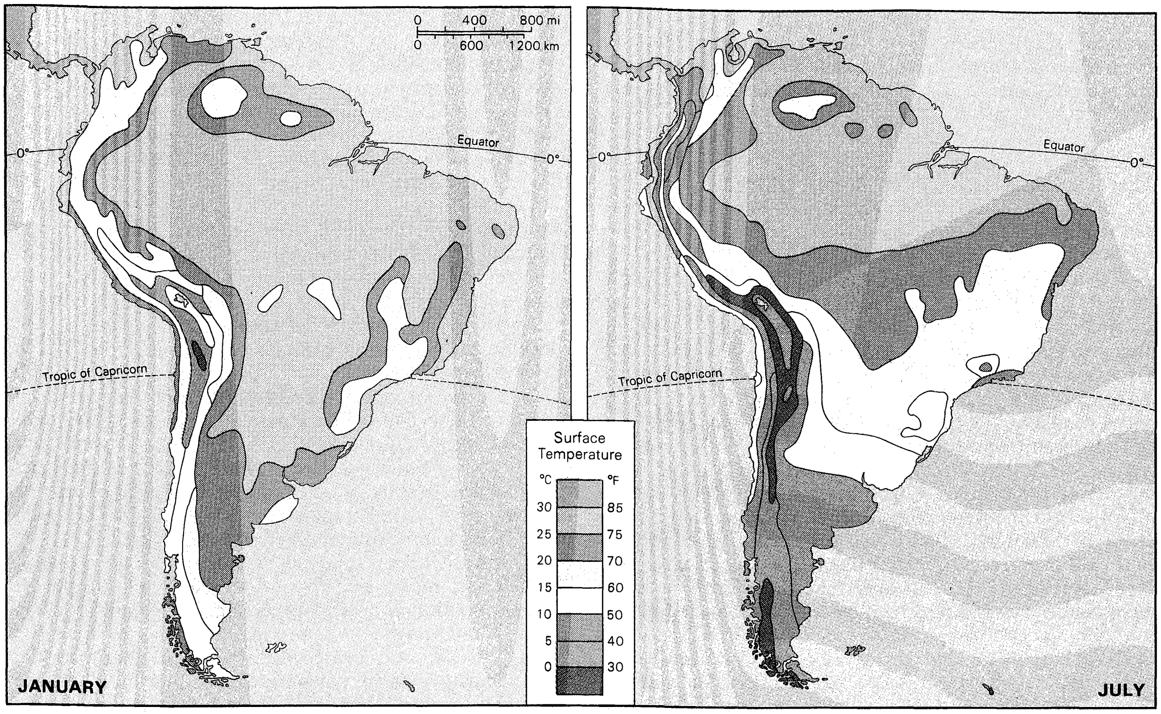
Among the tropical climates, the tropical rain forest type occurs on the Pacific coast of Colombia, in the Amazon Basin, and on part of the coast of Brazil. The average temperature is around 86° F (30° C), with monthly and annual variations of less than about 5° F (about 3° C); heavy rainfall, evenly distributed all the year round, averages about 98 inches (2,490 millimetres) in Pará, about 108 inches in Iquitos, 71 inches in Manaus, and 354 inches in the Chocó region of Colombia—one of the most watered zones in the world, where it rains more than 300 days a year. In the Amazon region, rains do not fall regularly all over the basin. The southern part receives rain especially during the southern summer (from October to April), while the northern part has its rainy season during the northern summer from May to September. The "dry" season is neither lengthy nor noticeable; humidity is always high.

The second type of tropical climate—the savanna (grassy parkland) type—is characterized by high temperatures (with monthly minimum temperatures above 64° F, or 18° C) but receives less precipitation and experiences a definite dry season. It is found around the rain-forest belt, in the Orinoco Basin, on the Brazilian Highlands, and in part of western Ecuador. Temperatures are still high and annual variations small, but average daily temperature extremes are greater, ranging from 42° to 63° F (6° to 17° C). There is a prolonged dry season.

Temperate climates, characterized by lower winter temperatures, are found south of the Tropic of Capricorn (in Paraguay, parts of Bolivia, Brazil, Argentina, and Chile) and in the Andes, where, in their rainfall pattern, they resemble tropical climates. On the Atlantic side, temperatures in the warmest month average 77° F (25° C), but cold-month averages vary from 63° F (17° C) in the north (Asunción, Paraguay) to 50° F (10° C) in Buenos Aires; rainfall is greater than one and a half inches each month in the east but decreases to the west. In central Chile, between latitudes 32° and 38° S, the features of the climate are similar to those of the Mediterranean, with mild winters and winter rains; summers are, however, cooler (69° F, or 21° C, in Santiago, Chile, in January—19° F, or 5° C, cooler than in the Mediterranean). In southern Chile, winter temperatures are lower, but not as low as the latitudes would seem to indicate. The islands and channels have a relatively uniform climate throughout the year, and winters are much less severe than in Labrador, which is at a comparable latitude north of the Equator. The presence of glaciers is the result of cold and cloudy summers during which ice does not melt. Rainfall is very high (102 inches in Valdivia, Chile, and probably twice this figure on the western slopes of the mountains), while the coast is one of the most heavily watered regions in South America. A short distance inland, rainfall decreases considerably (20 inches at Ushuaia, Argentina, which is protected from the westerly winds). Thus, in Patagonia there are surprisingly greater differences in climate from west to east than from north to south. Winds are very violent.

Arid zones

Arid climates are found in four areas. Patagonia east of the Andes is in the rain-shadow zone, and rainfall is low (about four inches in San Juan, Argentina). The annual range in temperature is more than 36° F (20° C)—the



Average temperatures for January and July for South America.

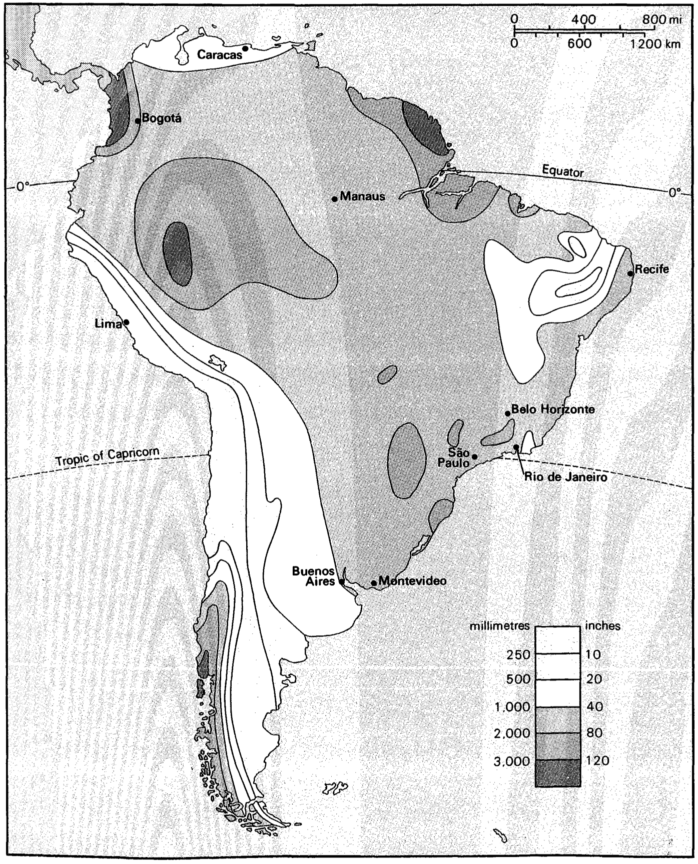
Adapted from Norton S. Ginsburg (ed.), *Aldine University Atlas* (1970), Aldine Publishing Co., Chicago; copyright © 1970 by George Philip & Son Ltd., London; with permission from the author and Aldine-Atherton, Inc.

highest in South America—and is caused by warm summers and cold winters. Another arid zone is found in a narrow coastal strip along the Pacific coast between latitudes 5° and 31° S. The cold seas (the Peru, or Humboldt, Current) and the proximity of the high Andes produce an inversion of normal atmospheric temperatures, as air in contact with the water cools more rapidly than the upper strata of air. The result is a cloud layer about 1,200 feet thick, lying at altitudes varying between about 1,000 and 3,000 feet, that prevents the warming of the air near the ground. Temperatures are consequently low: Lima has an average annual temperature of 64° F (18° C), ranging from about 72° F (22° C) in February to about 59° F (15° C) in August. The coast of Peru is thus the cloudiest desert in the world, with no sunshine for at least six months of the year. It almost never rains, except under abnormal circumstances, but condensation of fog (called *garúa* by the Peruvians) provides some humidity. Another desert extends from northeastern Colombia to Venezuela, covering a zone where rains are scarce and droughts prolonged.

Finally, an arid zone occurs in northeastern Brazil, between the Parnaíba and São Francisco rivers. The interior highlands act as a wedge separating the sea winds from the northeast and those from the southeast, all of which carry their moisture beyond the region. Average annual rainfall is less than four inches, and the dry season may last as long as seven months. The worst feature of the climate is the irregularity of the rainfall, as a result of which severe droughts plague the region.

Cold climates, with average annual temperatures of less than 50° F (10° C), occur in the southernmost parts of Argentina and Chile and in the high Andes above about 11,500 feet. Mean temperatures are relatively low, but daily variations are wide. There is a marked difference in humidity between the northern and southern parts of the upper Andean zone. In Colombia and Ecuador the climate at such altitudes is cool and damp. Temperatures, always low, may on the average vary daily from 54° F (12° C) during the daytime to 28° F (−2° C) at night. Rainfall is high and well distributed throughout a good part of the year. Clouds and mist are dense, and sun shines only for short periods. From central Peru to Bolivia and Chile, temperatures are still lower. Near Lake Titicaca, the average annual temperature is only about 34° or 36° F (1° or 2° C); November is the warmest month, with a temperature of 41° F (15° C), while the coldest month, July, has an average of 28° F (−2° C). Daily variations are considerable; the daily maximum of 68° F (20° C) drops to a nocturnal minimum of 5° F (−15° C). Annual rainfall varies from 24 to 56 inches but is concentrated during the southern summer, from November to April. The dry season is long and is characteristically accompa-

Adapted from Norton S. Ginsburg (ed.), *Aldine University Atlas* (1970), Aldine Publishing Co., Chicago; copyright © 1970 by George Philip & Son Ltd., London; with permission from the author and Aldine-Atherton, Inc.



Average annual precipitation in South America.

nied by drought. Winds are constant and often violent, accelerating the coldness and the dryness of the climate, which produces a harsh and hostile environment. As in any mountainous region, climate varies largely according to local conditions.

DRAINAGE

Rivers. Drainage has been deeply affected by the dissymmetry of the continent. The largest basins are found on the east side, and the main rivers run to the Atlantic Ocean.

'he ›mazon :iver

By far the largest river is the Amazon River, which, with a length of 4,000 miles, is second only to the Nile; the volume of water carried, however, surpasses that of all other rivers, comprising a fifth of all the flowing water of the world. About 6,350,000 cubic feet (179,800 cubic metres) of water per second is emptied into the Atlantic by the Amazon—10 times more than the Mississippi. The Amazon drains more than three-tenths of South America and has more than 1,000 tributaries, seven of which are more than 1,000 miles long. The source of the Amazon (although not the most distant of its headwaters from the sea) is in a small lake (Lauricocha) at an altitude of 15,875 feet (4,840 metres) in the Cordillera Huayhuash, near the Cordillera Blanca, Peru. It is named the Marañón in its upper course; after being joined by several rivers, it escapes from the Andes through narrow canyons. It is from the source of one of these rivers, the Ucayali, that the Amazon's length is measured. Below Iquitos, Peru, the river—often called Río Solimões—is joined by the Río Napo, which flows out of Ecuador; further on, near Manaus, it is joined by the mighty Rio Negro, which drains a large part of northern Brazil. The Amazon, then at full strength, winds through the low plains to pass between the Guiana Highlands and Brazilian Highlands before emptying into the Atlantic.

As rains do not fall regularly year round all over the basin, the river has two annual floods in its upper course, which is subject to the alternate influence of the tributaries that descend from the Peruvian Andes (where rains fall from October to January) and the tributaries on its left bank, which descend from Ecuador (where rains fall from March to July). This pattern of alternation disappears further downstream, the two seasons of high flow gradually merging into a single one, which progresses slowly downstream in a gigantic wave from November to June. The height of the floods is lower further downstream. At Tefé, the difference between the high and low water mark is 50 feet and, at Óbidos, Brazil, only 25 feet. The width of the river likewise varies; exceptional floods may extend more than 60 miles beyond the main river bed. Because of the low gradient of the basin, tides are perceptible as far as Óbidos, 600 miles from the Atlantic coast.

The Paraguay, Paraná, and Uruguay system

The second important drainage system, variously estimated to cover from 1,600,000 to 1,700,000 square miles, is formed by the Paraguay, Paraná, and Uruguay rivers, which empty into a vast expanse of the water, the Río de la Plata—actually not a river but a gulf. About 2,800,000 cubic feet of water per second flow from the common mouth of these rivers. The Paraguay River, with a length of 1,584 miles, rises in the Bolivian hills; it is a river of the plains, flowing across a wide stretch of marshes (the Pantanal) in its middle course; its lower course is, however, drier. The Paraná River, with a length of 2,485 miles, flows mainly among high plateaus.

The Orinoco River, 1,337 miles long, is the third largest drainage system, covering about 365,000 square miles. It flows west and then north; plunging down a series of steep slopes, it then flows along an almost flat basin. Near the ocean, it divides into a series of distributaries that form the Delta Amacuro, which has an area of about 5,000 square miles. The rainy season from April to October causes annual floods, which at Ciudad Bolívar, Venezuela, reach a height of 60 feet in July. By means of the Casiquiare Canal, it communicates with the Amazon system through the Rio Negro.

The São Francisco River, with a total length of 1,800 miles, rises in the State of Minas Gerais, Brazil, and flows northward for 1,000 miles before curving eastward to the Atlantic. The river has been an important artery of communication since colonial times and is now the location of large hydroelectric projects in operation since 1955, at Cachoeira da Paulo Afonso (Paulo Afonso Falls), and at Três Marias, north of Belo Horizonte.

Jacques Jangoux

Lake Titicaca, the world's highest large lake, located between Peru and Bolivia. The Cordillera Real of the Andes rise along the northeastern shore of the lake.

Among other important rivers are the Río Magdalena in Colombia (navigable in two sections separated by rapids at Honda), and the Essequibo, Maroni, and Oyapock rivers in the Guianas.

Drainage to the Pacific is different because of the proximity of the Andes from the Pacific coast and the scarcity of rains from southern Ecuador to northern Chile. The rivers are consequently very short, and few convey any large quantity of water. The only important rivers are the Río Guayas in Ecuador, on which the port of Guayaquil is located, and the Río Santa in Peru. Some torrents had great importance since early times, when good water management permitted ancient civilizations to develop. In central Chile the valleys of Río Aconcagua and Río Bío-Bío are very fertile.

Detailed discussion of the Amazon River, the Orinoco River, the Rio de la Plata system, and the São Francisco River can be found at the end of this article.

The Andean lakes

Lakes. Most of the lakes are mountain lakes located in the Andes or along their foothills. Because of the complex topography, water has accumulated in closed basins. Among permanent lakes, the largest is Lake Titicaca, which lies at an altitude of 12,500 feet (3,810 metres) between Peru and Bolivia; the lake measures 110 by 35 miles and was much wider in the past. Lago de Junín in central Peru; Lake Sarococha, also in Peru, between Puno and Arequipa; and Lago Poopó in Bolivia also rank among the larger Andean lakes. These create uniform physical conditions throughout the year in terms of temperature and percentage of dissolved gas; up to an altitude of 16,000 feet they never freeze. In consequence, the climate of their shores is temperate.

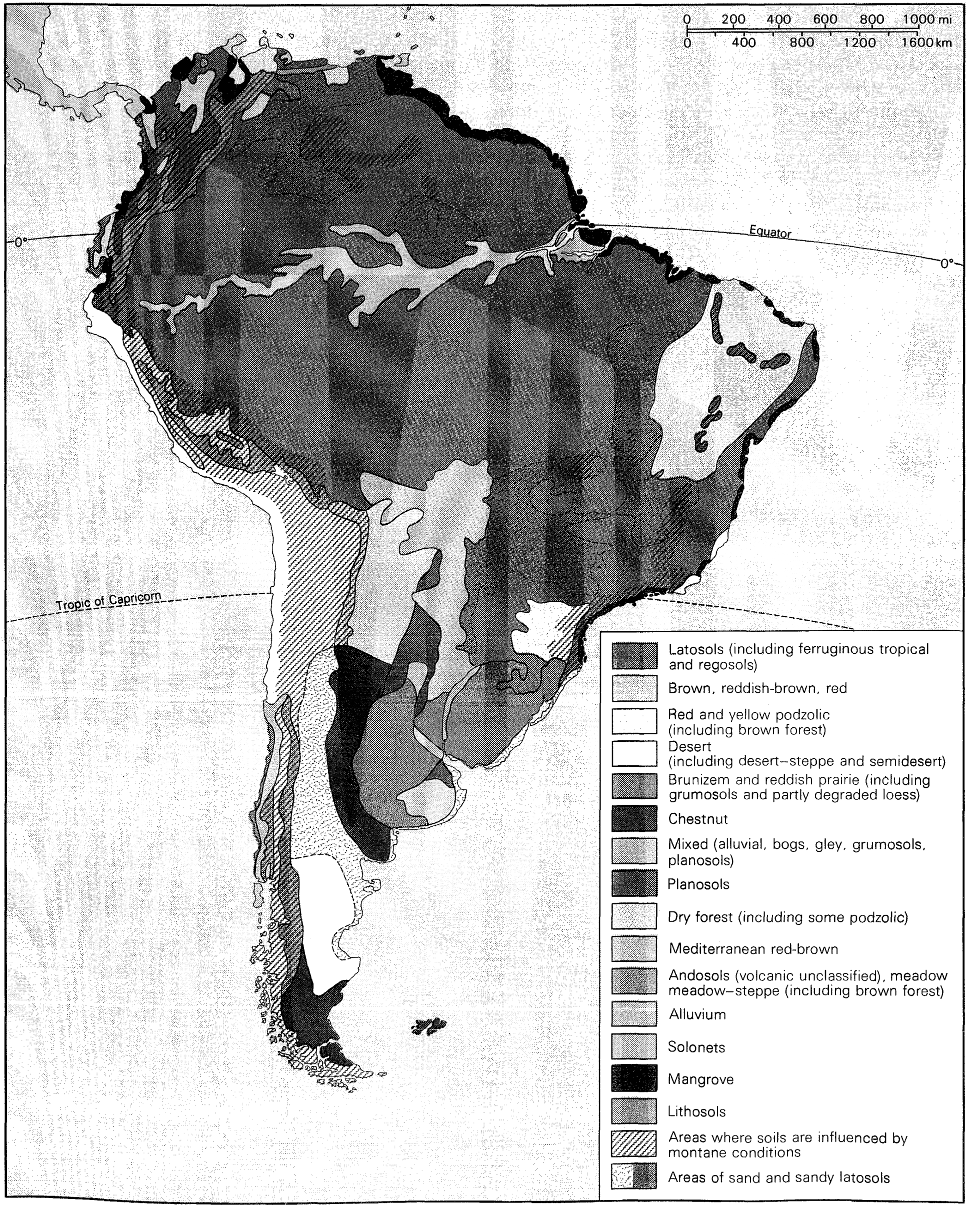
Another type of lake is to be found in Patagonia where, in the wake of melting glaciers, lakes have formed, such as Lago Buenos Aires, Lago Argentino, and Lago Nahuel Huapí. Their eastern parts, which stretch to the end of the Argentinian plateau, generally have gently sloping banks bordered by low mountains, while their western parts form a series of narrow arms, like fjords, lying between steep slopes.

Marshes and swamps. Marshes are to be found in depressions in many parts of the continent. One of the widest marshy areas is the Pantanal, in the middle valley of the Paraguay River; it is subject to flooding in December, reaching its highest watermark in June, when it becomes an immense swamp; similar habitats are to be found in the Orinoco Basin.

Swamps of another type are found in rain forests, mostly in the Amazon region and in northwestern Colombia. The ground is inundated or very marshy throughout the year or else only at the time of the annual flood.

Finally, the mouths of the Orinoco and of the Amazon

Adapted from R. Ganssen and F. Hadrich (eds.), *Atlas zur Bodenkunde*


0 200 400 600 800 1000 mi
0 400 800 1200 1600 km
Equator
0°
0°
Tropic of Capricorn
Latosols (including ferruginous tropical and regosols)
Brown, reddish-brown, red
Red and yellow podzolic (including brown forest)
Desert (including desert–steppe and semidesert)
Brunizem and reddish prairie (including grumosols and partly degraded loess)
Chestnut
Mixed (alluvial, bogs, gley, grumosols, planosols)
Planosols
Dry forest (including some podzolic)
Mediterranean red-brown
Andosols (volcanic unclassified), meadow meadow–steppe (including brown forest)
Alluvium
Solonets
Mangrove
Lithosols
Areas where soils are influenced by montane conditions
Areas of sand and sandy latosols

Soils of South America.

rivers form wide marshy deltas, while mangrove swamps of various types are to be found from southern Ecuador to Santos in Brazil.

SOILS

As a consequence of its geological history, geomorphology, and climate, no less than 26 soil regions, many of which are split into several subdivisions, can be distinguished in South America. They consist of three major groupings, which correspond to the continent's three major geographical components—the lowlands, the uplands, and the Andes.

Low natural fertility is an outstanding feature of South American soils. About 50 percent of the continent's soils are made up either of unconsolidated and nutrient-poor sediments (such as china clays and quartz sands) in the river basins, latosols (red residual soils, also called laterite, leached of silica and containing concentrations of iron and aluminum hydroxides), red-yellow podzols (highly bleached soils that are low in iron and lime), and regosols (azonal soils consisting chiefly of imperfectly consolidated material and having no specific morphology). About 20 percent of the continent is covered by desert soils of various types in which agriculture is hazardous because of lack of water. Other regions, representing about 10 percent of the total area, suffer from poor drainage, the soils being gleys (layers of soil in which the material is blueish grey, generally sticky, and often structureless, due to excessive moisture), groundwater laterites, grumosols (soils formed of granules), and planosols (a type of soil found in humid climates in which soluble salts and minerals are leached out of the upper layers and are cemented or compacted at a lower level). In the Andes, slopes are often steep, and lithosols (soils with no clear morphology, consisting of imperfectly weathered masses of rock fragments) abound, accounting for about 10 percent of the continent's surface. In the inter-Andean valleys and on some of the foothills, nevertheless, eutrophic soils (deposited by lakes, and containing much nutrient matter, but often shallow and subject to seasonal oxygen deficiency) can be found.

ıe fertile
il
gions

Fertile soils extend over only about 10 percent of the surface of South America. The most important of these are brunizems (any of a zonal group of deep dark prairie soils, developed from loess; *i.e.,* unstratified loam deposited by the wind), chestnut soils, ferruginous tropical soils, red latosols, and reddish-brown lateritic soils. On the low coastal ranges and in the foothills of the western Andes and on the nearby plains and terraces of Colombia and Ecuador, the soils consist mainly of red-yellow latosols, podzols, and alluvial soils. Soils in southern Brazil and Uruguay are not uniform and consist of brunizems, reddish prairie soils, and planosols. The Argentinian Pampas, the largest fertile area on the continent, is uniformly covered with the so-called pampean loess, which is calcareous, rich in minerals, and partly of volcanic origin. Other rich soils are found in the uplands of northeastern and central Brazil; they consist of sandy regosols in the north and red latosols in the south.

The agricultural development of South America reflects exactly the distribution of soils according to their fertility. It is mostly confined to the eastern temperate plains, in which is concentrated the production of wheat and corn and the grazing of cattle; to the subtropical and temperate parts of the Andes, from Colombia to Chile, where grazing takes place and a variety of crops are cultivated; and to eastern and southeastern Brazil, where coffee, cacao, and sugarcane are grown, while the interior plateaus are devoted to cattle grazing. The Amazon Basin is still a region of elementary agriculture, though it produces cacao and rubber.

Soil erosion has ravaged a large part of the continent. According to some estimates, in several countries up to 50 percent or more of the presently arable land has been severely affected or ruined by bad land use. In the Andes, land that once produced up to 25 bushels of wheat to the acre is now abandoned; mountain forests are still cleared for cattle grazing and cultivation, which ruins the soil of the region for years after. Soil damage has been slight in the areas in which the topography is relatively smooth. Campaigns for soil conservation or restoration are now in progress in most countries.

PLANT AND ANIMAL LIFE

South America, with its distinct vegetation and animal life, is the very heart of the neotropical region (that part of the Americas that extends from the Tropic of Cancer southwards). With Australia it is, because of its isolation from the rest of the world during the Tertiary period, the landmass with the strongest biological originality.

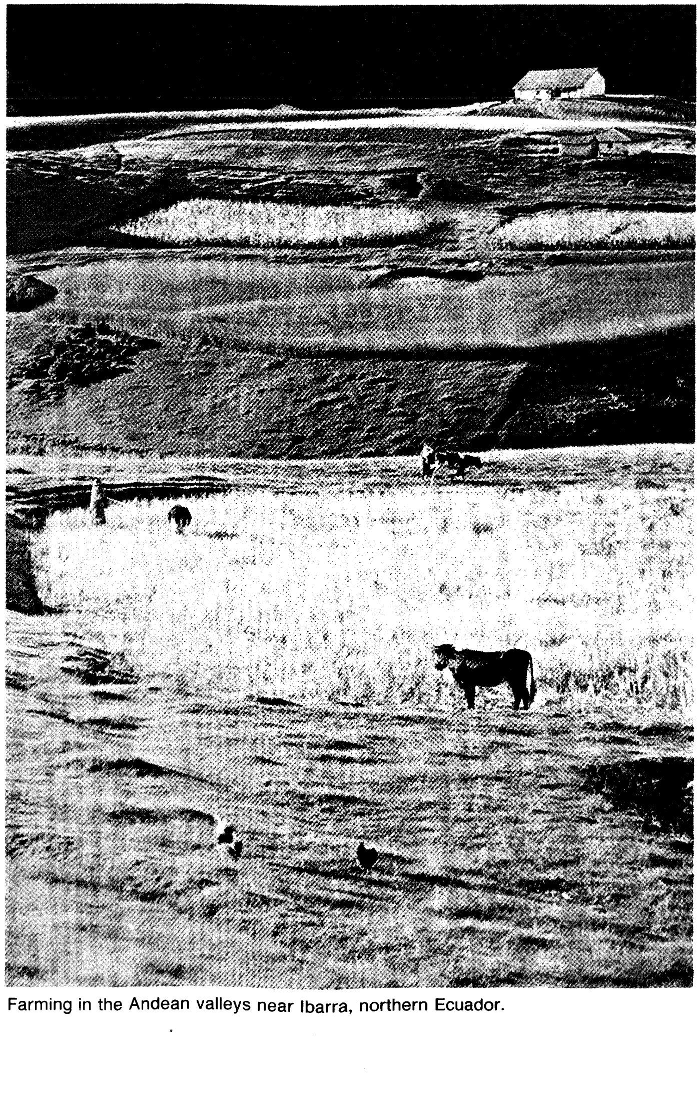
Vegetation and animal life are a mixture of elements of various relationship. Ancient groups of plants and animals including mollusks, chilopods, some fishes, reptiles, and amphibians show affinities with the animal life of Africa, Australia, and New Zealand. More recent elements migrated from or through North America during the late Tertiary Period and the Pleistocene Epoch; most of these are vertebrates.

The pattern of distribution within the continent is much diversified because of the variety of climatic and ecological zones. The northern tropical regions are the richest, while the southern part and the Andean highlands are much impoverished, despite some differentiation.

Vegetation. The origin of South American vegetation is much diversified, as is shown by the relationship of its components. The proportion of endemic plants is very high, even at the family level. Among angiosperms (plants having seeds enclosed in an ovary) no less than 25 families and 3,500 genera are endemic in the neotropical region. Others are related to African plants or belong to southern plant groups also distributed in Southern Africa and in Australasia.

Vegetation is by no means uniform all over the continent; its distribution is determined by climatic, geographic, and

Reflejo—EB Inc.



Farming in the Andean valleys near Ibarra, northern Ecuador.

soil differences. The main geographic plant zones are as follows:

Tropical and subtropical rain forests. Rain forest covers the largest part of the Amazon region, most of the Guianas, southern and eastern Venezuela, the Atlantic slopes of the Brazilian Plateau, and the Pacific coast of Colombia and northern Ecuador. The huge Amazon region is the largest and probably the oldest forest area in the world; it also ascends the slopes of the Andes until it merges with subtropical and temperate rain forest. On its southern border it merges with the woodlands of Mato Grosso, with galleries of its trees extending along the rivers. Consisting of enormous trees, some exceeding a height of 300 feet, it is composed of an almost incredible number of species growing side by side in the greatest profusion and arranged in different strata. In the region of Manaus, 1,652 plants belonging to 107 species in 37 different families were found on about 1,900 square feet. Amazonian trees number about 2,500 species.

The Amazon rain forest

The forests of the *igapós* (swamps, where the ground is inundated or very marshy throughout the year) cover the lowlands. Characteristic trees are, among others, jacareúbas (*Calophyllum brasiliense*), which is a tall tree with hard reddish-brown wood used for heavy construction, araparis (*Macrolobium acaciaefolium*), abiuranas (*Lucuma* species); piranheiras (*Piranhea trifoliata*) and louros-do-igapo (*Nectandra amazonum*). Undergrowth is dense.

The *várzea* regions are those that are inundated only at the annual flood. Trees are higher and quite diversified; they include oeiranas (*Alchornea castaneifolia*), a euphorbia (having a milky juice) of the chestnut family, and imbaúbas (*Cecropia* species), a rapid-growing tree of the nettle family with a light wood. Palms and hevea (a kind of rubber plant) grow in these forests. The forests of the nonswampy areas are rich in hardwoods, of which acapu (a tree with dark-chocolate-brown wood), pau-amarelo (*Euxylophora paraensis*), pau-santo (*Zollernia paraensis*), massaranduba (a Brazilian tree with light-reddish-brown wood that yields a milky juice), jarana (a tree with hard, heavy, durable wood), and matamatá (a tree with hard, heavy, durable wood, used for pilings) are the best known. Caucho (a wild rubber tree) and the castanheiro (*Bertholletia excelsa*) are characteristic of these forests where spiny palms cover the ground.

Epiphytes (nonparasitic plants that grow on other plants, deriving moisture and nutrients from rain and air) are very numerous, mostly Bromeliaceae (a species having leaves that are often spiny, and spiny flowers), orchids, and ferns. Lianas abound, particularly in drier forests.

Tropical deciduous forests. These forests, dominated by trees of moderate height, notably of leguminous species, are found widely dispersed through northern South America, where the climate is characterized by a prolonged dry season, notably in Venezuela, Colombia, and in the Brazilian Plateau.

Caatinga. Caatinga (white forest) is stunted and somewhat sparse. Thorny vegetation and trees that are leafless for long periods, and able to resist drought, are characteristic of the dry interior of northeastern Brazil, particularly in the basin of the São Francisco River. Dominant species are leguminous trees, particularly catingueiras (*Caesalpinia*), juremas (*Mimosa*), and joazeiros (*Zizyphus joaseiro*), Euphorbiaceae (a family of herbs, shrubs, or trees with usually milky and often poisonous juice), crotons (herbs and shrubs of the spurge family), and Bombacaceae (a family of tropical trees with palmate leaves and large, dry, or fleshy fruit). Undergrowth consists of thickets, bromeliads (plants with basal, often spiny leaves), and innumerable cacti, among which is the xiquexique (*Cereus gounellei*), the complicated intertwinings of which cover the soil.

South Brazilian forests. These parklike forests, sometimes very dense but interspersed with savanna, occupy vast expanses from the border of the Amazonian rain forest to the marshes of the Upper Paraguay River. The typical landscape is a grassland strewn with smaller trees. In effect, it includes a mosaic of associations, from hygrophilous (living or growing in moist places) to xerophilous (adapted to life in the presence of minimal amounts of water) forests and even desert.

Notable is the Paraná pine forest (*Araucaria angustifolia*), which covers a vast area between the Río Paraná and the Atlantic Ocean, stretching from Curitiba, Brazil, to northern Argentina. These trees dominate a dense forest of numerous species including hardwoods, Podocarpus, and the yerba maté (*Ilex paraguariensis*), or South American holly.

Xerophytic associations. Thickets of small trees and shrubs, often very thorny, among which prosopis (a genus of tropical or subtropical branching shrubs or trees), acacia, and mimosa predominate, cover regions that experience alternate dry and relatively wet seasons; these regions particularly include coastal Venezuela, northeastern Colombia, southwestern Ecuador, and northern Peru.

In Peru this association merges with the desert, which extends along the coast to northern Chile, having a width of from 50 to 100 miles. Only a few shrubs and some terrestrial (as distinct from epiphytic) bromeliads grow in this area, which becomes greener only in the Andean foothills, where cacti and other xerophytic plants are to be found, and along the valleys of rivers flowing down from the Andes. In some areas, called *lomas* by the Peruvians, winter mists bring some humidity, and a specialized type of vegetation, consisting of annual and bulbous plants, grows for a short period.

Subantarctic beech forests. Temperate rain forests—similar to those found in British Columbia and in the northwestern United States—grow in southern Chile at low and moderate altitudes, thanks to very high rainfall. The most typical trees belong to the genus *Nothofagus* (timber trees found in the cooler parts of the Southern Hemisphere), the northern species of which are evergreen, whereas the southern species are deciduous. Various conifers, notably the larch and araucarias, mingle with the leafy trees. A dense undergrowth of shrubs, lianas, bamboos, ferns, mosses, and epiphytes grow in the northern districts but disappear toward latitude 49° S. In southern Chilean Patagonia, forests consist of twisted, creeping trees merging into a kind of heath.

The temperate rain forests

Savannas. True savannas are found mostly in Venezuela where the plains known as the Llanos of the Orinoco Basin are covered with tufts of grasses and sedges. Trees and some palm trees are fairly abundant, especially along the rivers.

Pampas. The flat plains called Pampas, which constitute the greater part of eastern Argentina, are covered with grasses. The Pampas were originally covered with trees, but man has acclimatized various species, including pines, eucalyptus, oaks, and poplars.

To the south the Pampas merge with the Patagonian steppe, where grasses are mixed with rosette and cushion plants; the vegetation becomes much poorer as one proceeds south.

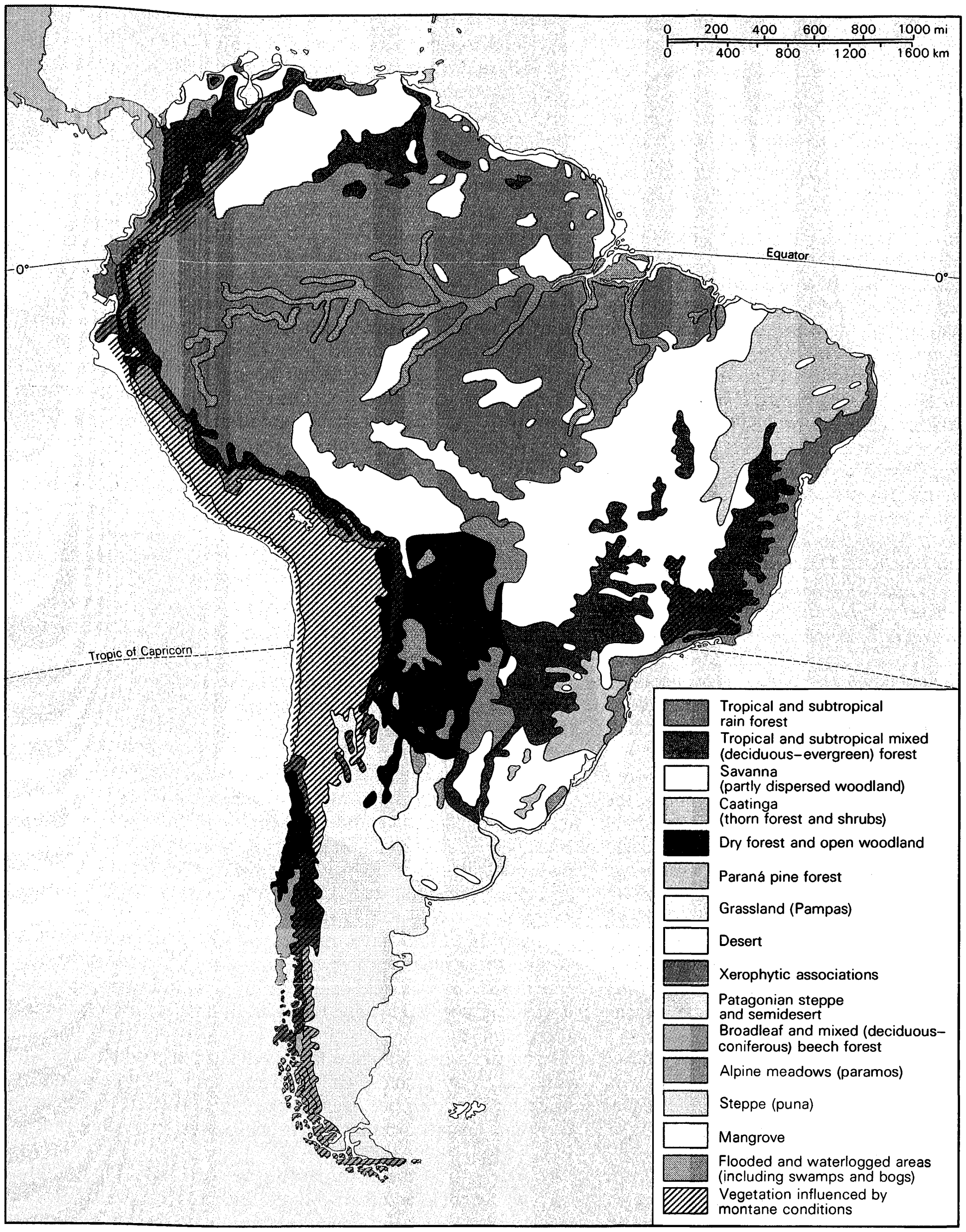
Mountain vegetation. In the high Andes, a temperate zone extends from the upper limit of the subtropical rain forest, around 6,000 feet, up to the timberline, which occurs at an altitude of between 10,500 and 12,500 feet. This zone is very humid on the Amazonian side, where both the epiphytes and the undergrowth are dense.

The upper zones have a peculiar vegetation that reaches the snow line. In the wet northern Andes in Colombia and Ecuador, Alpine meadows called *páramos* consist of grasses and of herbaceous plants, often with bright flowers; these are surmounted by taller plants, especially good-sized plants called *Espeletia,* or frailejones, which are 15 to 20 feet high and are crowned by enormous bouquets of long, lanceolate, hairy leaves.

In the south the *páramos* merge with the *pajonal de puna*—a typical steppe vegetation consisting of rough grasses, between which grow a variety of herbaceous, cushion, and rosette plants, shrubs, and cacti. Vegetation extends to the limit of permanent snow but becomes very scarce at higher altitudes where soil is often barren.

Human activity has transformed the original vegetation cover to a large extent throughout South America, particularly in forested areas. The forests of eastern Brazil were ravaged in the process of clearing the ground for sugarcane plantations as well as to grow other crops. The forests of *Araucaria* (a genus of tall trees of the pine

Destruction of the forest cover


0 200 400 600 800 1000 mi
0 400 800 1200 1600 km
Equator
0°
0°
Tropic of Capricorn
Tropical and subtropical rain forest
Tropical and subtropical mixed (deciduous–evergreen) forest
Savanna (partly dispersed woodland)
Caatinga (thorn forest and shrubs)
Dry forest and open woodland
Paraná pine forest
Grassland (Pampas)
Desert
Xerophytic associations
Patagonian steppe and semidesert
Broadleaf and mixed (deciduous–coniferous) beech forest
Alpine meadows (paramos)
Steppe (puna)
Mangrove
Flooded and waterlogged areas (including swamps and bogs)
Vegetation influenced by montane conditions

Vegetation zones of South America.

family) in the southern states of Brazil have been rapidly vanishing. The slopes of the Andes have also been severely deforested. New roads have been cut through the Amazon region, which is no longer the enormous inviolate mass it once was. In Patagonia the practice of burning to convert forest into pasture steadily increases. Animal herders have destroyed grasslands through overgrazing, notably in the Venezuelan Llanos and in the high Andes. The process of the destruction of habitats continues to accelerate alarmingly throughout the continent, where previously primeval conditions had remained stable for centuries.

Animal life. South American animal life is particularly rich and well diversified as a result of the wide range of habitats. A number of animal groups well distributed over the rest of the world are nevertheless missing, because of South America's lengthy isolation. Many animals belong to exclusive groups, and even at the family level the percentage of endemic forms is very high. Speciation has reached a much higher degree when compared to other parts of the world—a phenomenon that makes attempts to establish systematic classification difficult in South America. Neotropical animal life nevertheless evidently consists of several well-defined elements.

Freshwater fishes are very numerous, numbering about 2,700 species, though deriving from only a few ancestral groups. Amazonian fishes may approach 1,500 species in number. Characins (800 species), which are tropical fishes with deep scaly bodies and strong jaws; gymnotidae, South American cyprinoid fishes that include the electric eel; catfishes; cyprinodonts, a large family of small scaly-headed soft-finned fishes; and cichlids (a family that consists chiefly of fishes that have a resemblance to sunfish) are the dominant groups. Amphibians are well represented by caecilians (small wormlike, burrowing amphibians), salamanders, toads, and a number of varieties of frogs, including clawed frogs, the most aquatic of all; the tree frogs, arboreal amphibians, are particularly abundant through the Amazon Basin, and are very different from their African and Asian counterparts, although the frog faunas of Australia and South America are often surprisingly alike. Reptiles include a great variety of turtles and tortoises, crocodiles, caimans, (endemic crocodilians), geckos, many iguanas, teiids (a family of mostly tropical American lizards), amphisbaenids (a genus of harmless, limbless lizards) and many snakes, including boas, colubrids (a very large family of nonvenomous snakes), coral snakes, and vipers. Birds are represented by 89 families and about 2,700 species—a much higher figure than in Africa or Asia, which justifies the application of the name bird continent to South America. Twenty-five families

Savanna vegetation in the Guiana Highlands of eastern Venezuela.
Jacques Jangoux

are endemic to the neotropical region, including rheas (large, tall, flightless birds that resemble ostriches), curassows (large arboreal birds distantly related to the domestic fowl), hoatzins (a brownish crested bird, having claws on the digits of the wing when young), oil-birds, motmots (bright-coloured birds related to kingfishers), jacamars (small, bright-coloured, clamatorial birds), toucans, manakins, cotingas (related to manakins), and many passerine birds. Hummingbirds are differentiated in the neotropical region, where parrots, pigeons, cuckoos, tyrants (a kind of flycatcher), woodhewers, and orioles are among the dominant groups. Remarkably, the nonpasserine birds are more numerous in proportion to passerine birds than in other parts of the world.

Mammals include types that immigrated to the continent before its complete isolation in the early Tertiary Period. Among these are marsupials (pouched animals), sloths, anteaters, and armadillos; other kinds, now extinct, persisted until the Pleistocene Epoch. More recent elements include hystricomorph rodents (a suborder including porcupines and guinea pigs) and monkeys among the oldest arrivals, and tapirs, deer, carnivores, and bats among the newer arrivals.

Animals are distributed according to the pattern of vegetation zones and several well-defined communities can be distinguished. They include regions as diverse as the Amazonian forests and the high Andes.

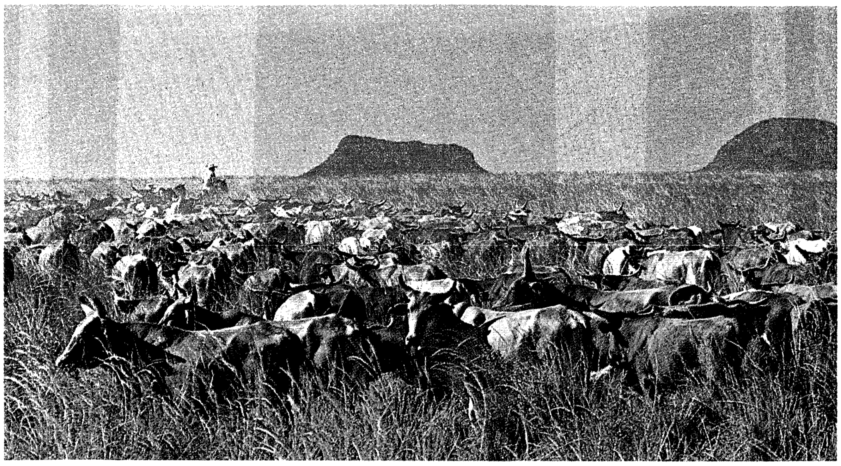
The Amazonian and Guianan forests. The most diverse community is found in the Amazonian and Guianan forests where abundance of water and trees makes life easy. Rivers are the realm of large numbers of invertebrates and fishes, such as pacu (*Metynnis*), which is a big brownish flat fish, the meat of which is highly appreciated; coumarou (*Curimato*), which is a vegetarian fish without teeth resembling the marine mullet; electric eel (*Electrophorus electricus*); arapaimas, or pirarucu, which attain a length of 15 feet and a weight of 200 pounds; and piranhas that are provided with teeth so sharp that they can cut through flesh like a razor, as well as a wealth of small fishes, many of which are vividly coloured. The manatees (a chiefly tropical, aquatic, herbivorous animal with a broad tail) and the inia, a primitive dolphin-like cetacean (an order of completely aquatic mammals), frequent the larger rivers of the region. Arraus (*Podocnemis expansa*), a water turtle, is still extraordinarily numerous. There are large populations of crocodiles and caimans inhabiting the main waterways.

Arboreal forms of the Amazonia forests

Amazonian forests constitute an environment to which most animals responded by becoming arboreal. Tree frogs can move across the surface of the leaves thanks to adhesive pads; lizards have very elongated fingers; monkeys, sarigues (a close relative of the opossum), and kinkajous (a kind of nocturnal, arboreal, carnivorous mammal) have prehensile tails. Birds are numerous and,

(Left) Harrison Forman, (right) Jacques Jangoux

Amazon and Guianan rain forests.
(Left) Traffic on the Amazon River in Brazil. (Right) Dense tropical vegetation in the southern Guiana Highlands region of Venezuela.



Range cattle grazing in shoulder-high grasses that grow in the Pampas of eastern Argentina.
E. Hartmann—Magnum

because of the enormous amount and variety of available foods, well diversified. Antbirds, tyrants, cotingas, tangaras (brilliantly coloured birds related to the finches), hummingbirds, toucans, woodpeckers, barbets (loud-voiced tropical birds closely related to honey guides), parrots, and tinamous (quail-like terrestrial birds) are the dominant groups. Many of them never leave the forest canopy, where they display their brilliant colours, which contrast with the more modest plumage of those birds that live in the undergrowth. Mammals are represented by a number of terrestrial or half-aquatic species, such as very small deer; some large rodents (including the agouti, paca, and capybara); tapirs; and carnivores (including the jaguar). Monkeys range from pygmy marmosets to larger douroucoulis (small, round-headed, stocky-bodied, bushy-tailed monkeys), woolly monkeys, spider monkeys, and howler monkeys. Bats are also numerous, including fruit bats. Sloths feed on the leaves of certain trees among whose branches they remain much of their lives. The predators are represented by carnivorous mammals, a series of snakes—including anacondas and boas—and large raptors (birds of prey), such as the great harpy eagle, the most powerful bird of prey to be found in the world. Insects, including a great variety of butterflies and ants, are innumerable.

The Brazilian Plateaus. The Brazilian Plateaus have an impoverished animal life, from which species that are strictly adapted to the dense forest are excluded. The plains of Uruguay and the Gran Chaco have a varied animal life that includes some particular species, such as the maned wolf. The marshes are inhabited by a wealth of waterfowl, as well as by a species of lungfish (*Lepidosiren paradoxa*) that is related to its African and Australian counterparts.

imal
e of the
mpas

The Argentinian Pampas. The Pampas of Argentina are inhabited by a very limited number of indigenous animals such as rheas, a series of smaller birds, including the popular ovenbird (*Furnarius rufus*), the name of which comes from its globe-shaped nest made of mud, and mammals such as the mara (*Dolichotis patagona*), a long-legged, long-eared rodent; the plains vizcachas (*Lagostomus*), a burrowing rodent related to the chinchilla; guanacos (*Lama guanacoe*), a South American mammal related to the camel but resembling a deer; and Pampas deer (*Blastoceros campestris*). The restricted number of the larger herbivorous mammals is quite remarkable and illustrates the scarcity of recent mammalian types in the neotropical region.

The southern Chilean forests. The forests of southern Chile are inhabited by a specialized animal life, with a high percentage of endemic species. Parakeets and hummingbirds are found as far south as Tierra del Fuego. A marsupial, the rincolesta of Chiloé (*Rhyncholestes raphanurus*), is one of the most primitive mammals still in existence.

The high Andes. The high Andes have been settled by a very impoverished animal life, which had to adapt to the harsh and cold environment, to the scanty vegetation, and to low oxygen pressure. The great quantity of lakes in the region has attracted numerous aquatic birds, including flamingoes, which nest up to 16,000 feet in northern Chile, and amphibians such as the giant toads of Lake Titicaca, which spend their entire life in water. Mammals are represented by guanacos, vicuñas (wild ruminants related to the domesticated llama), deer, and numerous rodents, including vizcachas, chinchillas, and guinea pigs. Predatory species include foxes, pumas, and many birds of prey, notable among which is the condor, the giant of living birds, with a wingspan of more than 10 feet.

The arid west coast. The arid coast of Peru and northern Chile is inhabited by a terrestrial animal life impoverished when compared to its rich offshore marine life. The zones washed by the Peru Current, cold and rich in salts, are swarming with life, from plankton to fishes, including an anchovy (*Engraulis ringens*); these small forms of life provide food for higher levels of the marine community, represented, for example, by sea lions and birds, many of which are endemic to the area. Bird life includes a penguin, many species of gulls and terns, shearwaters, petrels, cormorants, pelicans, and boobies (a kind of gannet). Three kinds of these birds—the guanay (*Phalacrocorax bougainvillii*), the variegated booby (*Sula variegata*), and the brown pelican (*Pelecanus occidentalis*)—nest by the millions on small islands off the coast, where their droppings accumulate to form guano, a highly prized source of fertilizer.

Protection of wildlife. Overhunting and the destruction of habitats have wrought great harm to South American animals. Until the mid-20th century South America had suffered less from these problems than had the rest of the world, but since that time the situation has changed very rapidly. Many kinds of animals are on the decline, and some even appear to be threatened with extinction. Destruction of the Amazonian forest, which in recent years has reached significant proportions, could be disastrous for its animal life.

National
parks

In order to remedy the situation, a number of nature reserves and national parks have been established throughout the continent. Argentina was indisputably the pioneer in wildlife protection in South America. Nahuel Huapí National Park, the original nucleus of which was established in 1903, preserves a part of the Andes, featuring glaciers, beech forests, and many animals. Two national parks, one on the Argentinian, the other on the Brazilian side, protect the Iguaçu Falls and the surrounding rain forest. Many national parks and reserves, including the Itatiaia National Park, are maintained in Brazil. In Venezuela the Henri Pittier National Park, in the coastal cordillera, is famous for its subtropical forest, bird life, and migrations of birds and butterflies. Other reserves and parks have also been established in Ecuador, Peru, and Chile in order to preserve representative areas of the habitats of the Andean slopes and plateaus. (J.P.D.)

The people

ETHNIC ORIGINS AND MIGRATIONS

Four main components have contributed to the present-day population of South America—the American Indians (Amerindians), who were the pre-Columbian inhabitants; the Iberians (Spanish and Portuguese who conquered and dominated the continent until the beginning of the 19th century); the Africans, imported as slaves by the colonizers; and, finally, the postindependence immigrants from overseas, mostly Europeans.

Indians. During the epoch of European discovery and conquest in the early 16th century, Indian societies existed at three different cultural levels; this to a large extent determined the composition of the population during the colonial period and afterward.

Andean civilization

Andean society, a civilization of archaic type, comparable to Egyptian, Mesopotamian, and other Old World pre-Greek civilizations, existed on the Pacific coast with its focal centre in Peru. It is estimated that agriculture (based on corn [maize]) had been developed in this area from at least 2500 BC, while the beginnings of the formative period of Andean civilization date to the 1st millennium BC. At that time the agricultural village was the basic community; different arts and crafts were also developed, and, toward the end of the period, religious–ceremonial buildings began to appear.

The culture of the Classic Andean period (approximately from AD 1 to 1000) reached a higher refinement in the architectural arts; palaces and elaborate multiroom buildings were constructed, metallurgy (the smelting of copper and gold alloys) was developed, and agricultural techniques, such as terracing and irrigation, were evolved. Urbanization, which is to say the emergence of inhabited cities, as distinct from religious–ceremonial centres, also came into existence during the Classic period. With a relatively high level of political and military organization, these civilizations established several kingdoms. Neither the Classic nor the post-Classic Andean civilizations, however, had a written language. The post-Classic phase, moreover, was marked by further extension and centralization. The Inca Empire of Peru, for example, expanded from its homeland in Cuzco both to the north (to the area occupied by present-day Colombia, with its relatively advanced Chibcha cultures), and to the south (present-day Chile, with its primitive Araucanian tribes). The Incas began their conquests around AD 1200, but the expansion of their empire was greatly accelerated after 1400, until the process was halted by the Spaniards in 1536.

The remaining South American societies—among which lower and intermediate levels could at the same time be distinguished—had not evolved beyond the earlier phases of social development. The lower level occurred in the Guanaco area, inhabited by nomadic hunting tribes. It was located on the present-day territory of Argentina and Uruguay and included the extreme south (Tierra del Fuego and Cape Horn). The intermediate level occurred in the Manioca area, which covered the Amazon Basin to the Atlantic coast, occupying part of the present-day territory of Brazil. Although the intermediate level was mainly represented by hunting people, they also practiced agriculture, using the slash-and-burn technique (*i.e.,* cleaning ground for temporary cultivation by cutting and burning the vegetation). Andean civilization and the peoples dominated or influenced by it included perhaps one-half of South America's indigenous population and also covered the most densely settled areas of the continent. Elsewhere, the population was sparse, and large areas remained deserted.

The contemporary Indian population

The number of Indians at the time of the conquest is uncertain: estimates vary from 8,000,000 to 100,000,000 for North, South, and Central America combined (for the Incas, from 3,000,000 to 32,000,000). Some authorities assign to South America 6,800,000, of which a little less than half belonged to the Inca Empire or were under Inca influence. Other more recent estimates that put South America's preconquest population at about 14,000,000 seem more realistic.

Equally controversial is the origin of South America's Indians. Most anthropologists believe that they arrived in America from Asia, in successive waves, from about 25,000 BC onward—probably across the Bering Strait, which separates the extremities of northeastern Asia and northwestern North America.

Iberians. Until the end of the era of their domination, only the Spanish and Portuguese were admitted to their South American colonies. The rigid exclusion of all other foreigners had but few exceptions, though a small number of other Europeans settled as a result of illegal or tolerated immigration. The Iberian ethnic background itself was, however, highly diversified. Apart from the Romans themselves, many pre-Roman peoples—Visigoths, Jews, Arabs, Berbers, and Moors—contributed to form both the Spanish and Portuguese population. Most of the Spaniards came from Castile and the southern regions. Very little is known about the principal regions from which the Portuguese came. It is estimated that the total number of *licencias* (authorizations to emigrate) granted by Spain was about 150,000 for the whole colonial period, which lasted from the 16th to the 19th century; it is possible that the number of illegal immigrants also come to this number. Of these, no more than two-fifths of the emigrants went to South America.

Africans. A few African servants accompanying their Spanish or Portuguese masters were the first slaves to enter the continent. Larger-scale importation of slaves from Africa developed two or three decades later: the slave trade was authorized by Spain for the first time in 1518, though reliable quantitative information is lacking. An idea of their demographic contribution may be given by some estimates: 4,000,000 for Brazil, and 3,000,000 for all Spanish America, of which only a minority went to the Andean region and an even smaller number to the southern cone (consisting of present-day Argentina, Uruguay, and Chile). Paradoxically, the slave trade was supported by the defenders of the Indians. African slaves were also considered to be more efficient than American Indians, particularly for working on tropical plantations. Most of the slaves imported into South America were from West Africa, including Angola. The slave trade ceased in the early 19th century.

Postindependence overseas immigrants. With the attainment of independence by most of the South American countries in the early 19th century, the legal exclusion of foreigners came to an end. Mass immigration to the continent, however, did not begin until the late 19th century, acquiring momentum in the last three decades of the century and continuing until 1930, when it decreased abruptly. Some 11,000,000 to 12,000,000 people arrived in South America; the great majority of these went to Argentina (more than 50 percent) and Brazil (about 37 percent). Although many later left, the demographic and sociocultural impact of this influx was tremendous in Argentina and (to a lesser extent) in southern Brazil. Immigration to other countries was numerically insignificant (although socioculturally meaningful), except in Uruguay, where because the preexisting population was not numerous, the percentage of the foreign born was high—18 percent in 1908, and even higher in the 19th century. In Argentina this proportion reached nearly one-third of the total, and stayed at that level for many years. In both cases the contribution of post-independence immigration was proportionally much higher than in the United States at the peak of mass immigration.

Sources o immigration

The great majority of the immigrants were Europeans—Italians (forming nearly one-half of the immigrants in Argentina, one-third of those in Brazil, and probably the majority of immigrants in Uruguay), the Spaniards (one-third in Argentina), and the Portuguese (30 percent in Brazil). Other small but socially relevant immigrant streams arrived from central and eastern Europe. This source of immigration became more important after the turn of the century and especially during the 1930s and 1940s, when it included more middle class and educated people, among whom were many Jews and other refugees. After World War II another smaller wave of immigration arrived from Europe, directed mostly to Venezuela and Argentina.

POPULATION AND ECOLOGICAL DISTRIBUTION

The present population. The present population of South America is the result of four centuries of race mixture among these four components—Indians, Iberians, Africans, and overseas immigrants—and their mixed descent. Race mixture began when the first Iberians touched South America. Their own previous traditions and their basic values and attitudes, coupled with other characteristics of their conquest and colonization, facilitated intermixing not only with the Indians but in general among all the various racial and ethnic groups even although its intensity, extent, and frequency varied both among different groups and at different epochs.

Legal marriage between Iberians and Indians was tolerated, often permitted, and even, in some special circumstances, promoted. It was possible—and in certain epochs easy—to recognize mestizo (generally, mixed European and American Indian) children, though often a mestizo was considered automatically illegitimate. With regard to mixing between Europeans and Africans and between Africans and Indians and their offspring, such social permissivity did not exist—a fact that nevertheless failed to prevent generalized race mixture. This prolonged process created a great variety of physical types, resulting in the emergence of a complex terminology to describe them.

The more important types are the mestizo (called caboclo in Brazil); mulatto (European–African), zambo (African–Indian), and cholo (mestizo–Indian). When the postindependence European immigration occurred, other national groups contributed to race mixture. As a result, Argentina and Uruguay completely modified their ethnic composition.

ıe three
ɔlogical
ȝions

Ecological distribution. For these ethnic groups, three regions of ecological distribution may be distinguished—Indo-America, Mestizo-America, and Euro-America.

Indo-America consists of the highlands from north to south on the western side of the continent. It corresponds roughly to the territory of Venezuela, Colombia, Ecuador, Peru, and Bolivia, where—together with a minority of white or "whitish" groups (mostly of Spanish origin)—mestizo and Indian groups are predominant. Mestizo-America is situated in the lowlands of tropical and subtropical South America. Both on the eastern and western coasts, the African influence is strongest; the area is then characterized by European groups (mostly Portuguese in Brazil), many of them with various proportions of black ancestry, as well as by Africans and mulattos. Euro-America, or the temperate zone, consists of southern Brazil, Argentina, Uruguay, and central Chile. There the Europeans are overwhelmingly predominant, and there is little African or Indian influence (with the exception of Chile, where the original Indian contribution was much more significant, although it is not any more visible from the cultural or even the somatic point of view). Geographical, historical, cultural, and socioeconomic factors account for this ecological distribution.

Indo-America. It must be remembered that the Indo-American zone was the site of the higher civilization in South America and was the most densely settled at the time of the conquest. There the social and economic organization of the Spanish empire could be founded on the relatively advanced institutions of the pre-Columbian cultures, and, at the same time, Indian labour could be easily exploited. Although the original Indian population suffered what has been called a demographic disaster (during the first century of Spanish domination the Indian population of this area decreased by at least 30 to 60 percent), a substantial Indian population survived, and race mixture was intense; the Indians subsequently began their demographic recovery at about the end of the 17th century. The new nations in this area did not receive substantial European immigration in the 19th and 20th centuries, so that the European component in the population comes mostly from the colonial period.

Mestizo-America. In the lowlands of Venezuela and Colombia, however, and much more especially in Brazil, African slave labour was important. There the sparse indigenous tribes were much less well adapted to work efficiently; they also resisted enslavement and were decimated by disease and warfare. Indians were therefore replaced by imported Africans, mostly for the purpose of cultivating plantation crops. European–Indian mixture became less frequent, and the African racial and cultural influence became significant, particularly in Brazil where the Negroid (*i.e.,* African and mulatto) population constitutes perhaps 30 to 40 percent of the total. The other two South American countries with a substantial proportion of blacks (or with black components) are Colombia (no more than 30 percent black and mulatto) and Venezuela, where the black or Negroid proportion of the population is probably smaller.

Euro-America. In the temperate zone, the indigenous tribes, already scarce, were destroyed or reduced to an insignificant size, and the mestizo population was mostly overwhelmed by the avalanche of European immigration.

Another factor in the ecological distribution of ethnic groups in South America is the degree of urbanization. People of European origins, who were white or "whitish," tended to be more urban than rural and to live in larger cities rather than smaller ones. This pattern, however, was being blurred in the middle and later years of the 20th century by internal migrations. In effect, from World War II onward, these internal population movements, as well as some inter-American migrations (mostly between contiguous states and mostly from the less to the more developed states) represented a new contribution to the South American melting pot.

Racial
criteria

Sociocultural factors. As a general comment on the foregoing discussion, it must be remembered that racial criteria in all of Latin America are sociocultural (that is, generally based on language, clothing, and behaviour patterns) and not genealogical (that is, based on descent, as in the United States, South Africa, and elsewhere). In Latin America, "passing" may be no more than a question of buying a European dress and moving elsewhere. This does not mean lack of prejudice, discrimination, and exploitation—which indeed characterize the whole history of South America—but it involves significant differences for the life chances of the individuals. "Race" classification becomes in any case very subjective, being based also on the particular social status of the person. (The higher the social class, the "whiter" the person is considered to be.)

Linguistic patterns. The linguistic diversity and multiplicity of South America is said to be unmatched anywhere in the world. Thousands of languages and dialects have been cataloged, including all those existing since the conquest. Classification systems vary a great deal—from more than 100 "linguistic families" and many unrelated languages at one extreme to an extremely simplified scheme including four main families and a few subfamilies at the other.

There is also considerable disagreement on the composition of these "stocks" and how many languages should be classified. Most are now extinct, either because the tribes who spoke them have disappeared or because of acculturation into a European language or, in some instances, into another indigenous tongue. Relations between the South American languages and other New and Old World languages are still highly controversial. There are limited evidences of ties between some southern Colombian languages and one large North American linguistic family. Hypotheses have been advanced about connections between South American Indian languages of the Pacific coast and the Malayo-Polynesian languages, but linguistic and cultural evidences fail to support them.

The
surviving
Indian
languages

The survival of Indian languages follows the same pattern as the ecological distribution of the population and is equally determined by the level of civilization. Quechua and Aymara (both languages of the Andean highlands, the former being spoken originally by the Cuzco Incas) are the most diffused languages; they are used by an estimated 60 percent of the population in Bolivia, 40 percent in Peru, and a sizable proportion in Colombia and Ecuador (where both Quechua and Jíbaro are spoken). Another important tongue is Guaraní, which, together with Spanish, is diffused most of all in Paraguay, where it is also spoken by the non-Indian population. Contrasting with this, the languages of the lower level Indian cultures have in general disappeared or remain only as scanty remnants. African dialects have been wiped out, although they did exercise some influence on Portuguese and Spanish—as also was the case with both existing and extinct Indian languages. Such countries as Argentina and Chile, therefore, have only a small segment of the population still speaking an indigenous language. A higher proportion speak such languages in Brazil, and perhaps 10 percent do so in Venezuela.

Sociological changes. All or most legal discrimination against the Indians and other ethnic sectors of the population was nominally abolished at the time of independence, or during the 19th century. The real conditions of the Indians (and to a certain extent of the Africans after the abolition of slavery in Brazil in 1888), however, remained the same or became worse, since, on one hand, liberal legislation tended to eliminate all communal property and the legal existence of the Indian communities, while, on the other, various forms of exploitation continued unchanged. These de facto conditions were also reinforced by 19th-century biological doctrines, which held Indians and Africans to be inferior races.

In the 20th century, however, a complete change in intellectual attitudes resulted in a series of public and private initiatives toward ameliorating the conditions of these groups. Illustrations of this change in attitude are to be seen in the creation by Latin American governments of the Instituto Indigenista Americano (1940); in the (formal) restoration of the "legality" of the Indian communities, in Peru (1919); and in the many projects of agrarian reform, as well as in other legal and administrative innovations discussed and sometimes approved in different countries, even though these are seldom seriously enforced. Institutions such as the Serviço de Proteção ao Indio (Protective Service for the Indians) in Brazil, as well as missions, centres, and the like were established with the purpose of providing educational, sanitary, and other social services to the Indians. Special areas were established as reservations, particularly for tribal Indians, although these received little attention of any kind from the governments. At present, two types of situation may be distinguished—that of the Indian tribes and that of the Indian communities.

The status of Indian tribes

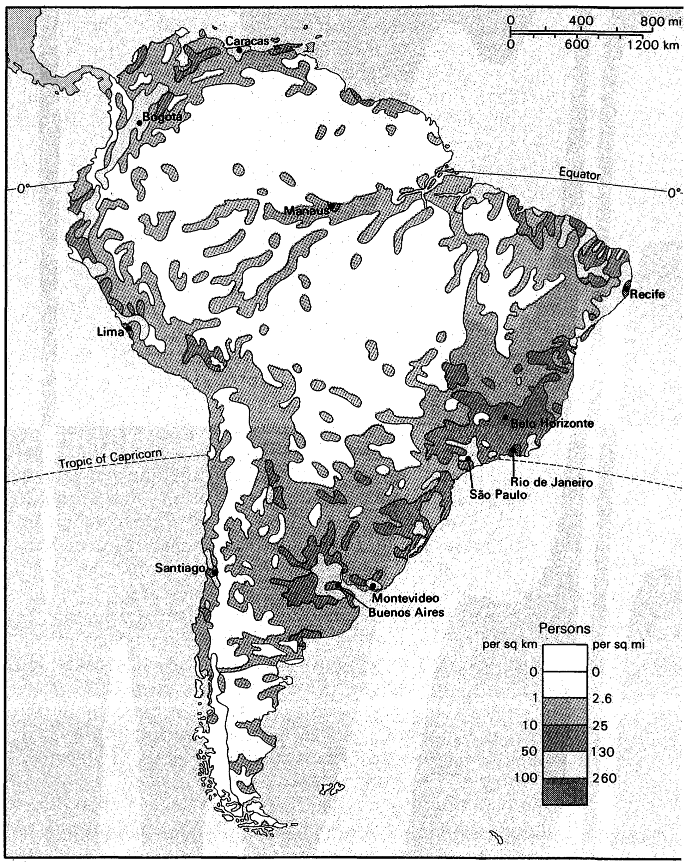
Although the Indian tribes should nominally be under the "protective" regulations existing in several Latin American countries, they must nevertheless be considered as totally or mostly marginal to the life of the national societies. Statistically, they represent an insignificant proportion of the total population (between 30,000 and 200,000 where they exist at all). They can be found in the Gran Chaco and Amazon regions and also in the selva (tropical rain forest) and mountain regions of Peru and Bolivia. Some Indian tribes are also located in Chile, Colombia, and Venezuela. These groups are disappearing—in most cases either through physical extinction, caused by poor living conditions, illness, or violent death, or through absorption into some sector of the national society.

The Indian communities (as distinct from tribal Indians) are related to the pre-Conquest communities belonging to the higher cultures but have been deeply modified by colonial domination and Hispanic culture. They can be found mostly in Peru but also exist in smaller proportions in the other areas influenced or formerly dominated by the Inca, such as parts of Bolivia, Colombia, or even Venezuela. These communities—some of which in Peru were recognized by the state—are nominally under the formal national bureaucracy but sometimes maintain their traditional Spanish-Indian authorities, such as alcaldes (mayors) or *regidores* (council members). Those that maintain more strongly their identity and traditional nature have been characterized as "corporated" (*i.e.,* incorporated) communities (in the sociological, not the legal, sense of the term). Often, however, the communities are (or were) de facto tributaries of the haciendas—the large landed estates that until recently exercised quasifeudal dominion over the Indians. This situation, however, has been in the process of rapid transformation, particularly since the Bolivian revolution (1952) and since the social movements and political changes that have taken place in Peru. (G.Ge.)

DEMOGRAPHIC PATTERNS

South America is at the peak of a phase of demographic transition, a modernizing process that includes three main stages. The first of these is the traditional stage, in which the population balance is maintained by both high birth and death rates; second is the transitional stage, in which decreasing mortality combines with a stable high fertility (or else with a delayed or slow decline, or both) to generate a high growth; finally is a modern stage, in which low mortality and low fertility produce a small demographic increase. Areas that developed earliest underwent the same process, but with two differences: they experienced first a lower fertility rate in the traditional stage and, second, a slower or delayed decline, or both, in the mortality rate.

It must nevertheless be remembered that some 60,000,000 Europeans emigrated overseas. Natality in Europe's preindustrial phase rarely exceeded 30 to 35 per 1,000; in South America it has been 25 to 40, with a number of exceptions. Argentina and Uruguay have been at the modern stage since the 1930s, and Chile has had declining fertility since the 1920s. Mortality decline in Europe often took 100 to 150 years. In 1900 the South American average life expectancy at birth was 25 to 30 years (not including Argentina and Uruguay). By the late 20th century it was estimated at a little more than 60 years. The great change occurred in all countries (except the two "first comers," Argentina and Uruguay) between 1920 and 1940 and is continuing. Mortality declined spectacularly in 40 or 50 years, and, in Chile, in 20.



Population density of South America.

The fertility rate. The population explosion is the result of a traditional type of fertility rate (among the highest ever known), combined with a quasimodern mortality rate. The reason for this discrepancy in comparison with the rest of the Western world lies in the fact that while mortality can be reduced through the relatively fast diffusion of sanitary and medical innovations a spontaneous fertility decline requires complex social transformations (in social stratification, and in urbanization, mobility, family, and life-style) correlated with psychological changes. These demand much time. It is also possible that cultural values peculiar to the Latin and Indian societies produce additional obstacles to spontaneous limitation of fertility. Government-sponsored policies to induce family planning have met two ideological obstacles: the attitude of the Roman Catholic Church and that of the political parties toward birth control. Though resistance has declined in the church, politicians have remained indifferent or hostile to birth control. The fact that the region is underpopulated disguises the enormous strain imposed on the economy by too rapid a rate of demographic growth.

ɔstacles
family
ınning

Spontaneous changes are nonetheless taking place. As in Europe, birth control is initiated by the urban population and by the middle classes. In South American cities the average number of live births per married woman becomes lower with increased education. Because of the large rural population and the limited size of the educated group, however, the overall impact of the differentials is not yet reflected in national fertility rates. The process, if left to its own spontaneous evolution, may require several decades to take place, while mortality is meanwhile expected to decline further in various countries. If fertility is not curbed or economic development is not accelerated, the operation of Malthusian laws (that when population outstrips its means of subsistence further growth will be checked by disease, famine, or war) would seem to ensure that the mortality decline would be interrupted or reversed. In this case the expected growth would be reduced.

Effects of rapid population increase. Rapid population increase has important demographic and social effects. Two examples are especially illuminating.

At the peak of population growth, the proportion of children tends to be very high, while in the third stage it is low. In South America the proportion of the population under 15 years is relatively high. As a consequence, the working-age group is greatly reduced. This high ratio creates a heavier burden for the working group, while the economy is not able to raise the productivity needed to compensate for it.

pulation
ɔwth in
ꞔ cities

Another crucial consequence is the so-called urban explosion. Argentina and Uruguay are among the most urbanized countries in the world, but their urban growth has been due to mass foreign immigration. The dramatic increase in urban concentration began approximately in the 1930s. In all of Latin America, urban centres with more than 10,000 inhabitants increased from one-fourth in 1950 to an estimated one-half 20 years later.

South America is one of the most urbanized regions in the world, following the industrially advanced areas. Although the rate of growth in larger cities has decreased since 1950, the urban population continues to be concentrated in the bigger cities: a large proportion of the urban, and in some cases of the total, population lives in a single urban centre. This situation prevails in Uruguay, Argentina, Chile, Paraguay, and Peru.

Demographic growth has both a direct and an indirect influence on the urban population explosion. Contrary to widespread opinions the principal component of urban growth on the continent is natural increase, not internal migration. This is the case in Venezuela, Chile, and other countries. Internal migration, nevertheless, remains important in South America; the demographic influence is indirect and operates in conjunction with other factors—such as unfavourable conditions in the countryside, including the concentration of land ownership in a few hands and the rejection by the rural population of the resulting extremely low standard of living.

High urbanization is coupled with a very low population density in South America as a whole, with variations among individual countries that ranged from a very low density in such countries as Suriname and Guyana to an extremely high density in Ecuador and Colombia. The averages disguise the existence of sharp internal inequalities in density. High density is found in South America only in urbanized areas—that is, in large cities or metropolitan areas and their immediate hinterland. South America is increasingly coming to resemble a vast desert ocean, dotted with a few scattered densely populated islands. Though geographical and physical environmental factors play a role, high concentration is largely the consequence of uneven internal development caused by international and domestic factors. The trend in South America continues to be toward an even higher imbalance, although in several countries the state has begun to intervene in an attempt to achieve a more rational use of the national territory. (Ed.)

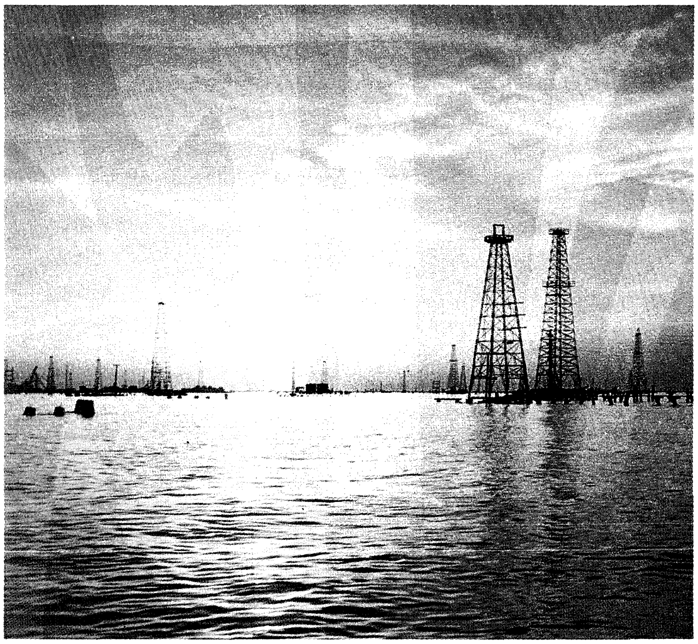
The economy

RESOURCES

Mineral resources. *Mineral fuels.* South America is relatively poor in coal, which is found in Colombia, as well as in Brazil, Argentina, Peru, Venezuela, and Chile (where lower-grade fuels such as lignite and peat are also available). Petroleum is present in great quantities in sedimentary terrains around Lake Maracaibo in Venezuela, and in lesser quantities in the coastal plains of Brazil, Ecuador, and Peru, along the eastern Andes of Bolivia, and in southern Patagonia. The greatest and most important of these mineral fuel reserves are located in Venezuela, one of the largest oil producers in the world. Other reserves are to be found in Argentina, Colombia, Peru, Brazil, Chile, Ecuador, and Bolivia. Natural gas also is found in Venezuela, Argentina, Colombia, and Chile. Uranium can be found in significant quantities in the form of various ores in Argentina, Brazil, and Chile, as well as in Colombia.

Oil
reserves

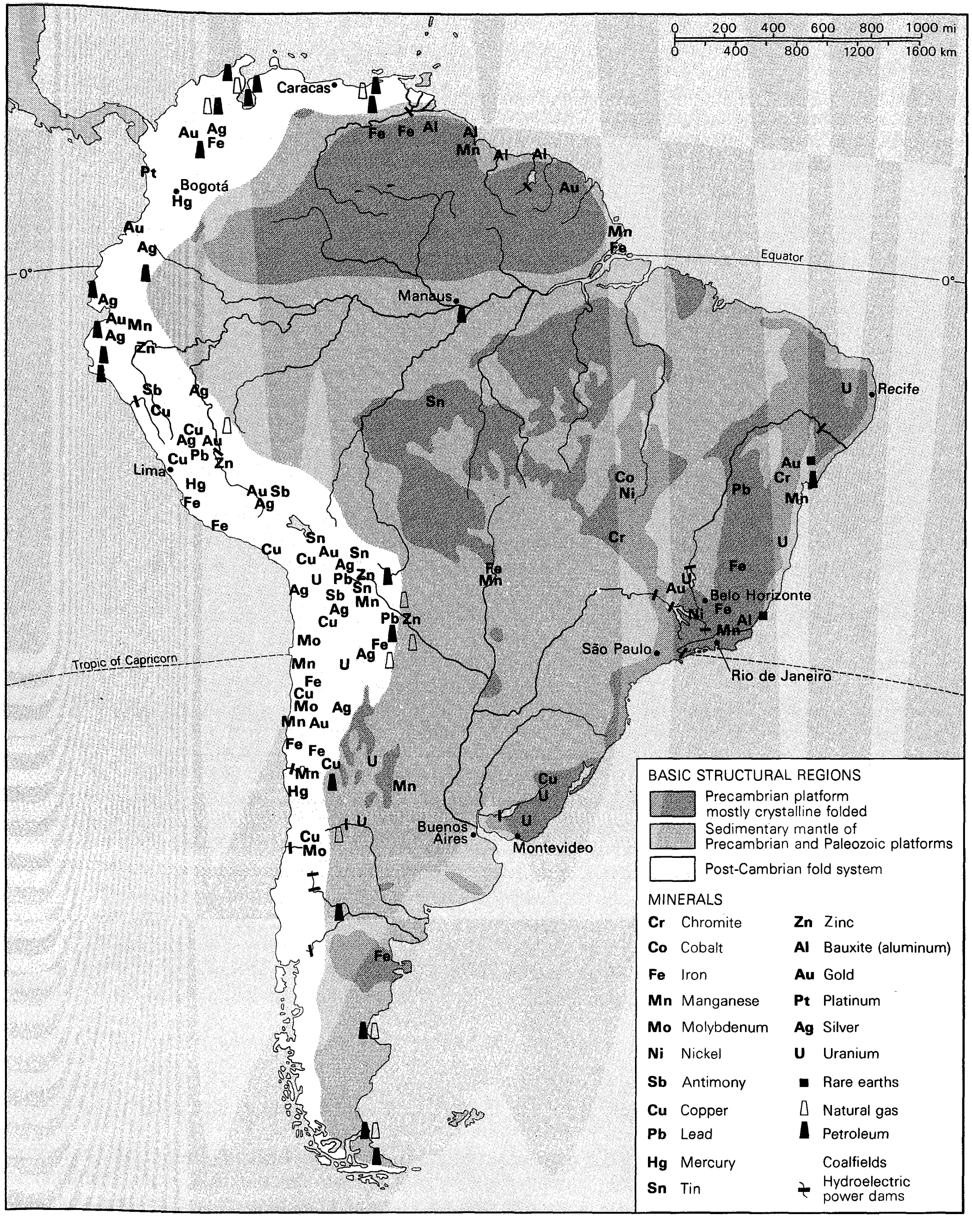
Shostal—EB Inc.



Oil derricks on Lake Maracaibo, Venezuela, centre of the petroleum industry in Latin America.

Metallic deposits. Iron ore is very abundant; South American reserves represent one-fifth of the world total. The most important beds are found in Brazil and Venezuela, representing a fabulous wealth for these countries. In Brazil, the states of Minas Gerais, Pará, and Mato Grosso do Sul contain lodes of magnetite and hematite with 50 to 65 percent iron, which represents the vast majority of the South American reserves. In Venezuela, at the base of the Guiana Highlands, the Sierra de Imataca, particularly Cerro Bolívar and El Pao mountains, have reserves of ore containing a high percentage of iron. Other beds of this type are to be found in Mutún, Bolivia and Chile. Oolitic iron ore (*i.e.,* consisting of small round grains cemented together) is located in Argentina (Sierra Grande) and in Colombia (Paz del Río). Other iron ore deposits are found in Peru (Marcona) and Chile (along a belt from Taltal, Antofagasta, to Ovalle, Coquimbo). Lateritic deposits of ferrous hydroxides are widespread, mainly in Colombia, Brazil, and Argentina.

Among ferroalloys, manganese is found in sedimentary forms in Brazil—mostly in Amapá, Bahia, Minas Gerais,



Basic structural regions and principal mineral and hydroelectric sites of South America.

and Mato Grosso do Sul—in Bolivia, and in much lesser quantities in Chile, Ecuador, and Uruguay. South America is remarkably deficient in nickel, chromite, and cobalt, though small quantities of all are found in Peru, Brazil, and Argentina. Of the non-Communist nations, Chile is third, after the United States and Canada, in the size of its reserves of molybdenum.

Nonferrous metals

Nonferrous metals are particularly abundant in South America, where copper reserves represent more than one-fourth of the world's known reserves; almost all of the South American reserves are found in Chile and Peru. In Chile, the Chuquicamata deposits contain the largest amount of copper known in the world, with ores containing 2.5 percent copper; El Teniente also has reserves. In Peru the most important deposits are found in the Central Andes (at Yauricocha, Morococha, Casapalca, Cerro de Pasco, and Huarón), as well as in the southern Andes, where the Toquepala mine opened in 1960. Copper is also found in Bolivia and in Brazil (at Caraíba, in the state of Bahia).

Tin, in the form of sedimentary deposits and in veins, is found in the eastern Andes of Bolivia in a narrow belt that is approximately 500 miles long by 95 miles wide, mainly running from Oruro to Potosí. Lead and zinc, two metals which are frequently associated in nature, are dispersed among many countries but are mostly found in Peru, as well as in Argentina, Brazil, and Bolivia. Aluminum, in the form of bauxite, is exploited mostly in Suriname. Reserves are found elsewhere, mostly in Guyana, French Guiana, Brazil, and Venezuela. Mercury is found in small quantities in Chile, Colombia, and Peru; antimony is found in Bolivia and in lesser quantities in Peru.

Among the precious metals that are to be found in South America, gold represents only a small percentage of world production, although it is found throughout the continent in varying quantities. In Brazil the states of Minas Gerais, Goiás, and Mato Grosso were well known in colonial times for their gold, but production diminished as a result of the exhaustion of the lodes. There are also gold-mining operations in Colombia, Brazil, Peru, and Chile. Silver is found mainly in Peru, but is also found to a lesser extent in Bolivia, Chile, Argentina, Colombia, and Ecuador. Colombia has the largest reserve of platinum.

Precious stones

Many parts of South America, mainly Brazil, have been known for their gems. The ancient bedrocks are rich in precious stones, which are found in veins of pegmatite (a coarse granite) or are washed out by streams. Diamonds are spread over a good part of the Brazilian Plateau, notably in the region of Diamantina. Some are suitable only for industrial use, but others are of considerable value as gems. Brazil, however, contributes only a small percentage of world production.

Other precious or semiprecious stones abound in the same region, notably topazes, tourmalines, beryls, aquamarines, chrysoberyls, garnets, and sapphires, as well as quartz of sufficiently high grade to be used in electronics (the only source apart from those in Communist countries). Colombia is famous for its emeralds, found in the Muzo mines.

Various mineral elements used in industry—such as beryllium, columbium (niobium), tantalum, thorium, lithium, rare earths, and mica—are found in Brazil and Argentina.

Nonmetallic deposits. South America has neither a significant output nor important reserves of the principal fertilizers—phosphates and potash. In the Atacama Desert, located in northern Chile, there are important deposits of nitrates to be found in the form of caliche, a grey rocklike substance with a high content of sodium chloride, sodium nitrate, and iodine salts. These deposits were formerly of some importance, but synthetic fertilizers have tended to supersede them; they continue, however, to be used as a source of iodine. Many various saline deposits found throughout the continent await investigation as potential sources of salts.

Biological resources. Biological resources are abundant and much diversified, although they are not evenly distributed throughout the continent. Southeastern Brazil, Uruguay, the Argentinian Pampa, and central Chile are areas that are suitable for wide-scale agriculture.

Botanical resources. The plant diversity of the South American continent is enormous; many of the South American plants are useful to humans. In pre-Columbian times, Indian cultures domesticated numerous plants, some of which now play a major role in the world's food supplies.

Though fewer than 200 species constitute the bulk of the neotropical forests, their mixed character is a major obstacle to large-scale exploitation of timber. Many species are used as cabinet woods, including the highly prized mahogany (*Swietenia macrophylla*) from Venezuela, Brazil, Peru, and Bolivia, and several leguminous species such as rosewood (*Dalbergia nigra*); other species are exploited as general utility woods and are mainly used domestically, often as fuel. Balsa (*Ochroma* species) is the lightest and fastest growing wood in the world. Other species produce tanning materials (such as the bark of the quebracho tree), various drugs, or edible products.

Forests are extensive, covering about 2,340,000,000 acres (946,000,000 hectares), an area that represents more than one-half of the entire continent. Brazil is the most densely forested country; its total forested area is greater than that of all the rest of South America put together, and it possesses more than two-fifths of the continent's accessible forests.

Indian corn, or maize (*Zea mays*), probably native to tropical America, is still cultivated to a large extent throughout the continent. Beans (including several species of the genus *Phaseolus*), widely cultivated by small-scale methods, form an important food item in most countries. Cassava, or yuca (*Manihot esculenta*), and sweet potato (*Ipomoea batatas*) are indigenous to the New World, from which they have been introduced to the other warm regions. The potato (*Solanum tuberosum*) originates from the high Andes, where related wild species are still found. Some other plants have been domesticated in this particular environment, such as the quinoa (*Chenopodium quinoa*) and the canahua (*C. canihua*), both giving very small seeds used as cereals, and tuberoses such as ulluco (*Ullucus tuberosus*) and oca (*Oxalis tuberosa*). Squashes and pumpkins (*Cucurbita* species) have been cultivated since pre-Columbian times and have spread throughout the world, as has the tomato (*Lycopersicon esculentum*), indigenous to the west coast. The peanut, or groundnut (*Arachis hypogaea*), originated in the Paraná River Basin, where the wild species has apparently become extinct. Cashews (*Anacardium occidentale*), now cultivated in most tropical countries, and Brazil nuts (*Bertholletia excelsa*), harvested from wild trees in the Amazon region, are widely used as delicacies but are also eaten locally raw or cooked. Cacao (*Theobroma cacao*), native to the Amazon region, was prized by the ancient Indian civilizations and is cultivated in South America, particularly in the State of Bahia, Brazil. Avocado (*Persea americana*) is native to the same region. Pineapple (*Ananas comosus*), probably indigenous to southern Brazil and the Paraná River Basin, was cultivated in pre-Columbian times throughout tropical South America and the West Indies. Papaya (*Carica papaya*) and guava (*Psidium guajava*) are both native to tropical America.

Humans also introduced a great number of plants to the continent. Sugarcane (*Saccharum officinarum*) has been cultivated in all humid tropical countries since early colonial times. Bananas have become an important local food item and are exported to the rest of the world in enormous quantities. Mango, oranges, lemons, and grapefruits are widespread. Coconut is very common, although its origin is much disputed.

Cereals

Among cereals, rice, which was introduced from Asia, is grown in Colombia, Peru, Brazil, and Uruguay but not in sufficient quantities. Wheat, introduced to the southern part of the continent, is, along with other cereals, widespread in Argentina, but it is also raised locally in Colombia.

The two most important native spices are allspice (*Pimenta officinalis*) and red or chili pepper (*Capsicum annuum*).

Grazing is extensively practiced on all grassland areas throughout the continent. Apart from the Llanos of the Orinoco, the main grazing regions extend through eastern and southern Brazil, as well as across Uruguay and Argentina, where various breeds of cattle are raised. Stock raising also flourishes in the temperate zones of the Andes, where cattle are raised, while the upper, drier zone is mainly devoted to raising sheep, alpacas, and llamas. Foraging is also widespread in savanna areas, and even in regions such as the caatinga, where xerophytic plants grow.

Coffee (*Coffea arabica* and related species), imported from the Old World, is extensively cultivated, particularly in Colombia and Brazil. The most celebrated native plant is the yerba maté, or maté (*Ilex paraguariensis*), indigenous to the Paraná Basin and still exploited in its wild state; its leaves are used to prepare a drink that is very popular in some parts of South America. Tobacco (*Nicotiana tabacum*), smoked by Indians in pre-Columbian times, is cultivated in many countries, but mainly in Brazil and Colombia.

A great number of South American plants provide valuable drugs, among which are quinine (obtained from the bark of several trees of the genus *Cinchona,* indigenous on the eastern slopes of the Andes) and cocaine (obtained from the leaves of the coca tree, *Erythroxylon coca,* found in the eastern Andes from Peru to Bolivia). Among a great variety of aboriginal remedies still in use, some may contain medically valuable substances.

Besides peanuts and some introduced plants, South America has a great number of oil-producing plants, such as the babassu palm, a native of Brazil, the nuts of which are rich in oil and are used for soapmaking. The production of essential oils has much development potential.

Vegetable waxes are produced mainly by the carnauba palm (*Copernicia cerifera*) of Brazil, the leaves of which are covered with a waxy secretion. Vegetable ivory is yielded by the hard seeds of the tagua palm (*Phytelephas macrocarpa*) found in northern South America, notably in Ecuador.

Cotton is partly indigenous to South America and has been used since prehistoric times; large quantities are produced in Brazil, Argentina, Peru, and Ecuador. Kapok and plants producing stem or leaf fibres, such as sisal, are extensively grown. The iraca (*Carludovica palmata*), a plant looking like a stemless palm, is cultivated in Colombia and Ecuador for its fibres; extracted from the young leaves, they are woven into "Panama" hats.

Several plants furnish latex, from which rubber is extracted. Para rubber, or seringa (*Hevea brasiliensis*), and related species, a native of the Amazon region, was known by the Indians. Ceará and manihot rubbers, found in the driest part of Brazil, have the potential for use if given suitable agricultural management. Panama rubber, or caucho (*Castilla elastica* and related species), native to the Amazon region, has yielded poor results.

Balata (*Manilkara bidentata*) yields a nonelastic rubber;

chicle is a gum extracted from the latex of sapodilla (*Achras zapota*) and is used in the preparation of chewing gum.

Domesticated animals

Animal resources. Animal resources are less diversified than botanical resources. Llamas and alpacas were domesticated in the high Andes in Inca times and probably earlier; both derive from the guanaco. Llamas are kept as beasts of burden and alpacas for their wool. Guinea pigs were domesticated in pre-Inca times in the same area and raised for meat. The keeping of native domesticated animals did not extend beyond the geographic limits in which their wild ancestors were found, with the exception of the Muscovy duck (*Cairina moschata*). Among animals introduced to the continent were cattle brought by Columbus on his second voyage. Numerous in Brazil, Uruguay, Argentina, Venezuela, and Colombia, cattle have become distributed throughout the continent where there is suitable grazing land. Horses are widespread and play an essential role in the economy of most countries. Sheep are raised successfully in the Argentinian region of Patagonia, in the Andean highlands, and on the Falkland Islands. Pigs and smaller domestic animals are present everywhere.

Game is plentiful in most habitats, though mammals are few and specialized. Deer are represented by several species; red deer have been acclimatized in Argentina. Specialized rodents (such as agoutis, pacas, capybaras, and maras), waterfowl, and a large series of terrestrial birds (such as tinamous, guans, and curassows) provide excellent game. Many Indians, mainly in the rain forest, live upon birds and monkeys. Many animals are collected alive and exported as pets or for zoos, the heavy demand being justified by their strangeness and beauty.

Several neotropical animals provide world-famous fur or wool. Chinchilla (*Chinchilla laniger*) were native to the high Andes of Peru, Bolivia, Chile, and northern Argentina, where thousands lived in flourishing colonies. They were hunted for their delicate gray fur to the point of extinction in their wild state, except possibly in some districts of Bolivia and Argentina. Vicuñas provide the finest wool on Earth and have been hunted since the time of the Incas; they have dwindled alarmingly in the upper Andes and are still poached despite laws and a ban placed upon trade in their fur. The giant otter (*Pteronura brasiliensis*) of the Amazon; several spotted cats, including the jaguar and the ocelot; and rodents, such as the nutria, also provide highly prized furs, but overexploitation of many species has depleted the stocks.

Freshwater fish, abundant in most of the rivers, have been exploited on a large scale since the earliest time, notably in the Amazon region and in the Guianas. Trout have been introduced into Andean lakes and torrents—sometimes to the detriment of the endemic species.

Marine fisheries are very prosperous, since South America is surrounded to a large extent by highly productive fishing areas. Peru is one of the world's major fishing nations.

South America is well known for colonies of seabirds along the Peruvian and Chilean coasts, resulting in the accumulation of dung (guano), a highly prized fertilizer. Pinnipeds (a suborder of aquatic carnivorous animals, including seals and walruses) are exploited for their oil, particularly in Uruguay. Fur seals are distributed along the southern coasts of South America, but their numbers have been considerably reduced by hunting. (J.P.D.)

INDUSTRY

Stages of industrialization

The first stage of industrialization, which might be called the pre-factory stage, has been left behind by most South American countries. The second stage, commonly known as that of the traditional industries (textiles, food, beverages, tobacco, sugar), is still present, although its importance varies from country to country as a result of special conditions or because of historical circumstances. The final stage, the development of basic industries and the production of simple machinery, is in progress in a number of countries in South America, including Argentina, Brazil, and Venezuela.

The process of import substitution that began in some countries even before the Great Depression and that later accelerated during World War II gave birth to many industries producing intermediate and capital goods; in some instances these industries have survived only because they have been protected from foreign competition by subsidies, tariffs, and trade restrictions. Import substitution has shown signs of weakening, however, in both the traditional industries sector, and the production of consumer goods.

Manufactured products are sold at high prices not only because of trade protection but also because of the small scale of production, the use of inadequate technology, and defects in the economic infrastructure (structure of basic services), mainly in the form of insufficient transport and energy.

The slow growth of the industrial sector and the adoption of production techniques requiring a high ratio of capital to labour resulted in a very limited contribution of industrialization to the creation of new sources of employment, especially in view of the rapid growth of the labour force.

The gross domestic product (GDP) originating in the industrial sector is approximately one-third of the gross national product. Of this total, about four-fifths represents manufacturing and the rest civil construction. About one-fourth of the labour force is occupied in the manufacturing sector.

In the two most industrialized countries—Argentina and Brazil—the production of foodstuffs, beverages, and tobacco accounts for only about one-seventh of the total manufacturing output. The production of metallurgy and mechanical industries represents more than one-third of the total manufacturing output, while chemicals and petroleum refining contribute about one-fifth, and textiles, footwear, and apparel contribute about one-tenth.

During the late 20th century, South American industrial production made substantial gains: the production of cement and steel (ingots, rolled, plates, and sheets), pig iron, automobiles, and household appliances has risen greatly.

Unlike some manufacturing industries that have become sophisticated, the construction industry in South America is in an elementary stage. Construction techniques use much labour, quality is low, and costs are quite high. Despite the existence of a huge housing shortage, building has proceeded at a very slow pace.

POWER AND IRRIGATION

Electrical energy output

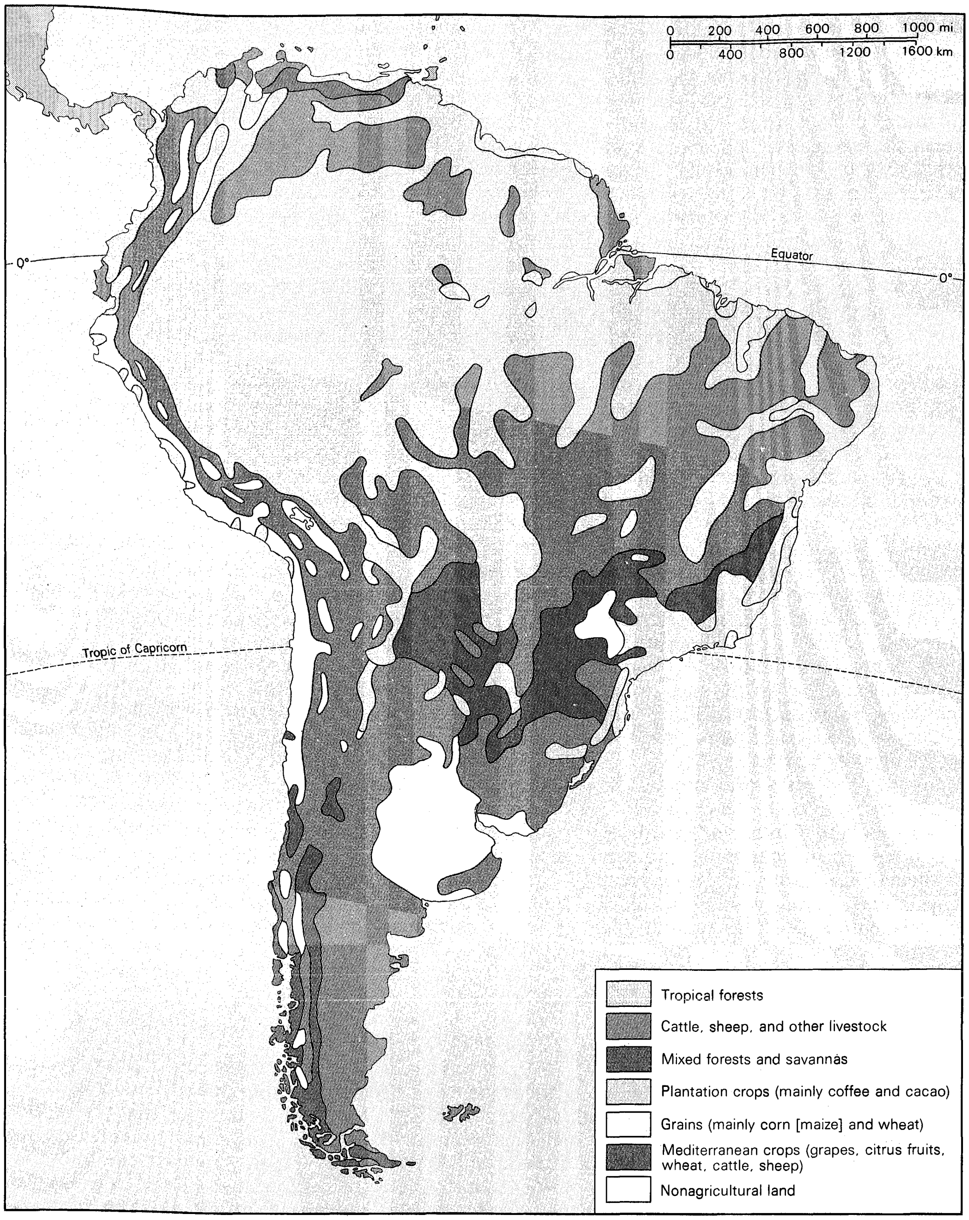
The total annual generation of electricity in South America has increased considerably, but there is much disparity between countries; Venezuela had more than twice the regional average; Argentina, Chile, and Uruguay about average; and Guyana and Peru only about one-half the average. Industry consumes almost two-fifths of the electric power generated in South America. Household consumption takes another two-fifths, and the rest represents either transmission losses or other uses.

Private installed capacity, which consists mainly of thermoelectric plants operated at high cost, represents a low percentage of total installed capacity. The trend has been toward increases in hydroelectric power—in part because of the increasing importance of the public power services as distinct from private power generation.

Argentina, Brazil, Uruguay, and Venezuela have ambitious electrical installation programs that can take advantage of the region's abundant hydroelectric resources. This will increasingly require extensive use of interconnecting systems that have already been effectively developed in such countries as Brazil, Chile, and Uruguay.

Most of the hydroelectric projects are also used for irrigation, but there are many instances in which the first purpose has been accomplished, but the construction of distribution canals for irrigation has yet to be achieved.

Only a small percentage of arable land and of land under permanent crops is irrigated. Notwithstanding the fact that the largest agricultural countries account for most of South America's irrigated land, Chile, Peru, and Ecuador show the highest proportion of irrigation in relation to cultivable land. In Peru irrigation activities are confined to the narrow strip of coastal plain. In Chile, about one-fourth of arable land uses irrigation water; this land consists mostly of farms of less than 25 acres in size. Because of the relative scarcity of farmland in these countries, it


0 200 400 600 800 1000 mi
0 400 800 1200 1600 km
Equator
0°
0°
Tropic of Capricorn
Tropical forests
Cattle, sheep, and other livestock
Mixed forests and savannas
Plantation crops (mainly coffee and cacao)
Grains (mainly corn [maize] and wheat)
Mediterranean crops (grapes, citrus fruits, wheat, cattle, sheep)
Nonagricultural land

Agricultural regions of South America.

may be said that large-scale irrigation is practically a condition of national existence.

Argentina accounts for almost one-fourth of the land under irrigation in South America. Irrigated areas in Argentina are devoted mainly to the growing of fruit.

AGRICULTURE

Even though agriculture is a large sector in South America's economy, several countries have to import agricultural commodities, including products that could be supplied by expanding local or regional production. The largest crop in terms of the area harvested and in output is corn, the second largest is wheat, and the third is rice. Cotton, potatoes, barley, sweet potatoes, rye, millet, and sorghum are also important agricultural products. The number of cattle raised in South America represents a good percentage of the world's total, and horses constitute another important livestock. There are also large numbers of pigs and sheep in South America.

Land productivity is, on the whole, far below that prevailing in other regions. The increases in production are chiefly attributable to extensions of the cultivated area.

ral
employ-
nt

In spite of the large shifts of population toward the cities in recent years, a large proportion of the widespread unemployment prevailing in South America is still found in the rural areas. About one-third of the economically active population works in the agricultural sector, as compared with almost one-half of the population for the world as a whole. The lack of educational skills among farm workers, the widespread existence of unemployment, and the limited use of modern methods and management techniques in the countries of South America have resulted in a very low average product per worker for the region.

South America extends over an area of 4,400,000,000 acres, of which only about one-eighth is arable land, land under permanent crops, or land occupied by meadows and pastures. More than one-half of the total land area is covered by forests, and another one-eighth has other uses. It is a general conviction throughout South America that land utilization is inefficient. Inflation control, land taxes based on productive capacity, agrarian reforms that consist of more than changes in the system of land tenure, as well as other policies, are all felt to be helpful in improving the efficiency with which land is used. The great shortage of farm capital (as well as the high cost of obtaining it) is, however, one of the most serious difficulties facing agricultural development in the countries of South America.

In short, the one factor preventing a rapid growth in agricultural output is an unfavourable relation between product prices and input costs. Capital, fertilizer, modern machines, and fuel command high prices, while prices received by farmers tend to be low.

TRADE

Banks and financial institutions are usually large enterprises with branches in many cities and towns; a high proportion of these are government owned. Wholesale and particularly retail business enterprises, on the other hand, are mostly individual concerns and in many cases are family shops. Department stores or chain stores are seldom found in South America.

Characteristics of commerce in South America

In contrast with some highly sophisticated industries in the manufacturing sector, distributive and retail firms use rather elementary methods and are old-fashioned in organization. This is probably due to the scarcity of managerial skills and to the relatively small share of credit that this sector receives in comparison with industry and agriculture. Furthermore, the development of commerce depends heavily on an efficient system of communications and transportation, which is still lacking in much of South America.

Internal trade. The relative importance of intra-regional trade continues to be small, in spite of its rapid growth since World War II. Intra-regional trade accounts for about one-sixth of total exports. In South America there exists a firm conviction that intensification of intra-regional trade is a necessary condition for rapid economic growth because it would help to reduce the region's excessive dependence on a few foreign markets, diversify exports, and help to solve balance-of-payments problems.

South American trade with the rest of Latin America is centred in a few countries. Argentina, Brazil, and Venezuela account for more than one-half of exports; these same three countries also absorb about one-half of the imports from the rest of Latin America.

The Latin American Integration Association

All of the independent South American countries except Guyana belong to the Latin American Integration Association (LAIA; formerly the Latin American Free Trade Association.) Despite formidable obstacles, the LAIA has directed its efforts toward designing a common trade policy for member countries, toward the gradual and automatic reduction of import duties and other restrictions on imports from the rest of the world, and toward arriving at agreements to compensate trade payments between member countries, as well as at making reciprocal credit arrangements between central banks. In addition, Bolivia, Colombia, Ecuador, Peru, and Venezuela have formed the Andean Subregional Group for the purpose of reaching agreement on common trade problems, such as an external tariff, reductions in tariffs applicable to subregional products, and coordinating their policies toward foreign investment.

More than 11,000 products have been granted some concession of one kind or another and have entered the intra-area trade. Trade among LAIA members consists mainly of basic commodities (about three-fourths) and semimanufactured and manufactured goods (about one-fifth). Among basic commodities the most important items are foodstuffs, beverages and tobacco, crude materials, and nonferrous metals. Among manufactured goods the main items of trade are chemicals, machinery, and transport equipment.

All these efforts toward integration are partially responsible for the rapid growth of the area's internal trade and for a large contribution toward a greater balance in trade among these countries. Since the 1960s intra-area trade has stabilized among South American countries, and surpluses and deficits have declined considerably.

External trade. External trade represents a key element in South America's economic growth. Essential imports, particularly capital and basic intermediate goods, are needed to accelerate the industrial process. A major problem has been that exports and net external financing have not been providing sufficient income to pay for those imports.

Despite a satisfactory increase in trade, South America's share of world trade has remained small, primarily because the growth in trade between major world industrialized countries has grown at an even faster rate. The share that South American countries contribute to the world's imports and exports is only a small percentage of the total. South America's share in total U.S. imports, also small, has been declining.

South America's major exports, in value terms, are foodstuffs, fuels, and crude materials. Within the first group the most important are sugar, bananas, cocoa, coffee, tobacco, beef, corn, and wheat; in the second group, petroleum and petroleum products; and in the third group, linseed oil, cotton, cattle hides, quebracho, fish meal, wool, copper, tin, iron ore, lead, and zinc. About one-third of these exports go to the United States, another one-third to western Europe, and one-eighth to Latin America.

Almost three-fourths of South America's imports consist of machinery, vehicles and parts, chemicals and pharmaceuticals, paper and paperboard, textile products, and other manufactures. More than one-third of South America's total imports are from the United States, almost one-third from western Europe, and about one-seventh from Latin America.

In general, South America's foreign trade sector has shown little diversification of exports, a heavy dependence on imports for domestic supplies of industrial goods, and a marked imbalance in trade with the industrialized countries.

TRANSPORTATION

In an area the size of South America, an efficient system of transportation is necessary for the expansion of national markets, as well as for the development of the hinterland. Unlike North America, South America does not have an adequately integrated transportation system. There are many separate road and rail systems, but they vary in extent and type. Since the mid-20th century, however, internal and international air routes have formed a reasonably well-integrated and complete transportation system.

Waterways. Traffic on the inland waterways faces many drawbacks, including dry seasons (and sometimes dry years), the direction of water flow, and difficult rapids. There are two systems of international importance—the Paraná–Uruguay basin (which includes territory in four countries) and the Amazon Basin (six countries), each having several thousand miles of navigable waterways. Furthermore, there are three other minor systems—the Magdalena in Colombia, the Orinoco in Venezuela, and the São Francisco in Brazil. The remaining rivers are unsuitable for navigation. In general, the volume of traffic on the waterways of South America is small, and the prospects for increasing it are limited. The major products transported by waterway are timber, mineral ores, and tropical fruits.

Maritime transport. Sea transport can be divided into three categories. The first of these is represented by the service linking the different ports along a country's coastline, as in Chile, Brazil, and Argentina, which have established regular cargo and passenger services of this type. Second are the services running between the various South American countries. Third are regular services running between countries in South America and other countries of the world.

A determined effort is being made by several countries to develop and enlarge national merchant marines, partly in order to arrest trends that in the past have resulted in their trade being carried by ships from outside the region and partly as an instrument to promote regional integration and to improve the national balance of payments.

Problems of rail transporta tion

Railways. The position of the railways in South America is unsatisfactory. Rail transport is plagued by operational problems and by a persistent loss of importance as a result of competition from road and air transport. Proportionally, railways handle only a small amount of goods because most bulky items moved are for export and are produced close to the coast. Almost all lines are single track, making traffic very slow and discouraging passenger traffic. Many countries have two or more gauges, hindering the possibility of an efficiently integrated rail system. Railways administrations, however, are making continuous efforts to improve services and reduce budget deficits. Except for some minor industrial lines, almost all railroads are government owned.

Roads. The length of roads used in South America is impressive. In many countries, however, only a small percentage is paved, while in other instances they are not wide enough for two vehicles to pass easily. The remainder of the system consisted of improved roads or simply dirt roads.

The improvement of existing roads and the construction of new ones has been proceeding rapidly—perhaps because some countries, such as Argentina and Brazil, manufacture motor vehicles in large quantities, while other countries are interested in expanding their automobile output. Furthermore, the inefficient service of the railroads has opened the way for the development of truck transportation, which has taken an increasingly large share of the volume of goods carried by land.

In developing international highways, particular attention has been paid to road-integration projects in which the Inter-American Development Bank and the World Bank have been heavily engaged. In the construction of the Brasília to Lima route, important progress has been made. Argentina has been investing large amounts to improve the integration of highways between its capital of Buenos Aires and Bolivia, from Mendoza to Santiago (Chile), and in the northeastern region bordering Paraguay and Brazil. Brazil continues to have the longest network of roads belonging to the Pan-American Highway system; the system itself extends throughout Latin America.

Air transport. Air transportation in South America has developed very rapidly since World War II, mainly because it is well suited to the area's requirements and avoids the forbidding geographical obstacles that hamper surface facilities. This is particularly true with respect to passenger traffic, but it applies less to the handling of bulky freight.

Each country has its own system of internal air services, operated generally by government-owned and sometimes by heavily subsidized private companies. This circumstance creates certain gaps in the otherwise reasonably well-integrated air transport system that has developed over the continent.

International air links

All of the South American capitals and most of the large cities are linked by direct air transport services with the major traffic centres of the United States and Europe. Domestic traffic has grown less rapidly than international traffic, but, together these developments resulted in maintenance of South America's share of total world air transport. (H.F.A.)

HISTORY

Only a brief historical survey of South American history is offered here. For full treatments see PRE-COLUMBIAN CIVILIZATIONS and LATIN AMERICA, HISTORY OF.

Pre-Columbian evolution

Before their discovery by the Europeans in the 16th century, the inhabitants of the South American continent lacked any form of script and, unlike the Maya and Aztecs of Mexico and Middle America, also had no type of recorded calendar. Consequently, knowledge of the prehistory of the continent is based on local traditions, accounts written by the first discoverers, and especially on archaeological field excavations. Because traditions of any significance are generally lacking except for the Incas of Peru and because historical accounts are limited to the conquest period, the dependence on archaeology is even greater.

EARLY CULTURES

The earliest inhabitants of South America were nomadic hunters, fishers, and gatherers who pushed southward from North America by way of Middle America and the Isthmus of Panama. There are, to be sure, other possible migration routes, including the transpacific, but the evidence for these is yet to be satisfactorily established. Remains of the early nomads have been found in such widely separated areas as highland Ecuador, eastern Brazil, and southern Chile near the Strait of Magellan. The remains include skulls, none particularly primitive, and stone and bone implements.

The hunting and gathering pattern survived in southern South America until historic times, but elsewhere the development of more advanced cultures and civilizations was based on farming. Various centres of plant domestication have been suggested, such as the marginal areas of the tropical forest, the high Andes, Middle America, and even the Old World for such plants as the bottle gourd, sweet potato, and perhaps cotton. Wherever the origin, the highest civilizations in South America were based on intensive agriculture, and this was best developed in the Andean mountains and along the Pacific coast, particularly in the Central Andes, which includes the mountains and coast of Peru and part of Bolivia.

The gross outline of cultural development in the Central Andes has been reasonably well established. The earliest remains, found on the northern coast of Peru, represent a population that depended heavily on fishing and gathering of shellfish for its subsistence but that also cultivated some domesticated plants, including the bottle gourd, squash, beans, chili pepper, and cotton. These people had no ceramics but did make baskets, mats, crudely flaked implements, and cloth from twined cotton fibres. They lived in small, subterranean houses and cooked their food by placing hot stones in some kind of water containers. Their remains are roughly dated as from 3000 to 1000 BC.

Ceramics were introduced at approximately 1000 BC, together with true weaving and domesticated corn (maize) and cassava (manioc), which initiated a long formative period during which many new techniques were tried and perfected. Systems of irrigation were established, new metallurgical techniques were invented, and many different shapes of sun-baked clay adobes were tried out in buildings. Religion, too, grew in importance, particularly in the Chavin culture. The stylized Chavin feline design, the dominant religious symbol, appears on stone carving, textiles, ceramics, and other media in many parts of the mountains and the coast.

Distinctive regional cultures

The long formative period was the basis for the distinctive regional cultures that flourished in almost every part of the Central Andes from approximately 400 BC to AD 600. This has been called the Early Intermediate, or Classic, Period because of the high artistic and technical achievements in ceramics, metallurgy, weaving, and architecture. Some of the outstanding local cultures are the Moche on the northern coast of Peru, noted for faithfully modelled and realistically painted ceramics; the Nazca on the southern coast, famed for polychrome embroideries and multicoloured clay vessels; Recuay in the northern highlands, with incised statues and small modelled figures on ceramic vessels; and Pucará and Tiahuanaco in the Bolivian highlands, noted for polished and fitted stone masonry and low-relief carving.

Each of the regional cultures started to expand as it grew stronger, and conflict followed. The Tiahuanaco culture became dominant and achieved the greatest extension. It can be traced throughout much of Bolivia, northern Chile, and the coast and highlands of Peru. There are indications that this Tiahuanaco expansionist movement was strongly motivated by religion, but it was nonetheless sufficiently organized to eclipse most of the local styles that it encountered. Although roughly dated from AD 600 to 1000, there is no evidence that any true unity, political or religious, was maintained throughout this time period, and as its influence diminished, regional cultures once again reappeared, each now definitely formulated into local political units. At least one of these, the Chimu on the northern coast, was well enough organized to be called an empire. Its arts and crafts reflected the traditions of the earlier Moche now mixed with Tiahuanaco. Its population was numerous and was settled in large ceremonial cities, such as Chan Chan, near the contemporary town of Trujillo, Peru, or assigned to live in strategically placed garrisons. The political and religious organization shows many parallels to the Inca Empire, which followed.

THE INCAS

The Inca Empire, discovered in full flower by the Spanish conquerors in 1532, was thus the culmination of many years of cultural growth. The Incas contributed their genius for political organization and built up an empire that not only united all the Central Andes but also extended north to Colombia and south to Chile and Argentina. For the Inca period the archaeological evidence is supplemented by traditions and by the written accounts from the early conquest years.

Unfortunately, the accounts preserved are not in agreement on all points, and it is not easy to reconcile discrepancies. For instance, one of the most important documents relating to Inca civilization is the account of Garcilaso de la Vega, son of a woman of the ruling Inca house by one of the conquerors, who had been brought up in his mother's family and had exceptional opportunities for collecting the ancient traditions. But his natural sympathies led him to stress the virtues of the Inca regime to the point of exaggeration. On the other hand, the "official" history of Pedro Sarmiento de Gamboa was written rather with the idea of showing the Incas as oppressors and the Spaniards as liberators.

In general terms, the history of the rise of the Incas to power would seem to be as follows: Their history begins with a frankly mythological account of a migration from the south to Cuzco, under the leadership of Manco Capac (Manqo Qhapaq). For years the Incas, living the same life as other Andean tribes, consolidated their power in the Cuzco valley and then began to make their influence felt among the surrounding tribes. A career of conquest, undertaken partly as a religious crusade, eventually raised the tribe to the position of a ruling caste, whose sovereign was regarded as a god on earth.

The first historical Inca (bearing the royal title of Sapa Inca) was Sinchi Roca (Zinchi Roq'a), by most authorities recorded as the "son" of Manco Capac. His date may be set at about AD 1200. The Inca power began to make itself felt outside the Cuzco valley in a southerly direction (among the Colla peoples) and was consolidated by Sinchi Roca's successor. Mayta Capac (Mayta Qhapaq) continued the policy of expansion toward the south, and the activities of these rulers led to the annexation and settlement of a large region, extending as far as Lake Titicaca, in succeeding reigns.

West and northwest of the Inca domain, the Chanca people, living under similar conditions, had built up a powerful confederation. A clash was inevitable, and the struggle was desperate; but the Inca, supported by their newly acquired tributaries from the Colla, emerged victorious. As a result, a large expanse of territory was added to the Inca sphere, inhabited by cognate peoples, practicing similar customs and religious observances and therefore amenable to the new conditions. The annexation of the Chanca district opened the way to the coast, the conquest of which, for climatic and religious reasons, was a long and laborious process. The reign of Pachacutec Inca Yupanqui (Pachakuti 'Inka Yupanki) seems to have been devoted to a reorganization of the Inca system in view of its new colonial responsibilities. But under Topa Inca Yupanqui (Thupa 'Inka Yupanki) and Huayna Capac (Wayna Qhapaq), the two great "conquerors" of the Inca dynasty, Inca domination was imposed throughout the highland region from Tucumán in Argentina to Quito in Ecuador: The conquest of the Ica (Arhuaco) and Chimu peoples of the coast was a more laborious process. (T.A.J./W.C.Be.)

OTHER PEOPLES

Elsewhere in South America the populations were far less numerous and far less advanced in civilization. In the northern mountains were the Chibchas, and along the northern coasts were the Caribs and Arawaks, the latter extending southward. In Brazil and the upper Plata Basin the Tupi-Guarani tribes predominated, with a semisedentary culture. The extreme south was sparsely populated by the barbarous Patagonians on the east and the warlike but disciplined Araucanians on the west.

European conquest and settlement

THE AGE OF DISCOVERY

Early European explorers

The continent of South America was first visited by Europeans in 1498, when Christopher Columbus, on his third voyage, touched near the mouth of the Orinoco. Others soon came after, and by 1509 the eastern coast had been followed as far as the Río de la Plata. In 1513 Vasco Nuñez de Balboa discovered the Pacific Ocean at the Isthmus of Panama; in 1520 Ferdinand Magellan navigated the strait that bears his name; in 1527 Francisco Pizarro landed on the coast of Peru at Tumbez, and, in the same year, Sebastian Cabot ascended the Río de la Plata and the Paraná River as far as the mouth of the Bermejo; in 1533 Cartagena was founded and the native town of Quito peacefully occupied; in 1535 Pizarro founded Lima, and, in the same year, Diego de Almagro invaded Chile, and Pedro de Mendoza established a settlement at Buenos Aires; in 1537 Gonzalo Jiménez de Quesada made his way from Santa Marta up the Magdalena River to Bogotá, the capital of the Chibchas; and in 1541 Pedro de Valdivia founded Santiago de Chile, and Francisco de Orellana travelled from eastern Ecuador to the Atlantic by the Napo and Amazon rivers. It is claimed that the bay of Rio de Janeiro was discovered by the Portuguese in 1502. Martim Afonso de Sousa visited it in 1531 but went on to found the first Portuguese settlement in Brazil at São Vicente (1532), near Santos.

THE COLONIAL PERIOD

Colonial settlement, western coast. The initial penetration of the Spaniards was into the Inca region. Its primary purpose was to secure precious metals, and, after seizing what the Indians possessed, the conquerors turned promptly to the mining of additional supplies. Agriculture was necessary, however, to provide food and clothing; and, in searching for gold and silver, the Spaniards occupied many places that were suited only for nonmining pursuits. The Spanish monarchs were zealous churchmen, too, and sought from the outset, through priests sent with the expeditions, both to convert and to improve the natives. Thus, many settlements were based on farming, and all kinds of European domestic animals were introduced and propagated with diligence.

Mining of gold and silver

The principal mining regions were in the high Andes inland from Lima and in south central Bolivia. While gold extraction was widely scattered, with Colombia eventually leading, silver was far more plentiful; and the discovery of the huge "silver mountain" at Potosí in 1545 made that place for many years the largest population centre in the hemisphere. The principal farming areas were along the Peruvian coast, where sugar, rice, and corn (maize) were grown on a large scale, and in the "central vale" of Chile, which was for a time the granary of Spanish South America.

Both mining and agriculture were based on the labour of the Indians, who had already been accustomed by their Inca masters to systematic toil. They were adapted without very great difficulty to Spain's *encomienda* system, under the direction of the mine owners and plantation operators who received grants from the Spanish crown. By the end of the 16th century, when the process of subjugation had been in good part accomplished, the Andean countries were organized in a wide system of agricultural and mining settlements. The number of Spaniards needed for troops and administration, as well as for mining, farming, industries, construction, and trade, was large, and these men at once formed unions with the Indian women, since relatively few families came from Spain. The hybrid, or mestizo, populations that resulted from this intermingling soon swelled to large proportions, particularly in the lowlands, and the mestizos played an important auxiliary role as overseers and superior employees in all activities.

As a basis for settlement the mining operations receded somewhat in the 17th century, because the best deposits became partially exhausted and because other occupations increased. In the highlands the Merino sheep gave the Indians a large new industry. In Chile, Peru, and western Argentina, wine grapes and olives were grown. In Ecuador, vegetable ivory, balsa wood, and, after 1700 cinchona were exploited. The numerous edifices and houses of European type were built of brick as well as stone or adobe, with the necessary wood brought from southern Chile. Boatbuilding was active.

Colonial settlement, eastern coast. Unlike the Andean western coastal countries, the eastern regions of the continent had rather few Indians and for two centuries were not known to contain minerals. Settlement from the beginning was based upon agriculture and natural products. The Spanish occupation in the Plata Basin was retarded by the southern Indians, with whom the Spaniards were at first unable to cope. They found in Paraguay, however, a well-watered rolling country peopled by the non-nomadic and nonhostile Guaranis, and Asunción in 1537 became the seat of Spanish operations. From it were founded towns farther down the river and eventually, in 1580, the city of Buenos Aires. By the late 18th century the rich humus of east central Argentina was coming into use for grains, and the plains were used for cattle, while in the north and in Paraguay, tropical crops and maté were produced. The absence of metals and inorganic combustibles handicapped other development, as did the Spanish prohibition of any trade except with Spain via Peru.

The Portuguese settlement of Brazil, almost one-half of the continent, was undertaken through 20 captaincies granted to nobles for tropical plantations. Because the Indians were found to be ineffective workers, blacks were imported in large numbers from Africa. Thus, along the northern littoral, despite uneven management and deficient rainfall, the production of sugar, rice, tobacco, indigo, and cacao soon became extensive; and dyewoods, cabinet woods, and animal skins were also exported. Southward from Rio de Janeiro, and particularly in the cooler tablelands of São Paulo, a more European type of farming came into existence. The hardy pioneer population showed energy in subduing and utilizing the Indians, thus opening the interior of central Brazil. At the beginning of the 18th century came the discovery of gold, diamonds, and a diversity of valuable ores in Minas Gerais, attracting an influx of population and making Brazil the world's main producer of gold for more than 100 years. In the south the settlement was more gradual but came to be based on cattle and sheep. (W.Ft.)

The legacy of Iberian colonialism. This basic sociocultural framework, with all its inherent ambivalences and contradictions, emerged during the colonial era and exercised a deep and lasting impact on the political, economic, and social structure of the new South American nations. Only after many decades was this impact blurred by the action of other external influences, change occurring from within, and a growing sense of national identity.

orces
ıaping
ɔlitical
:ography

The political geography of the region emerged to a great extent as a result of the administrative and judiciary divisions established by Spain and Portugal. The three Spanish viceroyalties of Lima (1542), Nueva Granada (1717), and Río de la Plata (1776), the captaincies general, and the *audiencias* (units that were partly administrative and partly judicial) shaped the political boundaries, while the isolation in which the Spanish metropolis maintained each unit was at least one factor resulting in the fragmentation of Spanish America into 10 (and later into 20) independent states; geography and foreign interests were probably other important causes. Though Brazil maintained its unity, its component states, as well as their pronounced regionalism, were at least in part a heritage of the Portuguese administration and the laxity of its central control. Although the political organization of the independent nations of South America was largely molded after the French and U.S. models, the political culture, values, attitudes, and behaviour—particularly during the 19th century—were to a considerable extent the reflection of Iberian politics.

The cultural heritage. Though intellectual life, and to a lesser extent education, were strongly influenced by France in particular and Europe in general (later also by the United States), much of the Spanish and Portuguese intellectual heritage was transmitted, together with the language, to the new nations. Spanish remained the official language of all the independent states except Brazil (Portuguese), Guyana (English), and Suriname (Dutch).

Both Spanish and Portuguese have been influenced by local conditions. Spoken Spanish presents some differences between the various South American countries and also exhibits differences from the Spanish spoken in Spain. The hypothesis that the Andalusian dialect exercised the greatest influence seems controversial. At least one linguist has distinguished four varieties of Spanish in South America and two main types of pronunciation. Indian influences (and, in Argentina and Uruguay, the influence of European immigration) have had a certain impact. The differences between spoken Portuguese in Brazil and in Portugal itself are perhaps greater than those between Spanish as spoken in South America and as spoken in Spain.

Economic and social factors. In the economy a similar continuity is discernible. Colonial South America was a producer of raw materials—precious metals and tropical crops—and a market for European (mostly Spanish) manufactures. Local manufacturing was discouraged, and the rest of the economy—based mostly on subsistence agriculture and on some artisanal production for local lower class consumption—contributed to the surplus that helped to support the local colonial administration and the higher stratum of colonial society, as well as the Spanish (or, in Brazil, the Portuguese) royal treasury. After independence this economic framework continued to exist in the form of the primary export economy that was predominant in the region until the 1930s and that still remains an essential part of it today.

The same process that may be observed with regard to social organization and stratification was closely related, both during the colonial period and afterward, to the system of land tenure and to the pattern of labour relations. The great landed estate, the hacienda (called fazenda in Brazil), was one of the cornerstones of colonial society. The hacienda became an autonomous, self-sufficient institution not only in economic terms but also politically and socially. It was a self-sufficient productive unit, with the great landowner, or hacendado, generally exercising absolute authority—protective, oppressive, or both—over all the groups living on the hacienda; these included his own family, as well as the numerous servants, serfs, and slaves who provided labour for the agricultural, artisanal, and other activities.

Institution
of the
hacienda

The role of the hacienda was reinforced during the 17th century by the weakening of the colonial administration centred in the cities. Increasingly, the countryside—that is to say the great landowners—acquired importance. For many decades after independence, the role of the hacienda was further reinforced by the breakdown of state organization and authority. The increasing concentration of landed property that emerged during the colonial period typically occurred against the will of the crown; this displeasure was usually expressed in legislative measures, and sometimes there were actual attempts to enforce them. Labour relations established in colonial times also reflected the contradictory attitudes of Iberian rule. The *encomienda* (which meant "to entrust" a number of Indians to a Spaniard) was theoretically established as an institution for the protection of the indigenous population and for its conversion to the Christian faith. In fact, however, it became a means of reducing the Indians to slavery or serfdom. This was an important factor in the abolition of the *encomienda* and similar institutions, but the same semifeudal relationships were reproduced in the hacienda and continued under various partially modernized or transformed forms through the 19th century. In certain areas these relationships did not disappear until the first half of the 20th century, when social change in rural areas began to gain momentum.

In colonial times three main estates (social groups) were legally recognized—the Spaniards, the Indians, and the *castas* (castes), the latter being composed of those of mixed blood (such as the mestizo and mulatto) and of African slaves. In fact, however, the hierarchization was different: the European-born Spaniards, who had a monopoly of

The
colonial
hierarchy
and its
effects

most government offices and of authorized foreign trade, as well as other advantages, formed the highest stratum; below them were the Creoles (white or "whitish") followed by the mestizos; in the lower strata were the mulattoes, zambos, free Africans, slaves, and—at the bottom—the Indians.

This inherited social structure resulted in the failure to create a viable modern political system in the new nations. In addition, the central economic role played by raw-material exports not only reinforced the *latifundium* (landed estate system), maintained economic dependence and vulnerability, and provided an obstacle to industrialization, but it also determined forms of social relationships that in turn tended to perpetuate the archaic social structure. (G.Ge.)

The nations of South America

POSTINDEPENDENCE DEVELOPMENT

With the independence of the Spanish colonies in the early 19th century, the last Spanish restrictions upon trade and local enterprise were swept away. Freedom to export to world markets aided the expansion of agriculture, even though factional strife hampered productivity. Road construction progressed, and after 1850 railways were built from many coastal points into the interior, usually with British and other external assistance. Wheat and hides from Argentina were sold in quantity to Europe. Coffee was introduced into São Paulo and within a few decades gained for Brazil the preeminence in that staple that it thereafter maintained. Cacao became important in Ecuador, Brazil, and Venezuela.

Immigration was actively promoted and reached considerable volume in the Atlantic countries. Agriculture on the western coast showed less expansion, because of remoteness and the exiguity of arable land, while mineral industries for a time flagged. In no part of South America did independence bring the breaking up of the vast landed estates, and agriculture, in consequence, did not have the stimulus of diffused ownership. A compensatory gain in Peru was the exploitation of guano on the Chincha Islands and in Chile the export of natural nitrate from the Atacama Desert. The arrival of manufactures from Europe at lower costs than in colonial days was not favourable to South American industrial development, especially as motive power, industrial training, and local capital were not available. Yet simple industries became widespread: textile mills multiplied greatly; cereals, sugar, and leather were processed; and the output of clothing, furniture, footwear, soap, ceramics, and beverages was everywhere augmented.

The decades just before World War I brought new developments. Natural rubber was exploited in the Amazon, but exports of nuts, waxes, and skins proved more permanent. The Plata region built large packing plants and shipped chilled meat, while quebracho extract was produced in the Paraguay area. Along the Caribbean, bananas were cultivated, and the mountain valleys of Colombia developed mild coffees. Chilean nitrate became a bonanza, and Chile began to utilize its coal and iron. Copper and tin mining were revived in the Andes, and manganese from Brazil and bauxite from the Guianas entered world trade.

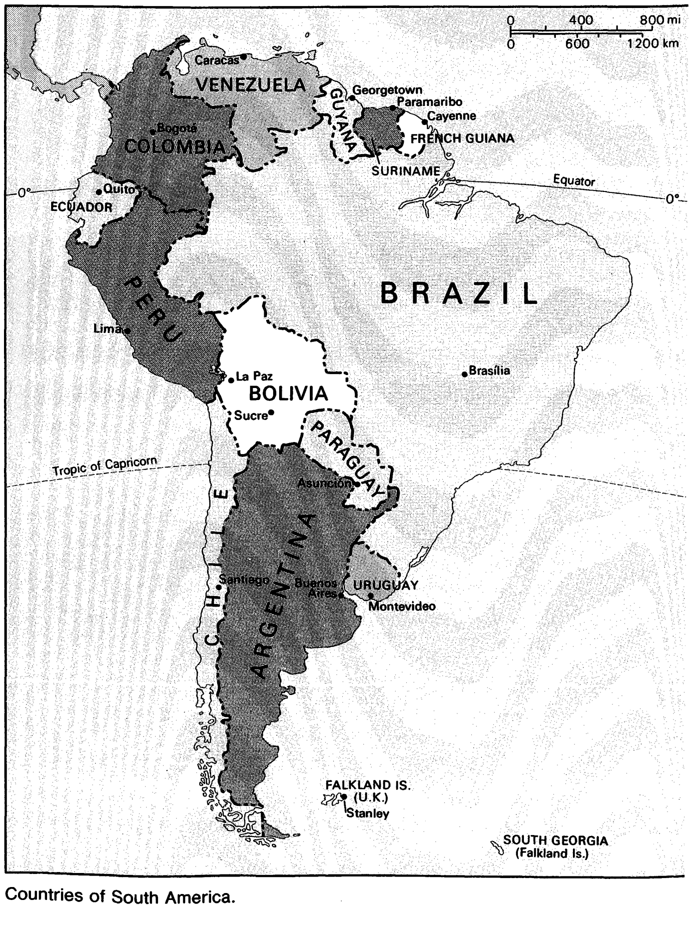
THE 20TH CENTURY

Economic developments. The opening of the Panama Canal benefitted the western coast, and World War I gave added importance to South American metals. Both tended to shift trade away from Europe to the United States. After the war, petroleum extraction was begun in Colombia and Venezuela: and postwar scarcity of manufactures, coupled with mounting nationalistic sentiment, stimulated factory projects for "national" goods. Governments secured large loans from New York City and London and improved roads, harbours, and education. Both immigration and the birth rate rose. The universities and press participated in the modernization movement. The world depression of the 1930s reversed these trends as to metal exports and immigrant imports, although a number of Spanish republican refugees arrived with skills and capital. But popular discontent arising from the Depression caused agrarian reform to be undertaken in several countries, to the ultimate advantage of agriculture.

World War II resuscitated all exports from South America, despite the German submarine campaign, and strengthened the continent's overall financial position. Postwar exports levelled off at better than prewar volume, and the continent could afford large purchases of equipment. Many countries created official development corporations to promote hydroelectric power, new lumbering and metallurgical activities, and general industrialization; and the indexes of industrial production showed gains of 50 percent or more as compared with the interbellum decades. There was a striking increase in the petroleum output of Venezuela, and that nation's vast iron-ore deposits were opened up. Agriculture and sanitation were improved, with technical cooperation from the United States and the UN. Fisheries were expanded, especially in Peru, and the crushing of vegetable oils was increased. (W.Ft.)

Development of resources and industries

Social and political changes. The economic expansion created by a growing demand for primary exports, the new services required as a result of this expansion, and the growing centralizing and coordinating role of the state all helped to transform South American urban society, particularly in the larger cities, and led to the emergence of an urban middle class and to the beginnings of a modern urban working class. These transformations were carried into the political system, leading to an enlargement of participation. The breakdown in international trade created by two world wars and the Great Depression generated conditions favouring industrialization. At the same time urban concentration increased enormously, while the growth of the mass media and of transportation networks broke down the isolation of the countryside and the large marginal areas that usually formed the greater part of the national territory. The timing, speed, and character of these transformations varied widely among the South American nations. In general, Euro-American countries initiated their transformation first, followed by the largest nations in Mestizo-America, and by some of the Indo-American states. Stagnation and involution (*i.e.*,



Countries of South America.

deterioration of already attained levels of modernization and development) have occurred, but mass mobilization and social change are the main characteristics of contemporary South America.

The emergence of new social sectors—first the urban middle classes and then the urban proletariat—deeply modified the cultural, social, and political scene. Nationalism, in different ways and degrees, characterized the outlook of the middle classes and of the populistic regimes that replaced the oligarchic regimes, which had been characterized by "limited" democracy, or (more often) undisguised authoritarianism. Military intervention continued, but its meaning changed. Prior to the 1930s it had represented an expression of factional conflict within the ruling elites, but instead it became a component of the "participation crises" generated by the entry of an increasingly larger proportion of the population into politics. Economic, political, and cultural "dependence" on foreign hegemonic nations—mainly the United States—became a central issue. A growing sense of national identity generated a revaluation of the Indian and mestizo heritage. Even where this element was virtually lacking—as in Euro-America—efforts were nevertheless made to re-create a national-popular tradition. Economic development and the integration of national society became universally accepted, but deep conflicts were created by the contrasting ideologies and interests of the various social sectors that were actively participating in the political arena. New regimes were established, varying from those oriented toward Marxism on the one hand to the modernizing authoritarianism of military rule on the other. The old colonial heritage was clearly disappearing. (G.Ge.)

Foreign relations and national identity

GEOGRAPHICAL FEATURES OF SPECIAL INTEREST

Prominent landforms

ANDES

A vast series of extremely high plateaus surmounted by even higher peaks, the Andes mountain system stretches for 5,500 miles (8,900 kilometres)—from the southern tip of South America to the continent's northernmost coast on the Caribbean. One of the great natural features of the globe, it separates a small western coastal area from the rest of the continent, affecting deeply the conditions of life within itself and in surrounding areas.

The Andes is not a single line of formidable peaks but a succession of parallel and transverse mountain ranges, or cordilleras, and of intervening plateaus and depressions. Distinct eastern and western ranges—respectively named the Cordillera Oriental and the Cordillera Occidental—are characteristic of most of the system. The western cordillera is within the great belt of crustal disturbances bordering the Pacific and is thus the home of both earthquakes and volcanoes.

The directional trend of both the cordilleras is generally south–north, but in several places the Cordillera Oriental bulges eastward to form either isolated peninsula-like ranges or such high intermontane plateau regions as the Altiplano (Spanish: "high plateau"), occupying adjoining parts of Argentina, Chile, Bolivia, and Peru.

Some historians believe the name Andes comes from the Quechua Indian *anti* ("east"); others suggest it is derived from the Quechua *anta* ("copper"). It is perhaps more reasonable to ascribe it to the *anta* of the older Aymara language, which connotes copper colour generally.

Physical features. *Physiography.* There is no universal agreement about the major south-to-north subdivisions of the Andes system. For practicality, the system is dealt with in this article in terms of six south-to-north subdivisions: Patagonian Andes, Central Andes, Peruvian Andes, Ecuadorian Andes, Colombian cordilleras, and Venezuelan Cordillera.

The Patagonian Andes begin on the mountainous Staten Island (3,700 feet [1,125 metres]) of the Tierra del Fuego archipelago, run to the west through Isla Grande, where the important ridges, including Darwin, Valdivieso, and Sorondo—all less than 7,900 feet high—then rise to the north beyond the Strait of Magellan. The physiography of this southernmost subdivision of the system is complicated by the presence of the independent Sierra de la Costa.

Numerous transverse and longitudinal depressions and breaches cut this wild and rugged portion of the Andean range, sometimes completely; many are occupied by ice fields, glaciers, rivers, lakes, or fjords. The crests of the mountains exceed 10,000 feet (Fitzroy reaching 11,070 feet [3,375 metres]) north to latitude 46° S but average only 6,500–8,400 feet from latitude 46° to 41° S, except for Tronador (11,657 feet [3,554 metres]). The line of permanent snow is found at 2,300 feet in Tierra del Fuego, 5,000 feet at Volcán Osorno (41° S), and 12,000 feet in Domuyo (36° 38′ S). A line of active volcanoes—including Yate, Corcovado, and Macá—occurs about 40° S to 45° S. Gigantic ice fields are located between Monte Fitzroy (in Chile, Cerro Chaltel) and Lago Buenos Aires at both sides of Baker Fjord; the Viedma, Upsala, and other glaciers are born there. Other notable features are the more than 50 lakes found south of 39° S. Those depressions free of water form fertile valleys called *vegas.* The passes are low and easy, reaching 6,000 feet at Pino Hachado, 3,800 feet at Lipinza, and 2,000 feet at Mua Mum.

aks in e Patanian ndes

On both sides, but more especially on the western side, the forests are untamed and magnificent, covering the mountains as high as the snow line, although at the higher altitudes toward the north the vegetation is smaller and less dense. Both Argentina and Chile have created national parks to preserve the area's natural beauty.

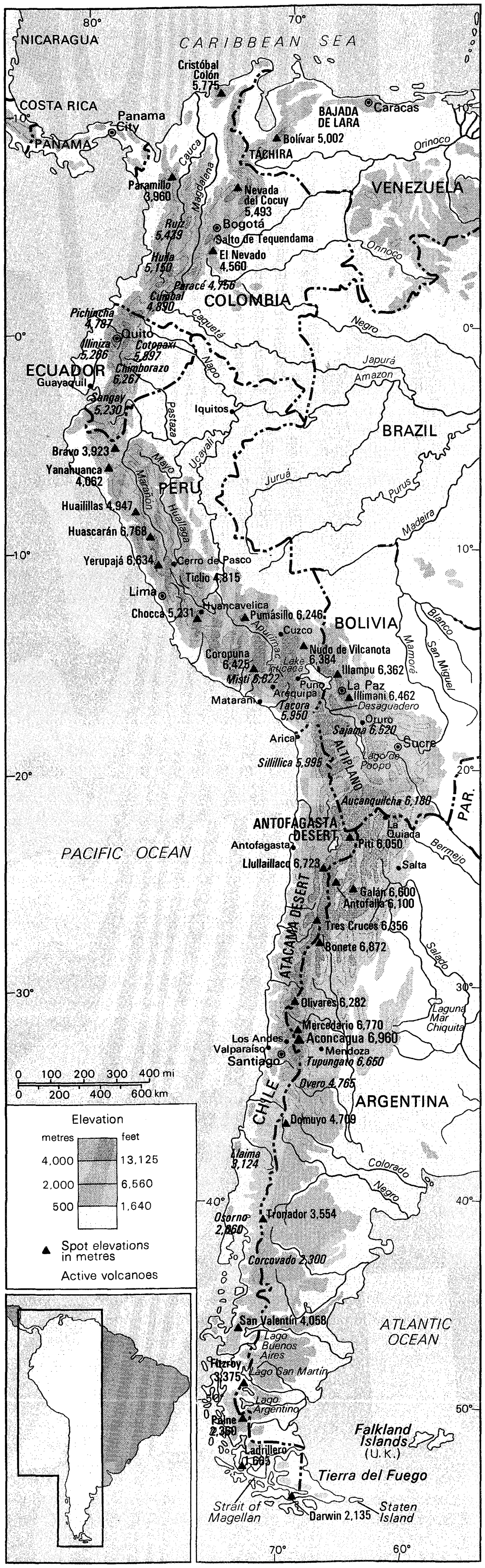
North of Lago Aluminé the axis of the cordillera shifts to the east up to a zone of transition between parallels 37° and 35° S, where the geographic aspect and geologic structure change. This marks the northern end of the Patagonian Andes.

The Central Andes begin at latitude 35° S where the cordillera undergoes a sharp change of character. Its width increases to about 50 miles, and it becomes arid and higher; the passes, too, are higher and more difficult. Glaciers are rare and very high, and there is virtually no vegetation. Lakes disappear north of latitude 39° S, forests, north of 37°. The main range serves as the boundary between Chile and Argentina and also is the divide between rivers flowing to the Pacific and the Atlantic. The last of the southern series of volcanoes, Tupungato (22,309 feet [6,800 metres]) is just east of Santiago, Chile. From there to the peak of Aconcagua, great snowcapped peaks, such as Juncal, rear up. The mighty Aconcagua itself, at 22,834 feet (6,960 metres), is the highest peak in the Western Hemisphere; to its north lies Mercedario (22,206 feet [6,770 metres]), and between them are the passes of Espinacito (16,000 feet) and Patos (12,825 feet). To the south the passes are Pircas (16,961 feet), Bermejo (over 10,000 feet), and Iglesia (13,400 feet). Farther north the passes are more numerous and higher, and the peaks of Bonete, Ojos del Salado, Incahuasi, and Pissis surpass 20,000 feet. The snow line is also much higher.

The peak of Tres Cruces (20,853 feet [6,356 metres]) at latitude 27° S marks the culmination of this part of the cordillera. To the north there is a transverse depression and the beginning of the high plateau called Puna in Argentina, Puna de Atacama in Chile, and Altiplano in Bolivia. The cordillera grows wider as it advances into Bolivia and Peru, where the great plateau is bounded by two ranges: the Occidental and the Oriental.

The beginning of the Altiplano

Northward, to latitude 18° S, the great peaks of El Cóndor, Sierra Nevada, Llullaillaco, Galán, and Antofalla all exceed 19,000 feet. The two main ranges and several volcanic secondary chains enclose depressions called *salares* because of the deposits of salts they contain; Sierra de Calalaste encompasses the big *salares* of Antofalla. Volcanoes of this zone occur mostly on a northerly line along the Cordillera Occidental as far as the Misti (latitude 16° S) in Peru.



Andes mountain system.

At about 18° S the Cordillera Occidental changes to a northwesterly direction to the west side, descending smoothly to the deserts of Atacama and Antofagasta and the southern Peruvian coast.

The eastern range, the Cordillera Oriental, lower and built on a broad bed of lava, is cut and denuded by rivers with steep gradients, fed by heavy rainfall. It has two sections. The southern is 150 miles wide and—with the exception of Chorolque (18,414 feet [5,614 metres])—not very high. The northern, named Cordillera Real, is narrow, with higher peaks and glaciers; and the most important peaks, at over 21,000 feet, are Illimani and Illampu (Nevado de Sorata).

At about latitude 22° S, the Cordillera Oriental penetrates into Bolivia, describing a wide semicircle to the north; the plateau becomes the broadest part of the Andes—the Altiplano de Bolivia.

The Altiplano is one of the largest interior basins of the world—500 miles long, 80 miles wide, and 11,200 to 12,800 feet high; it has no water outlet to the ocean. In about the middle of the plateau there is a great depression, running south to north. Lake Titicaca, the highest navigable lake of the world (110 miles long), is on the northernmost part of the depression; the Río Desaguadero follows the depression, taking the water of Titicaca to the small Lago de Poopó.

Canyons and passes of the Peruvian Andes

The Cordillera Occidental runs parallel to the coast of Peru, while the Cordillera Real from Bolivia ends in the rough mountain mass of Nudo de Vilcanota at latitude 15° S. From there, two lofty and narrow chains emerge northward, the cordilleras of Carabaya and Vilcanota, separated by a deep gorge; to the west of these and northwest from the city of Cuzco a third range appears, the Cordillera de Vilcabamba. The three ranges are products of erosive action of rivers that have cut deep canyons. West of the Vilcabambas, the Río Apurímac runs in one of the deepest canyons of the hemisphere. In the valley west of the Cordillera de Vilcanota stands the city of Cuzco (10,909 feet [3,326 metres]).

Traditionally, the Peruvian Andes have been described as three cordilleras, which join at Nudo de Vilcanota, Pasco, and Loja (Ecuador). Today, the region is considered as a simple plateau. The *nudos* (knots) are erosive products and hydrographic divides; the cordilleras stand on the plateau and were produced by crustal movements and erosion by the large rivers. The plateau in the south is 60 miles wide and 13,000 feet high; in the north it is narrower and lower (less than 12,500 feet).

The Nudo de Pasco is a large, high plateau. To the west it is bounded by the Cordillera de Huarochirí, on the west slope of which the Río Rímac rises in a cluster of lakes fed by glaciers and descends rapidly to the ocean (15,700 feet in 60 miles). The nearly 15,800-foot-high pass of Ticlio is used by a railroad. On the *nudo* there are many small lakes and ponds, the lake of Junín, about 20 miles long, being the largest.

Cordilleras Blanca and Negra

North of the Nudo de Pasco three different ranges run on the plateau: the Cordilleras Occidental, Central, and Oriental. In the Cordillera Occidental at latitude 10° S a deep, narrow valley (Callejón de Huaylas) separates two ranges—Cordillera Blanca (eastern) and Cordillera Negra (western); the Río Santa runs between them and cuts Cordillera Negra to drain into the Pacific. Cordillera Blanca is a complex highland with permanently snow-capped peaks, some among the highest of the Andes (for example, Huascarán, which is just over 22,204 feet [6,768 metres]). The glaciers that rise here are very often broken off by earthquakes and rush down the slopes, demolishing settlements and vegetation alike. Cordillera Negra, so named because it is without snow, is lower. The two ranges join together at latitude 9° S. The Río Marañón, which runs northward between Cordilleras Occidental and Central at about latitude 6° S, changes its direction to the northeast, penetrating into the region of the *pongos* (narrow transverse gorges that cut the cordillera to reach the Amazon Basin). These include Pongo de Rentema (about one and one-fourth miles long, 200 feet wide), Mayo, Mayasito, Huarcaya, and, the most important, Pongo de Manseriche (seven miles long).

Between Cordillera Central and Cordillera Oriental, the Río Huallaga runs in a deep gorge with few small valleys; it cuts the eastern cordillera in Pongo de Aguirre (latitude 6° S). Cordillera Oriental ends in the Amazon Basin at latitude 5° S.

The permanent snow line is 19,000 feet in Nevado Chanchami (about latitude 16° S) and declines to 15,000 feet in Cordillera Blanca and to 13,000 feet in Huascarán. Permanent snow disappears north of latitude 8° S, the puna ends and the so-called humid puna, or *jalca,* begins. Mountains become wider and smoother in appearance; vegetation changes to heathland and trees. The altitude diminishes, and passes are much lower, as at Cuello de Porcullo (7,000 feet).

:hains of olcanoes ı the :cua-orian ındes

A rough and eroded high mass of mountains named Nudo de Loja (latitude 4° S) marks the end of the Peruvian cordillera and the beginning of the Ecuadorian system. This is a large plateau running in a south to north direction and bordered by two chains of high and numerous volcanoes. To the west, in the Cordillera Occidental, geologically recent and relatively low, there is a line of 19 volcanoes, seven of them over 15,000 feet high.

The eastern border is Cordillera Central (formerly Cordillera Oriental), higher and older, with a line of 20 volcanoes. Five are over 16,000 feet.

The plateau in between (sometimes called Avenida de los Volcanes) is 8,000 feet high and 55 miles wide and is covered by volcanic rocks, andesites, and lava. In some places the accumulation of material forms *nudos* and secondary chains that divide the valley into several basins.

On the external slopes of the cordillera the rainfall is heavy, and vegetation flourishes. The high cordillera is cold and desolate, with peaks covered by mist; such regions are known as *páramos.* The central plateau is partially cultivated and supports the majority of the Ecuadorian population. In the transverse ranges there are some important peaks, including El Relado (13,642 feet [4,159 metres]), Imbabura (15,190 feet [4,630 metres]), and Ñaupán (14,852 feet [4,529 metres]).

'he 'ordillera 'riental

A third cordillera has been discovered recently in the eastern jungle of Ecuador; this is the Cordillera Oriental. It seems to be a very old alluvial formation that has been divided by rivers and heavy rainfall into a number of mountain masses, isolated or forming irregular short chains, and covered by luxuriant forest, like the cordilleras of Guacamayo, Galeras, and Lumbaquí. The altitude does not exceed 7,900 feet, except at Cordilleras del Cóndor (13,000 feet) and Cerro Pax (11,000 feet).

North of the international boundary there is a group of high, snowcapped volcanoes (Azufral, Cumbal, Chiles) known as Nudo de Huaca. Farther north begins the great Macizo Central Colombiano, with the volcanic mass of Nudo de Pasto (latitude 1–2° N), which is the most important Colombian physiographic complex and the centre of the country's river system.

Three distinct ranges, the Cordilleras Occidental, Central, and Oriental, run northward. Cordillera Occidental, parallel to the coast and moderately high (12,800 feet near Antioquia), ends in the mountain complex of the Cordillera de Chocó.

Cordillera Central, or Cordillera del Quindío, is the highest (mean altitude almost 10,000 feet) but the shortest (400 miles) range of Colombian Andes; its last spurs disappear south of the confluence of Cauca and Magdalena (latitude 8° N). Most of the volcanoes of the zone are in this range, including Tolima (17,105 feet [5,215 metres]) and Ruiz (17,844 feet [5,439 metres]) among others; Nevado del Huila (16,896 feet [5,150 metres]) is also in the Cordillera Central. At about latitude 5°80′ N, the range widens into the plateau of Antioquia, where Medellín stands.

Between Cordillera Central and Cordillera Occidental is a great depression, the Patía–Cauca Valley, divided into three longitudinal plains. The southernmost is the narrow valley of Patía that drains to the Pacific. The middle plain is the highest (8,200 feet) and constitutes the divide of the other two. The northern plain, the largest (15 miles wide and 125 miles long), is the valley of Cauca, which drains by the Rió Cauca to the Magdalena.

Páramos and *sabanas*

Cordillera Oriental has a slight northeast direction and is the widest and the longest of the system. The average altitude is 7,900 to 8,900 feet. North of parallel 3° N it widens, and after the depression of Uribe (6,130 feet) begins the Macizo de Sumapaz (altitude 10,000 to 13,000 feet), where the large Páramo (high plateau) de Sumapaz is located. The volcanoes Puracé, Picos de la Fragua, and Neiva stand south of this *páramo,* the first in the Cordillera Central, close to Popayán, the other two on the Cordillera Oriental. North of Sumapaz the range divides into two, enclosing a large plain 125 miles wide and 200 miles long, often interrupted by small transverse chains that form several peneplains, called *sabanas* (or savannas), commonly containing shallow lagoons and useless bogs. One-third of the population of Colombia lives there. The largest and most populated is the Sabana de Bogotá; other important *sabanas* are: Ubaté, Chiquinquirá, Tunja, and Sogamoso. East of Honda (latitude 5° N) the cordillera is a series of abrupt parallel chains running to the north-northeast: among them Sierra Nevada del Cocuy (18,000 feet) has snowed peaks. Farther north the central ranges end, but the external chains go on and diverge to the north and northeast; the western is Cordillera de Ocaña (6,500 feet), which on its northeastern side has the sierras of Motilones and Perijá; this is the boundary between Colombia and Venezuela and goes as far as parallel 11° N in the Península de la Guajira. The eastern chain bends to the east and penetrates to Venezuela at the Cordillera de Mérida.

The northern end of the Andes

Rivers on the Cordillera Oriental make their ways on the *páramos* and *sabanas* and cut the borders of the plateau through deep gorges; Salto de Tequendama of the Río Bogotá is an example.

On the border of the Caribbean, between Serranía de Perijá and Isla de la Aguja, stands Sierra Nevada de Santa Marta, an isolated triangular massif that rises abruptly from the coast to snowcapped peaks of 18,947 feet (5,775 metres). Geologically speaking, it does not belong to the Andes.

The Andes of Venezuela are represented by Cordillera de Mérida (280 miles long, 50 to 90 miles wide and about 10,000 feet high); it is separated from Cordillera Oriental of Colombia by the depression of Táchira. It begins in Pico el Cobre, extends first in a northerly direction for 35 miles and then northeast, up to the depression of Barquisimeto, or Bajada de Lara, where it ends. The cordillera is a great uplifted axis (where erosion has uncovered granite and gneiss rocks but where the northwest and southeast sides remain covered by sediments) formed by numerous chains with snow-covered summits separated by longitudinal and transverse depressions—Sierras Tovar, Nevada, Santo Domingo, de la Culata, Trujillo, and others. The Cordillera de Mérida separates the hydrographic regions of the Orinoco and the Caribbean.

To the north, Sierra Falcón and Cordillera del Litoral (called in Venezuela Sistema Andino) do not belong to the Andes but rather to the Guiana system.

Geology. The eastern side of the Andes is in general older than the western. Paleozoic (Cambrian, Silurian, Devonian, and Carboniferous) beds formed from 280,000,000 to over 570,000,000 years ago are found in the eastern Argentinean and Bolivian Andes; Cordilleras Real and Vilcabamba have even more ancient Archean (late Precambrian) as well as Paleozoic rocks, related to the Brazilian Massif. In Peru and Ecuador, the eastern ranges are of ancient gneiss and schists, covered mostly by Mesozoic, or Secondary, sandstones (from about 65,000,000 to 225,000,000 years old). In Colombia, the Cordillera Central shows ancient crystalline rocks (Cambrian to Carboniferous) with intrusion of heat-altered rocks.

The western ranges and the plateaus are mainly of Mesozoic formation, Jurassic and Cretaceous. Rocks of Cretaceous age (up to 136,000,000 years old) are widely distributed in the Patagonian Andes; in Chile there are Jurassic (up to 190,000,000 years old) and Cretaceous sediments. Puna de Atacama and the west ranges of Bolivia show Mesozoic beds with Cenozoic (Tertiary; 7,000,000 to 65,000,000 years old) volcanic material. In Peru, gray limestone covers large parts of the plateau and the Cordillera Occidental. In Ecuador the Mesozoic beds occur with deposits of volcanic material in the valley. In

Colombia the western branch is mostly of Cretaceous beds, as is the Cordillera Oriental.

The entire western coast of South America is included in the great circum-Pacific belt of seismic activity (earthquakes and volcanoes). The great geologic dislocation and the folding of the cordillera started in Cretaceous (Late Mesozoic) time and continued during the Tertiary (early to mid-Cenozoic). During the Mesozoic, the western part of South America was a geosynclinal basin, a vast downwarp in the Earth's surface, covered by the Pacific. Periods of upliftings and subsidences during the Tertiary were accompanied by great volcanic effusions. In the Quaternary (Late Cenozoic), beginning some 2,500,000 years ago, the disturbances continued in a process that is still going on. This involves a continuous raising of the cordilleras and a sinking in the ocean of a great part of the coast, producing an unstable condition caused by the tremendous abrupt difference between the heights of the cordilleras and the great depths of the Pacific. In this geologically recent period, volcanic activity and glaciation produced a period of uniform summit levelling (peneplanation) and denudation by filling depressions with glacial deposits and volcanic material and eroding the overall mountain relief. The modern volcanic system is Quaternary; three main groups were formed: that of the Patagonian and Central Andes (Chile); that between Chile and Peru; and that between Ecuador and Colombia.

Volcanic and earthquake activity in the Quaternary

Soils. The complex interchange between climate, geology, topography, and biology that determines soil types and their condition is deeply affected by altitude. In general, Andean soils are relatively young and suffer great erosion by water and winds because of the unimpeded steep gradients of the land.

In the deep south of the Patagonian Andes the formation of soils is difficult; the action of glaciers and winds has left almost-bare rocks in many places. Northerly peat bogs, podzols, and meadow soils, yellow-brown and red-brown, with thick horizons of humus, are found; drainage is poor. In the region of lakes, dark-brown soils occur. They are rich in organic material and well drained. North of latitude 45° S drainage becomes insufficient in surface as well as in profile. In the upper part of the cordillera, soils are formed directly on weathered rocks; in the lower zones reddish-brown soils with gravel and quartz are found; erosion is heavy.

North of parallel 37° S are found primary soils with water deficiency. In the Atacama Desert, where precipitation and vegetation are scarce, so is the organic material; desertic gray soils bearing mineral salts occur, heavily eroded. This type, with little differences, extends on Cordillera Occidental to north of Peru.

The plateau and the east side of the eastern cordilleras show characteristics of soil closely related to altitude, from Bolivia to Colombia.

In the Andean *páramo* are found rough topography and embryonic soils black with organic material; drainage ranges from very poor to very rapid.

At altitudes between 6,000 and 12,000 feet, red, brown, and chernozem soils occur in soft slopes of valleys and plains. In more poorly drained areas, soils are relatively fertile, being permeable in some places with sandy horizon B (the main layer of the soil profile, extending from the surface down to the bedrock). These soils are economically the most important in Bolivia, Peru, and Ecuador. The *sabana* soils of Colombia are gray-brown, with claypan in certain levels, which makes them impermeable, of very poor drainage.

In the high parts of the mountains are very poor soils, superficial and stony, as represented by Andean rendzina (dark grayish-brown soils) and by brown-black soils. On the east side of the eastern cordilleras, descending to the Amazon Basin, there are thin, dark, humid soils without evolution, suffering great erosion because of heavy rainfall and the steep gradient of the slopes.

The intrazonal soils include humic clay and solonetz (dark alkaline soils) close to the lakes and lagoons. Also included in this group are soils formed from volcanic ash in Cordillera Occidental, south of latitude 14° S (Chile, Peru, and Bolivia).

The azonal soils are alluvials, of incomplete evolution and stratified without definite profile, and lithosols (a group of shallow soils consisting of imperfectly weathered rock fragments) that occupy the major part of the Andean massif. Their superficial layer is directly on rocks. They have incomplete profiles on rough slopes. Erosion is heavy. In Colombia, in slopes and gorges, sandy yellow-brown azonal soils are the base of the large coffee plantations.

Climate. In general, temperature increases from Patagonia to the Equator, but combinations of such factors as altitude, proximity to the sea, the cold Peru (Humboldt) Current, heavy or scarce rainfall, and defenses against wind determine an extensive variety of climates. The hottest rain forests, deserts, or deep gorges are often separated from the polar temperatures of the high puna by a few miles. At identical altitudes, however, there is no parallelism between external (to the Pacific or to the Amazon Basin) and internal slopes of the cordilleras; the external slopes are under the influence of either the ocean or the Amazon Basin.

The isotherm (line of equal average temperatures) 10° C (50° F) extends from the Patagonian Andes to Colombia. As mentioned above, the line of permanent snow varies greatly. It is 2,600 feet at the Strait of Magellan, 7,000 feet at 39° S, 20,000 feet at 27° S, 19,000 feet at 16° S, 16,000 feet at 9° S, and 15,000 feet at the tropical Andes.

Precipitation varies widely. There is heavy rainfall and solid precipitation south of parallel 38° S (180 inches at Lake Quillen). On the Central Andes of Chile and the Atacama Desert, rainfall is scarce, with some precipitation in summer (December, January, February). On the Altiplano de Bolivia and the Peruvian plateau and in the valleys of Ecuador and *sabanas* of Colombia, rainfall is seasonal and not heavy; in some places there is very little. It rains in small amounts on the west side of the Peruvian Cordillera Occidental, somewhat more in Ecuador and Colombia. On the east side of the Cordilleras Orientales (Amazon Basin side), rainfall is usually seasonal and heavy.

Precipitation and temperature

Temperature varies with altitude. Up to 4,900 and 8,200 feet the climate is subtropical and tropical, respectively: dry with hot temperatures during the day and mildly warm at night. At heights of between 8,200 and 11,500 feet it is dry, with mild temperatures; but there are big differences between night and day; this realm constitutes the most populated area of the Andes. At 11,500 to 14,800 feet it is dry and cold, with great differences between day and night and between sunshine and shadow, and it is freezing at night. At 13,500 to 15,700 feet (the puna), occurs the climate of the *páramo,* very cold and dry, with subfreezing temperatures. Above 15,700 feet on the peaks and high ridges, climate is polar with very low temperatures and ice-cold winds.

As in all other areas of the world, microclimates (extremely local developments) are widely varied because of the interplay of latitude, altitude, and other factors.

Plant and animal life. Life on the Andes depends upon altitude; within certain levels, however, the existence of vegetation is determined by climate, watering, and soil and of animal life by the food that is available. The permanent snow line is the upper limit of both. In the Andes, some plants and animals can live at any altitude, and others can only live at certain levels. Cats live only exceptionally above 13,000 feet; white-tailed mice usually do not stay lower than 13,000 feet and live up to 17,000 feet. The Peruvian camelids (llama, guanaco, alpaca, and vicuña) are animals of the Altiplano (11,200 to 12,800 feet), although they can live well at lower altitudes. The condor is believed to fly up to 26,000 feet. Probably the low barometric pressures of high altitudes are less important for vegetation, but altitude imposes a number of climatic factors, such as temperature, wind, radiation, and dryness of atmosphere, that determine the kind of plants growing at the different parts of the Andes.

Dependence of life on altitude

In general, the Andes can be divided into latitudinal bands, each with typical main vegetation and fauna; but latitude imposes differences between south and north; and proximity to the Pacific and to the Amazon Basin creates differences between the external and internal slopes of the Cordilleras Occidental and Oriental.

A zone at about parallel 35° S separates two different regions of the Andes. To the south, in the Patagonian Andes, the flora is austral (of southern aspect) instead of Andean. Magnificent forests of araucaria (a pine), oak, coigue (an evergreen used for thatching), chusquea, cypress, and larch occur.

Northerly characteristics are different. The Cordillera Occidental is very dry in the south, slightly humid (with moisture and scarce rainfall) in central and north Peru, and humid with heavy or moderate rainfall in Ecuador and Colombia. Vegetation follows the climatic scheme: in the south, it is poor and desertic; at higher altitudes, steppe vegetation occurs. Animals include the guemul, puma, vizcacha, cuy (guinea pig), chinchilla, camelids, mice, and lizards; birds include the condor, partridge, parina, huallata, and coot. Agriculture is poor. The east side of the Cordilleras Orientales northward from Bolivia has lush vegetation, most of it natural tropical forest; agriculture is difficult because of the steepness of the slopes, but the rich fauna of the jungle is present.

'lora and
ıuna of
ıe plateau

On the plateau (valleys, plains, ranges, and internal slopes of the cordilleras), life is again closely related to altitude. Tropical palms and eternal snows lie within a few miles of each other; altitude varies from 1,600 feet in deep gorges to more than 20,000 feet in peaks and ridges. Up to 8,000 feet, vegetation is of dry tropical and subtropical climate, and agricultture is important: the great coffee industry of Colombia is mainly located in the warm valleys of this zone. Between 8,200 and 11,500 feet lies the most populated zone of the Andes; some of the major cities of the Andean countries are there, and it supports the main part of Andean agriculture. Temperatures vary from warm in the valleys to moderate low (down to 50° F [10° C]) on the plains, *sabanas,* and slopes; and there is seasonal rainfall and water from rivers. This zone is also suitable for livestock and poultry.

At 11,500 to 13,400 feet relief is usually rough and difficult for agriculture; in Colombia this zone is *páramo* and sub-*páramo,* with seasonal rainfall; in Ecuador rain is abundant; in Peru and Bolivia the *páramo* has from moderate to scarce rainfall.

At 13,400 to 15,700 feet (puna) vegetation is formed by plants that resist the cold temperature and the freezing during the night; above 16,000 feet, vegetation is almost absent.

The people. Man's presence on the Andes is relatively recent; the oldest human bones found have been 10,000 to 12,000 years old, although human habitation may be much older. The lack of oxygen due to the altitude, especially above 12,000 feet, is so physiologically disturbing that it imposes deep adaptative changes even in the cells of the body. The highest altitude in the Andes at which man has been found as a permanent resident is 17,100 feet (shepherds in the south of Peru) and, as a temporary worker, 18,500 to 19,000 feet (Carrasco Mine, in Aucanquilcha, Atacama Desert, Chile).

From Patagonia to the southern limits of the Bolivian Altiplano, the Andes are unpopulated; a few small groups of shepherds and farmers live in the lower slopes and *vegas* of the cordillera.

Northward, the Andes hold the largest part of the population and the majority of the most important cities of Bolivia, Peru, Ecuador, and Colombia. The main populated area is localized between 8,200 and 11,500 feet. In Colombia 11,500 feet is the upper limit of human life, and above 11,500 feet population is sparse; whereas in Peru and Bolivia, most of the inhabitants live above 10,000 feet (4,000,000 Peruvians live between 10,000 and 13,000 feet, and 800,000 live above 13,000 feet).

ıdians
ıd
estizos

Half of the population of Bolivia are Aymara and Quechua Indians; the other half are mestizos (or cholos, mixed); in the Lake Titicaca district live the small and primitive group of the Urus. Population is mainly distributed between the high *páramos,* where, except for a seminomad population of shepherds, the principal occupation is mining, and lower narrow valleys, where the people engage in agriculture.

In Peru, mining is the most important human activity above 11,500 feet, but the great part of the Andean population is dedicated to agriculture and raising sheep, goats, llamas, and alpacas; more and more people are engaged in industry and commerce. In the south, around Lake Titicaca, there is a group of Aymara Indians; but the largest native population is Quechua.

The inhabitants of the Andes of Ecuador are mainly Quechuas and mestizos; in the south there are small groups of Cañaris and, in the north, of Aymara descendants (the Salasacas). Agriculture is the main occupation; some native groups are dedicated to ceramics and weaving.

In Colombia the largest part of the population lives between 5,000 and 10,500 feet. Only 7 percent are native Indians: a nucleus of Chibchas reside on the Altiplano of the Cordillera Oriental; Paeces and Guambianos on the Cordillera Central; Pastos and Quillacingas on the southern mountains. At about 3,000 to 6,500 feet, the zone of coffee plantations is the most densely populated (259 inhabitants per square mile). Agriculture is the main occupation, followed by cattle raising.

Agricultural patterns

The economy. *Agriculture and livestock.* Agricultural operations on the Andes are difficult, and the results are relatively poor. The water supply is inadequate; a large part of the plateau is dry or receives little and irregular seasonal rainfall. Temperatures of the high plains are too cold, and crops are subject to freezing. The terrain is rough, and soils are not well developed; and where fertile valleys do occur, they are narrow and small.

Some agricultural products of the Andes, however, can be grown in sufficient quantity to be exported: the main items are coffee (especially Colombia), cacao, coca, tobacco, cotton, and cinchona (quinine). There are possibilities of increasing the amount of arable land area by irrigation, but not to a great extent. The east side of the eastern cordilleras, facing the Amazon Basin, is in fact being developed, although its steep and rugged nature makes development difficult.

The natural pastures of the plateau are put to good use for cattle raising, and possibilities exist for further development of this activity. Colombia exports cattle, and Peru has a good milk-canning and livestock industry. Sheep, goat, llama, and alpaca raising are widespread in Peru and Bolivia, with both countries exporting sheep and alpaca wool.

Mining. The mining industry of the Andes is one of the most important of the world. Mining is especially extensive in the south. The industry includes mining for copper in Chile and Peru; tin in Bolivia; silver, lead, and zinc in Bolivia and Peru; gold in Peru, Ecuador, and Colombia; platinum and emeralds in Colombia; bismuth in Bolivia; vanadium in Peru; and coal and iron in Chile, Peru, and Colombia. Petroleum is extensively distributed along the whole eastern side of the Andes.

Transportation. The Andes are a formidable barrier for communication, with great effects on the economical and cultural development of the region. Production centres are, in general, far away from seaports, and the mountainous character of the land makes the construction and maintenance of railways and roads difficult and expensive. Most of the railways have been built to transport mining products. Agriculture requires roads because valleys are small and widely separated, making railways too expensive.

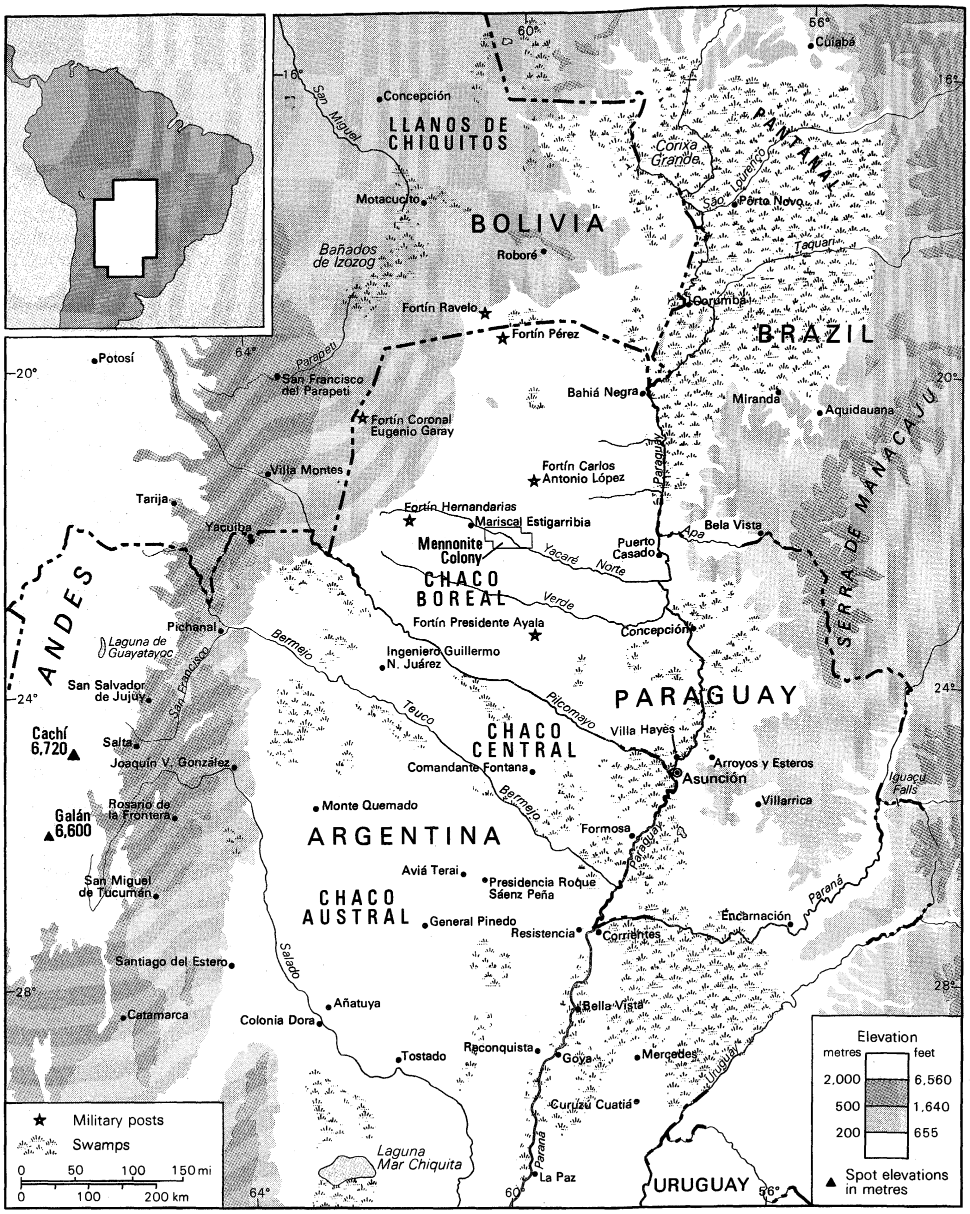
Pack animals

There is still a large use of pack trails between small communities and between farms and markets. Horses, donkeys, and mules are widely used; in Colombia the ox and in Peru and Bolivia the llama are also transport animals.

Railways are few and the internal systems little developed. There are two international railways between Chile and Argentina: the first, Los Andes–Mendoza, through the pass of Juncal, of narrow gauge, was opened in 1910 and connects Valparaíso and Buenos Aires; the second, Antofagasta–Salta (the former in Chile, the latter in Argentina), was opened in 1948.

Bolivia is connected with its neighbours by four railways: La Paz–Buenos Aires, crossing the Puna from Villagra to La Quiaca; La Paz–Oruro–Antofagasta, through Portezuelo de Ascotán; La Paz–Arica, crossing a point 14,000 feet high; and La Paz–Lake Titicaca, which continues on to Arequipa and Matarani (Peru).

In Peru there are two internal but important railways:



The Gran Chaco.

the Puno–Cuzco and the Lima–Cerro de Pasco and Huancavelica, the highest in the world, crossing Ticlio at an altitude of 15,797 feet (4,815 metres).

The main railroad in Ecuador is the Quito–Guayaquil and, in Colombia, the Ferrocarril del Atlántico, from Bogotá to the Caribbean coast, opened in 1961.

The construction of roads on the Andes is very active; thousands of miles are opened every year. The Pan-American Highway is built in the mountains in Colombia and Ecuador; it connects Caracas, Bogotá, and Quito. Various routes in Peru, Bolivia, Argentina, and Chile are included in the system.

Air transport is also important; it is especially developed in Colombia due to the difficulties of overland communication. (M.T.V.)

GRAN CHACO

The Gran Chaco is an immense lowland alluvial plain in interior south central South America. The name is of Quechua origin, meaning "hunting land." Largely uninhabited, the Chaco is an arid subtropical region of low forests and savannas traversed by only two rivers and practically unmarked by roads or rail lines. It is bounded on the west by the Andes mountains and on the east by the Paraguay and Paraná rivers. Its northern and southern boundaries are not as precise: its northern boundary is generally assumed to be the Llanos (high plains) de Chiquitos and Bañados (swamps) de Izozog in Bolivia, and its southern boundary the Río Salado in Argentina. Thus defined, the Gran Chaco covers about 280,000 square miles (730,000 square kilometres), of which slightly more than one-half lies within Argentina, one-third in Paraguay, and the remainder in Bolivia. Two great rivers, the Pilcomayo and the Bermejo, traverse the Chaco from their Andean headwaters to the Paraguay and Paraná rivers in the east. Regional divisions are the Chaco Boreal, north of the Pilcomayo; the Chaco Central, between the Pilcomayo and Bermejo; and the Chaco Austral, south of the Bermejo.

Area and boundaries

Physical features. *Physiography.* The Gran Chaco is the alluvial fill of a vast geosynclinal basin formed by downwarping or submergence of the area between the Andean Cordillera on the west and the Brazilian Shield on the east. Because of its alluvial character, the Gran Chaco is nearly stone free and is composed of extremely deep (up to 10,000 feet [3,048 metres]) unconsolidated sandy and silty sediments. The only rock outcrops of consequence are a few isolated remnants along the Paraguay River (in Paraguay) and some sandstone mesas in northern Paraguay and southern Bolivia.

Drainage. All but the extreme northwestern sector of the Gran Chaco is drained by west-bank tributaries of the Paraguay and Paraná rivers. The Bermejo (Teuco) and the Pilcomayo, even though they manage to traverse the Chaco, are still typical of most Chaco streams. Their courses are marked by countless sloughs, oxbow lakes, braided channels, sandbars, and vast swamplands; and they sustain such high losses from flooding, seepage, and evaporation that only a meagre portion of their full flow ever reaches the parent stream. Most of the Chaco is so poorly drained that the very shallow, irregular channels on the exceptionally level plain lead to rapid and extensive flooding during the rainy summers. At the peak of these floods, as much as 42,000 square miles, or 15 percent, of the area of the Chaco may be inundated, although some of this is as much due to improper drainage of the impermeable subsoils as to overflow of the streams. Potable groundwater supplies are poor. Saline water is common in both deep and shallow wells, and the location and maintenance of freshwater supplies is much a matter of chance. The problem appears to be greatest in the Chaco Boreal, although some hydrologists feel that large-scale studies might reveal a situation more like that of the remainder of the Chaco or like the Argentine Pampa, where groundwater problems are not now considered to be as severe as early settlers and explorers had postulated.

Climate. With a north–south extent of about 700 miles (1,125 kilometres), the Gran Chaco is subject to climates that vary from tropical in the north to warm temperate in the south. Most of the region is, however, subtropical. Average temperatures vary from 65° to 75° F (18° to 24° C), with an average humidity between 60 and 75 percent. Great temperature contrasts, however, exist. Average maximums are near 80° F (27° C), and absolute maximums may exceed 116° F (47° C). The average minimum is about 57° F (14° C), although freezing winter temperatures are known throughout the region.

The highest average annual rainfall is in the east, with 52 inches (1,320 millimetres), which gradually decreases to about 20 inches in the far west. Although the rainfall would normally be adequate for agriculture, from one-third to one-half the total comes in hot summer. Evaporation losses sharply reduce the effective precipitation and give the Chaco an arid nature lost only in the permanent swamps and forests along the Paraguay River.

Although light breezes are common, outbreaks of cool air from the south, called pamperos in Argentina, bring gusty winds of 32–45 miles per hour that occasionally exceed 60 miles per hour. The windiest season, however, is spring, during the transition from warm to hot weather. In the dry season dust storms are not uncommon.

Soils. Chaco soils range from sandy to heavy clay. Soils in the more humid east have more organic material and lateritic subsoils, whereas in the west are soils with less surface organic material and predominately calcareous subsoils. The local determining factor is drainage, whether a function of soil texture or relative relief. Sometimes elevation differences of less than three feet result in different soil types. Grasslands, or savannas, seem to be associated commonly with sandier soils, bushlands with poorly drained clay soils, and the forestland with better drained clay soils. In many cases the high concentration of dissolved salts in the soil water creates physiologically arid conditions in swampy sites, thus extending the arid appearance even into many areas with an abundance of water.

Plant life. The vegetation of the Gran Chaco is intimately associated with the pattern of soils and reflects the same general east–west division. The eastern Chaco is called Parque Chaqueño for its parkish landscape of clustered trees and shrubs interspersed with tall herbaceous savannas. To the west, a wide transition zone grades into the *espinal,* a dry forest of spiny, thorny shrubs and low trees. Chaco vegetation is adapted to grow under arid conditions and is highly varied and exceedingly complex. The climax vegetation is considered to be the *quebrachales,* the vast, low hardwood forests of which various species of quebracho tree are dominant and economically important as sources of tannin and lumber. These forests cover extensive areas away from the rivers; nearer the rivers they occupy the higher, better drained sites, giving rise to a landscape in which the forests appear like islands amid a sea of savanna grasses growing as high as a man on horseback. In the more arid western Chaco, thorn forests, the continuity of which is occasionally broken by palm groves, saline steppes, and savannas induced by fire or deforestation, are dominated by another quebracho tree that has a lower tannin content and is used most often for lumber. There is also a marked increase in the number and density of thorny species, among which the notorious vinal (*Prosopis ruscifolia*) was declared a national plague in Argentina because its thorns, up to a foot in length, created a livestock hazard in the agricultural lands it was invading.

ources of
nnin and
mber

Animal life. True to its native name, the Chaco has an abundance of wildlife. Among the larger animals are the jaguar, ocelot, puma, tapir, giant armadillo, spiny anteater, many foxes, numerous small wildcats, the agouti (a large rodent), the capybara (water hog), the red wolf, palustrian deer, the peccary, and the guanaco. It is one of the last major refuges for the South American member of the ostrich family, the rhea, or nandu, and has long been noted for its abundant and varied bird population. The streams are host to more than 400 fish species, among which are the salmon-like dorado and the aggressive piranha. Countless travellers' tales complain of the pestilent insects. Reptiles are also abundant, with numerous lizards and at least 60 known species of snakes, including many pit vipers and constrictors, while at least six species of poisonous tree toads have been identified.

The people. The indigenous tribes of the Chaco were numerous. Because of their subsistence as hunters, gatherers, and fishermen, tribal units were not much larger than extended families. Nevertheless, from among the diverse dialects, anthropologists have described a few major linguistic associations: the Guaycurú, Lengua, Mataco, Vilela, Zamuco, and Tupí. Most of these people lived under extremely primitive conditions, with settlement dictated by the availability of fresh water, making stream courses the most coveted sites. Implements were fashioned largely from wood and bones because of the absence of stones, while the spiny leaves of the pineapple-like groundcover *carraguatas* served as the universal fibre source. The harsh Chaco forest, quite surprisingly, contains more plant sources of human sustenance in the form of edible pods, fruits, berries, and tubers than surrounding areas, and this factor was well exploited by the native tribes. Game was gathered by trapping, netting, clubbing, and spearing, often in conjunction with large group drives. For those Indian groups still living outside the limits of European settlement, conditions are only slightly modified today, although these people now have domesticated animals and metal tools. Most tribes, however, exist as sort of a peasant pioneer fringe and practice some form of shifting subsistence agriculture.

Aside from the scattered, although successful, *reducciones* (agricultural communes) of the Jesuits, the Chaco defied effective European occupancy until well into the 19th century. Hostile Indian tribes, in concert with the forbidding nature of the Chaco itself, limited European influence in the colonial period to a situation much like a state of siege.

The
Mennonite
colonies

Since World War II, efforts have been made by the respective governments to spur colonization and settlement of the Chaco. Argentine efforts have concentrated along the railways out of Resistencia and Formosa, with pioneer settlements composed mainly of eastern European immigrants and based on cotton production. In the central Paraguayan Chaco, where the Trans-Chaco Road from Villa Hayes to Fortín Coronel Eugenio Garay was completed only in 1965, Mennonite immigrants from Canada had settled in the 1920s and were joined by co-religionists from the Soviet Union in the 1930s. These settlers established self-sufficient colonies and were joined by another large contingent of refugees from the Soviet Union after World War II. These colonies support a population of about 30,000. The primary land use in the Bolivian Chaco is still open cattle range. With nearby supplies of oil and gas in addition to hydroelectric and water storage on the fast-flowing piedmont streams such as the Pilcomayo, future development of the Bolivian Chaco may be more feasible than in some other parts of the Chaco.

The economy. The limited early colonization in the Argentine and Bolivian Chaco was based on exploitation of the longhorn criollo cattle that roamed half wild throughout the region. The western Chaco Austral, near Salta, was also exploited as a source of heavy timbers for the mines in the highlands of Bolivia and Peru. In the late 19th century, the Chaco in Argentina and southern Paraguay became a land of great ranches trading in the criollo cattle; and numerous, small, independent camps of woodcutters (*obrajes*) exploited the abundant hardwoods of the Chaco forests for lumber and firewood. Cattle grazing is still the most extensive land use, with few substantial changes from pioneer days. One of the key problems in improving industry is the apparent endemic nature of many serious cattle diseases and pests against which the criollo cattle have developed some immunity, whereas purebred cattle are fully susceptible.

The tannin processing industry

In the eastern Chaco, vast, highly capitalized industrial ventures established large plants to process the great quantities of tannin found in the various quebracho species. Unlike the *obrajes* of the woodcutters, these operations were large, centralized mills adjacent to rivers or rail lines from which the selective cutting of quebracho has proceeded at a systematic pace. The slow growth habits of the quebracho trees, however, may result in the demise of the tannin industry, as the pace of the harvest has greatly exceeded reforestation. It is uncertain if the relatively untouched Bolivian Chaco holds sufficient quebracho timber to offset diminished production in the exploited sectors or if synthetic products will supplant the present demand for quebracho tannin. Other forest products include lumber and heavy timbers from a variety of other species, firewood, and palo santo oil (holywood oil), from the wood of *Bulnesia sarmientii,* a tree found in the more arid portions of the Chaco.

Although wild cotton had been known in much of the Chaco since pre-Columbian times, it had never been raised as more than an agricultural curiosity until the 20th century. During World War I, with cotton prices at a peak, large areas in Chaco province of Argentina were turned over to cotton culture. Despite bad markets, insect plagues, and often poor weather in many years since that time, the crop area has increased to over 1,000,000 acres. Production methods are antiquated, however, with nearly total reliance on thousands of migrant labourers. The crop is used for both fibre and seed-oil production. Lesser quantities are found in the Paraguayan and Bolivian Chaco.

Discovery of oil in the Bolivian piedmont in the 1920s led within a decade to the disastrous Chaco War between Bolivia and Paraguay, whose leaders held hopes of finding more oil in the neighbouring Chaco Boreal. Paraguayan claims were eventually honoured but did not include any part of the oil-rich piedmont. Subsequent explorations have been disappointing, but hopes still exist for future discovery. (G.E.Ma./M.D.H.M.)

PATAGONIAN DESERT

The Patagonian Desert covers the greater part of the region of Patagonia, that is to say, nearly all of the southern portion of mainland Argentina. With an area of about 260,000 square miles (673,000 square kilometres), it is the largest desert in the Americas, extending from latitudes 37° S to 51° S. It is bounded, approximately, by Andean Patagonia (the southern extension of the Andes mountains) to the west; by the Río Colorado to the north, except where the desert extends beyond the river into the borderlands of Mendoza and La Pampa provinces; by the Atlantic Ocean to the east; and by the Río Coig to the south. The desert thus covers most of the provinces of Neuquén, Río Negro, Chubut, and Santa Cruz; some geographers also include the area north of the island of Tierra del Fuego, which lies at the southern tip of South America and is divided between Argentina and Chile.

Origin of the name

The name Patagonia is derived from the Patagones, as the original inhabitants were called by Spanish explorers in the 16th century. It appears that Ferdinand Magellan, the Portuguese navigator and first European to reconnoitre the coast in that era, coined that name because the natives, with their thick furs, bushy hair, and painted faces, reminded him of Patagon, a dog-headed monster of the Spanish 16th-century romance *Amadís de Gaula.* The phrase *Patagonia, tierra maldita* ("Patagonia, land accursed") is proverbial.

Physical features. *Physiography.* The desert covers the Patagonian tableland that extends from the Andes to the Atlantic Ocean. The general aspect of the tableland is one of vast steppelike (virtually treeless) plains, rising terrace fashion from high coastal cliffs to the foot of the Andes; but the true aspect of the plains is by no means as simple as such a general description would imply. Along the Río Negro, the land rises in a series of fairly level plains from about 300 feet (90 metres) at the coast to about 1,300 feet at the junction of the Limay and Neuquén rivers and 3,000 feet at the base of the Andes. The table land region rises to an altitude of 5,000 feet. South of the Río Negro, the plains are much more irregular; volcanic eruptions have occurred in this area down to fairly recent times. Basaltic sheets, apparently only recently cooled, cover the tableland east of lakes Buenos Aires and Pueyrredón. On the Chico and Santa Cruz rivers, the plains have spread to within about 50 miles (80 kilometres) of the coast and reach almost to the coast south of the Coig and Gallegos rivers. In places, basaltic massifs (mountainous masses) are the salient features of the landscape.

The coast consists largely of high cliffs separated from the sea by a narrow coastal plain. Thus, the plateaus are formed of horizontal strata, some of sedimentary rocks, others of lava flows. Areas of hilly land, composed of resistant crystalline rocks, stand above the plateaus.

Rivers and lakes

Drainage. The deep, wide valleys bordered by high cliffs that cut the tablelands from west to east are all beds of former rivers that flowed from the Andes to the Atlantic, but only a few now carry permanent streams of Andean origin (the Colorado, Negro, Chubut, Senguerr, Chico, and Santa Cruz). The majority either have intermittent streams such as the Shehuen, Coig, and Gallegos, which have their sources east of the Andes, or, like the Deseado, are completely dry, except for salt ponds in the deeper depressions, and so altered by the combined effect of wind and sand as to afford little surface evidence of the rivers that once flowed in them. They serve an important purpose in the collection of the scanty surface water. Alluvial beds of considerable depth but of poor development have been built up in them. The line of contact between the Patagonian tableland and the Patagonian Andes is marked by a chain of lakes found in glacier troughs or cirques, and "closed" in their lower parts by moraines (glacial deposits).

From Lago Nahuel Huapí northward, the lakes—except for Lago Lácar—still drain to the Atlantic. South of Lago Nahuel Huapí, all of the lakes except Viedma and Argentino now drain to the Pacific through deep canyons cut across the Andean cordillera.

Climate. Patagonia is influenced by the South Pacific air current, which brings humid winds from the ocean to the continent. The winds, however, lose their humidity as they blow over the west coast of South America and over the Andes and are thus quite dry when they reach the Patagonian Desert.

The two climatic zones

Climatologically, the Patagonian Desert can be divided into two main zones, the northern and the southern, by a line drawn from the western half of Neuquén province to a point just south of the Península Valdés in northeastern Chubut province.

The northern zone, semiarid, has annual mean temperatures between about 54° and 68° F (12° and 20° C); recorded maximum temperatures vary between about 106° and 113° F (41° and 45° C), and minimum temperatures vary between 12° and 23° F (−5° and −11° C). Sunshine, minimal along the coast, is most plentiful inland to the northwest. The rainfall varies from 3.5 to 17 inches (89 to 432 millimetres) a year. The prevailing winds, from the southwest, are dry, cold, and strong.

The climate of the southern zone of the Patagonian Desert is sharply distinct from the humid one of the Andean cordillera. In the north of this zone, the influence of the Atlantic is practically nonexistent—probably because of the heights of the coastal region (which reach 900 to 1,800 feet around the Golfo San Jorge)—although cold winds coming from the west and the cold Falkland Current both make themselves felt. In the southern part, which is practically peninsular, the Atlantic exerts some influence. The zone has a cold, dry climate, with temperatures that are higher along the coast than they are inland and with strong west winds. Mean annual temperatures range from 43° to 55° F (6° to 13° C), with maximum temperatures at about 108° F (42° C), and minimum temperatures between 16° and −27° F (−9° and −33° C). There are heavy snows in winter and frosts throughout the year; and spring and autumn are so short that summer and winter are the only seasons worth counting. Average annual precipitation (rain and snow) ranges between about 4.5 and eight inches, though as much as 18.5 inches has been recorded. In the central areas of the desert there is less precipitation, and there is more sunshine than on the coast or in the Andean cordillera.

Plant and animal life. Whereas the long, narrow strip of Patagonia's western border has a vegetation like that of the adjacent cordillera, the arid region of the Patagonian Desert comprises two zones, each with its own characteristic vegetation.

The northern zone includes most of Neuquén province, most of Río Negro province, and the northeasternmost part of Chubut. About 84,000,000 acres (34,000,000 hectares) in area, it consists of open bushland, covered with widely spaced thickets between about three and seven feet high. Grasses flourish in the sandy areas, while salt-tolerant grasses and shrubs predominate in the salt flats. In irrigated areas profitable crops can be grown; these include peaches, plums, damsons, almonds, apples, pears, olives, grapes, hops, dates, vegetables, aromatic plants and alfalfa. In the north, the zone merges into one of wooded steppe.

rops
f the
rigated
eas

The second zone, covering the southwesternmost parts of Neuquén and Río Negro provinces, three-quarters of Chubut, and nearly all Santa Cruz province, has an area of 116,000,000 acres. The vegetation is low and very sparse and needs almost no water.

Among the birds of the desert are herons and other waders; predators such as the shielded eagle, the sparrow hawk, and the chimango (beetle-eater); and the almost extinct ñandú, or choique, a flightless bird similar to the African ostrich but not so tall.

The typical marsupial (animal having a pouch for carrying young) of the desert is the comadreja (a member of the weasel family). Species of bats include a long-eared variety. Armadillos, pichis (small armadillos), foxes, ferrets, skunks, mountain cats, and pumas are to be found, as also are maras, or Patagonian hares, and different kinds of burrowing rodents, such as the vizcacha and the tucotuco. Of the larger mammals, the most interesting is the guanaco, a kind of llama, hunted almost to extermination. There are a number of poisonous snakes as well as tortoises and lizards.

Vinchucas (winged bugs), bloodsucker insects (transmitters of American trypanosomiasis, or Chagas' disease), scorpions, and 14 kinds of spiders (including one kind called *Mecysmanchenius,* not found elsewhere) are also to be encountered. The rivers and lakes are naturally poor in fish, but some have been stocked with salmon and trout. Sea fish, however, and crustaceans and mollusks are plentiful off the coast.

he
itagones

The people. The original inhabitants of Patagonia seem to have been Indians from Tierra del Fuego. The most ancient artifacts, such as harpoons, found in the caves along the Strait of Magellan suggest that these people were moving up the mainland coast about 5,100 years ago. Robust and very tall, they constituted the two principal ethnic groups of Patagones—the Puelche-Guennakin in the north, and the Chonik or Tehuelche in the south. The surviving descendants of these Patagones, namely the Yámanes and the Alacalufes, are few in number; a kindred people, the Chono, died out in recent times. The Spanish explorers found the Patagones living as nomadic hunters of guanaco or of ñandú.

Toward the end of the 16th century the Spaniards attempted to colonize the Patagonian coastal region to clear it of English pirates; but a Jesuit settlement on the Golfo San Matías came to nothing. In 1778, however, the English tried to settle on the same bay, and the Spaniards reacted by founding Patagonia's first two towns, San José and Viedma (originally named Nuestra Señora del Carmen). A Spanish settlement at Puerto Deseado lasted from 1780 to 1807; but three years later this region was again devoid of European settlement.

After Argentina became independent, an attempt was made to populate Patagonia to make it part of the national state. Immigration, however, was not massive, though people came for various reasons: some to exploit the economic resources, others (for instance, the Welsh) to enjoy religious or political liberties. By the late 20th century most of the immigrants were Chileans seeking temporary work rather than a fixed domicile. Such transitory immigrants form about 20 percent of the population. Apart from major concentrations at Comodoro Rivadavia and in the towns strung out along the upper valley of the Río Negro, Patagonia's sparse population is mostly rural.

The economy. *Natural resources and their exploitation.* Oilfields around Comodoro Rivadavia, Plaza Huincul, and Catriel are Patagonia's most conspicuous asset, apart from the great deposits of ore at Sierra Grande in Río Negro province, which supply all Argentina with iron. Río Negro province also has deposits of manganese, tungsten (or wolfram), fluorite (calcium fluoride, used as a flux in metallurgy), lead, and heavy spar (barite, the principal ore of barium); Neuquén province has deposits of copper and gold, vanadium (used to toughen steel), and zinc-lead ore; Chubut province has uranium and manganese in moderate quantities. There are also deposits of kaolin and gypsum.

Hydroelectric
development

To exploit the hydroelectric potential of the Neuquén and Limay rivers, a generating station was constructed to produce an estimated 1,650,000 kilowatts—as the centre of the El Chocón–Cerros Colorados complex. The project is also designed to irrigate the Comahue region, thus promoting farming and industry.

Transportation. Argentina's National Route No. 3 runs for more than 1,860 miles from Buenos Aires and Bahía Blanca through the Patagonian coastal region southward to Comodoro Rivadavia. The roads of the desert proper, however, are few and of poor quality. Four railroads run east–west from the coastal region; two, which reach the foothills of the Andean cordillera, are connected with the Bahía Blanca–Buenos Aires line. Air services are based chiefly on the towns of the coastal region.

The chief ports are Rawson, Deseado, Santa Cruz, and Río Gallegos (south of the Río Coig); San Antonio Este, Puerto Madryn, Camarones, and San Julián, all on protected bays; and Comodoro Rivadavia, an outlet for petroleum products. San Antonio Este and Puerto Madryn are being developed for overseas traffic. (E.F.G.D.)

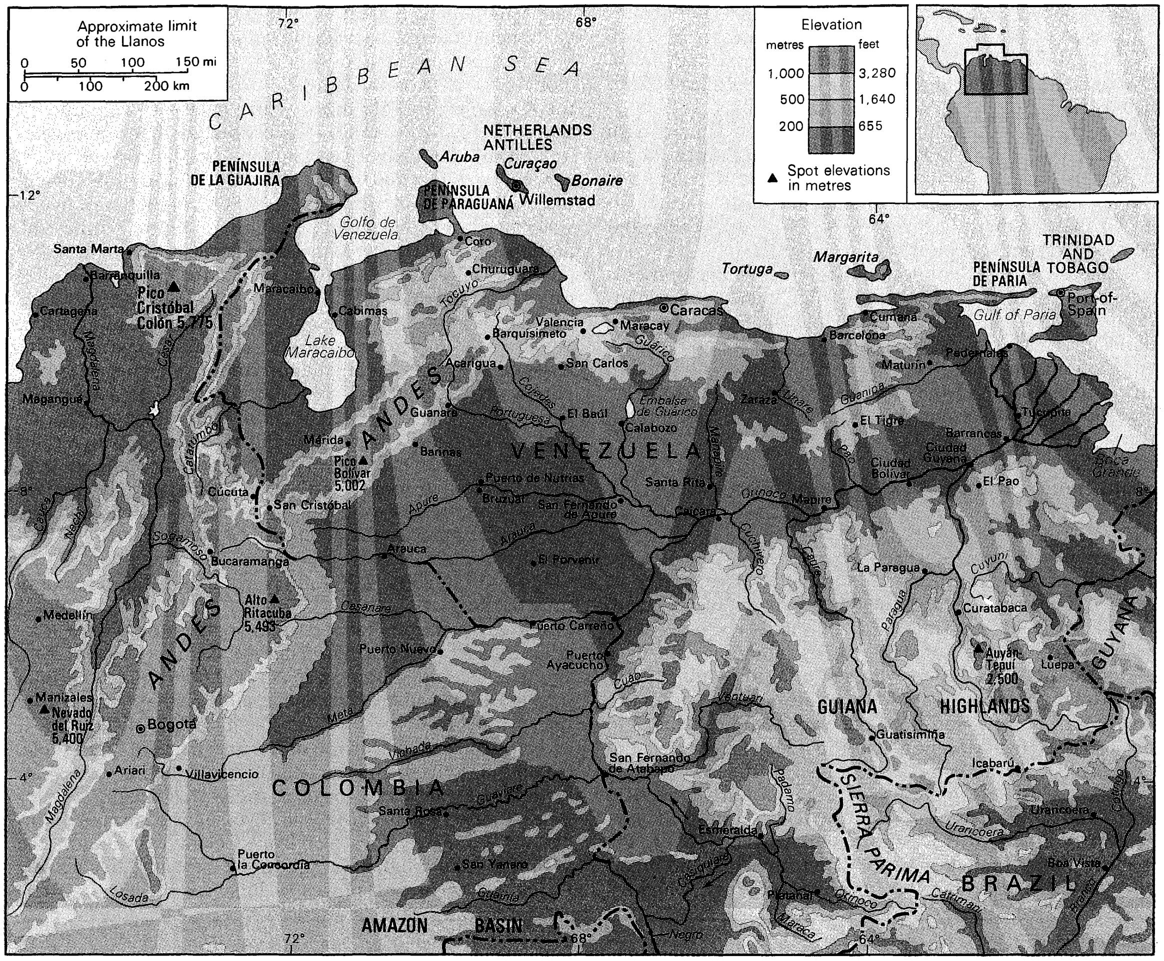
LLANOS

The Llanos (Spanish for "plains") is a grassland, or savanna, that stretches across northern South America, and occupies one-third of Venezuela and about one-fifth of Colombia. The Llanos are delimited by the Andes Mountains in the west and north, the Lower Orinoco River and the Guiana Highlands in the east, and the Río Guaviare and the Amazonian rain forest in the south. The region covers an area of some 220,000 square miles (570,000 square kilometres) and is comparable in size to France or Texas.

The savanna was named by the Spaniards in the 16th century and has been used as a vast cattle range since then. Until the mid-1900s, settlement was limited to widely scattered ranches known as *hatos* ("cow herds"), a few villages, and missionary stations along the lower courses of the region's rivers. Since the 1930s the region has experienced economic growth.

Physical
features

Physical features. *Physiography.* Most of the Llanos lies within 1,000 feet (305 metres) above sea level. The High Plains (Llanos Altos) are most conspicuous near the Andes, where they form extensive platforms between rivers and are some 100 to 200 feet above the valley floors. Away from the mountains they are increasingly fragmented, as in the dissected tableland of the central and eastern Venezuelan Llanos (the Sabana de Mesas)



The Llanos.

and the hill country south of the Río Meta in Colombia (the Serranía).

The Low Plains (Llanos Bajos) are defined by two rivers, the Río Apure in the north and the Río Meta in the south. The lowest portion of the Llanos is an area that lies to the west of the Orinoco Valley. This area is annually converted into an inland lake by the flooding of the region's rivers.

Soils and drainage. The Llanos are drained by the Orinoco River and its left-bank tributaries, including the Guaviare, Meta, Apure, and Cojedes rivers. Seasonal changes between saturation and dehydration have led to advanced laterization of the soil, the process in which the base minerals have been leached away or incorporated into insoluble iron and aluminum silicates. Fine-grained soils form hardpans (cemented layers of soils), and in gravel regions, iron-cemented quartz conglomerates underlie the surface. Excessive acidity and the lack of nutrient bases, organic matter, and nitrogen make virtually all mature soils infertile.

Climate. The wet and dry seasons result from the annual migration of the Intertropical Convergence Zone, a low-pressure trough between the hemispheric easterlies, or trade winds. The zone remains south of the Equator from December to March, bringing the entire Llanos under the influence of the Northeast Trades, which cause the dry and hot summer weather. The zone enters the Llanos from the south in April, reaches its northernmost position along the north coast in July, and moves south again until December. The passage of the zone brings the rainy winter period.

Monthly precipitation is seldom less than 10 inches (254 millimetres) in the Colombian Llanos between April and November. The rains peak about midyear in the Venezuelan north, with monthly totals of around 10 inches. Annual precipitation is highest near the Andes, where Villavicencio receives 180 inches; and there is a pronounced decrease toward the central plains, where Puerto de Nutrias receives 45 inches.

Mean daily temperatures are above 75° F (24° C) throughout the year, and the annual range does not surpass 7° F (4° C). Daily maximum temperatures rise above 95° F (35° C) in the dry period; the dry winds and nocturnal cooling bring relief with normal minimum temperatures between 65° and 75° F (18° and 24° C).

Plant and animal life. Most of the Llanos is treeless savanna. In the low-lying areas, swamp grasses and sedges are to be found, as also is bunchgrass (*Trachypogon*). Long-stemmed grass dominates the dry savanna and is mixed with carpet grass (*Axonopus affinis*), the only natural grass to provide green forage during the dry season.

The most conspicuous trees in the Llanos occur in the gallery forests along the rivers and in the narrower files of trees known as *morichales,* after the dominant moriche palm that follow minor water courses. Broadleaved evergreens originally occupied the high-rainfall zone in the Andean piedmont. There is also a handful of xerophytic trees (*i.e.,* those adapted to arid conditions), including the *chaparro* (scrub oak) and the dwarf palm scattered on the open savanna.

The Llanos have few animals. Most mammals nest in the gallery forests and feed on the grassland. The only true savanna dwellers are a few burrowing rodents and more than 20 species of birds (among them the white and scarlet ibis, the *morichal* oriole, and the burrowing owl). There are several species of deer and rabbit, the anteater and armadillo, the tapir, the jaguar, and the largest living rodent, the capybara. Crocodiles, caimans (a crocodile like amphibian), and snakes, including the boa constrictor, inhabit the rivers, which also teem with little-known varieties of fish. Insects include butterflies, beetles, ants, and mound-building termites.

The people and economy. Cattle raising remains the mainstay of the economy, the base of which was widened

The discovery of oil

by the discovery of petroleum in the 1930s. Oil strikes in the eastern and central Venezuelan Llanos at El Tigre (1937) and Barinas (1948) initiated industrial and urban development. Several of the "boom towns" of that period, such as El Tigre, have grown into sizable cities.

An expansion of intensive agriculture has occurred with the settlement, which began in the 1950s, of pioneer farmers in the Andean piedmont and along the river valleys. Major concentrations of these small farms are located in the Barinas–Guanare–Acarigua district of Venezuela and the Ariari region in Colombia.

Population growth connected with these developments has been impressive. There is a high degree of urbanization; more than half of the people of the Venezuelan Llanos live in cities of more than 10,000 inhabitants. A few thousand Indians (of Carib and Arawak origin) are left on reservations in the Lower Orinoco area.

Population increase has been modest in the Colombian areas. (Di.B.)

Prominent drainage systems and waterways

AMAZON RIVER

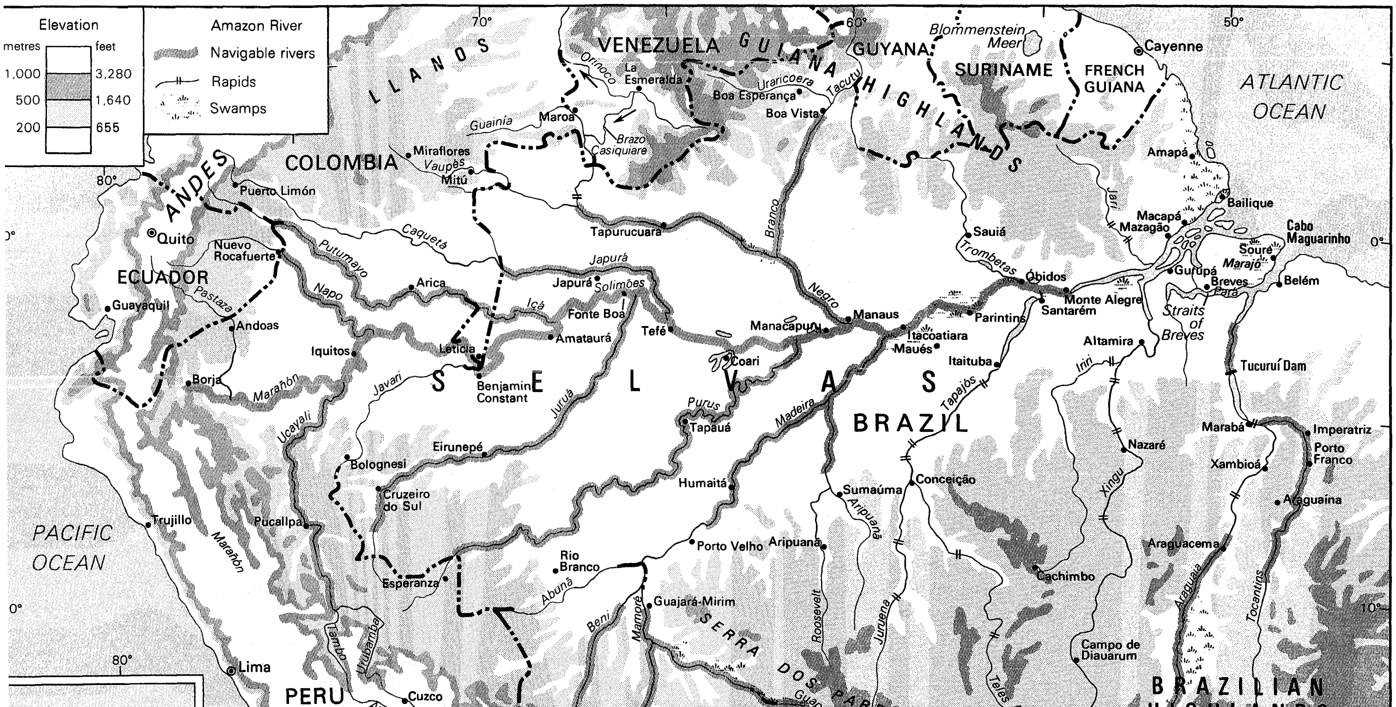
The Amazon (Portuguese and Spanish Amazonas) is the greatest river of South America and the largest river in the world in volume and in area of its drainage basin. Although the name Amazon is popularly employed for the whole of the main stream, in Peruvian and Brazilian nomenclature it is properly applied only to sections of it. In Peru, the upper stream (fed by numerous tributaries flowing down from sources in the high Andes) down to Iquitos, is called Marañón (Portuguese, Maranhão) and from there to the Atlantic, Amazonas. In Brazil the name Solimões is used from Iquitos to the mouth of the Río Negro and Amazonas only from the Río Negro to the sea.

The first European to explore the Amazon River, the Spanish soldier Francisco de Orellana, gave it its name after reporting pitched battles with tribes of female warriors. The length of the great river that he partially explored is about 4,000 miles (6,400 kilometres) slightly less than the Nile but still the equivalent of the distance from New York City to Rome. Its westernmost source is high in the Andes, within 100 miles of the Pacific Ocean; and its mouth is in the Atlantic. It is estimated that about one-fifth of all the water that runs off the Earth's surface is carried by the Amazon. At its mouth, 150 miles wide, there is a mean discharge of 170,000,000,000 gallons of water per hour into the ocean. This is some 10 times the amount carried by the Mississippi River, and it turns the ocean's water from salt to brackish for more than 100 miles from shore.

Physical features. *Physiography.* There are more than 1,000 known tributaries of the Amazon that flow into it from the Guiana Highlands and from the Brazilian Highlands, as well as from the Andes. Seven of these tributaries—the Japurá (Caquetá in Colombia), the Juruá, the Madeira, the Negro, the Purus, the Tocantins, and the Xingu—are more than 1,000 miles long; and one, the Madeira, is more than 2,000 miles from source to mouth. The largest oceangoing steamers can ascend the river 1,000 miles to the great city Manaus. Freighters and small passenger vessels drawing 18–20 feet of water can sail to Iquitos, some 2,300 miles from the Atlantic, at any time of the year.

The Amazon Valley

The vast Amazon Valley (Amazonia) is the largest lowland in Latin America. The drainage basin of the Amazon, about 2,722,000 square miles (7,050,000 square kilometres) in extent and nearly twice as large as that of any other river, stretches over 25 of the Earth's 180° of latitude. It includes the greater part of Brazil and parts of Venezuela, Colombia, Ecuador, Peru, and Bolivia. The Amazon Valley is not an immense swamp; most of Amazon country is terra firma, well above the level of the annual inundations. Going by the river from Belém to Manaus, the tidal forests of the delta are soon left behind; after leaving the Straits of Breves, ranges of hills several hundred feet



The Amazon River Basin and its drainage network.

high begin to appear in the north. At Monte Alegre the hills come down almost to the water's edge; beyond them extends a wide expanse of open *campos,* or grasslands, bounded by ridges along their northern edge. These natural *campos* are characteristic of the Amazonian landscape in many places—the largest of them, the *campos gerais* of the Rio Branco, far to the west, being several thousand square miles in extent.

The Amazon over the millennia has meandered freely over a wide flood plain, on which there appear a series of meander scars, oxbow lakes, and recently abandoned channels. Whenever the deposition of silt or alluvium is great enough to lessen the strength of the main current of the river, it will, during high water, overflow its temporary natural levee and carve out another channel. The new channel will in turn gradually silt up over a period of years or even decades, and the stream will then again shift its course.

The areas that are annually flooded, the *várzeas,* are rejuvenated each year by the deposit of fertile silt left as the waters recede. It is estimated that these valuable soils occupy some 25,000 square miles. (R.E.Cr.)

Geology and hydrography. The Amazon Valley is a great structural depression, a subsidence trough filled with Tertiary sediments, which flares out to its greatest dimension in the upper reaches. It lies between two old and not very high crystalline plateaus, the rugged Guiana Massif on the north and the lower Brazilian Massif, lying somewhat farther from the main river, on the south. The Amazon Valley was occupied by a great freshwater sea during the Pliocene Epoch. Sometime during the Pleistocene an outlet to the Atlantic was established, and the great river and its tributaries became deeply entrenched into the Pliocene surface.

The modern Amazon and its tributaries occupy a great drowned valley. With the rise in sea level that followed the melting of the ice caps, the steep-sided canyons that had been eroded into the Pliocene surface during the lower sea stand were flooded. The old depositional surface is the soil of the terra firma on which much of the Amazon forest has developed. In the upper part of the valley, in the Peruvian and Bolivian Oriente, more recent outwash from the Andes has covered many of the older surfaces.

At the Óbidos narrows, where the river is constricted to a width of a little more than a mile, the average depth of the channel below the mean water level is more than 200 feet (60 metres); in most of the Brazilian part of the river its depth exceeds 150 feet. Depths of more than 300 feet have been recorded at several points ascending the river from Belém. Yet, at the Peruvian border, 2,500 miles (4,000 kilometres) from the Atlantic, the elevation above sea level is less than 300 feet. The maximum free width (without islands) of the river's permanent bed is 8.5 miles, upstream from the mouth of the Xingu. During great floods, however, when it completely fills the floodplain, it spreads out in a band 35 miles wide or more. The average velocity of the Amazon is about 1.5 miles an hour, a speed that increases greatly at flood time.

The great river gradually rises from November to June and then falls until the end of October. The rise of the Negro branch does not occur at the same time, for the steady rains do not commence in its valley until February or March. By June the Negro is full, and then it begins to fall with the Amazon. The flood levels are, in places, from 40 to 50 feet above low river. Taking four roughly equidistant points, the rise at Iquitos is 20 feet, at Tefé 45, near Óbidos 35, and at Belém 12.

The Amazon's course

The Amazon does not meander in the fashion of the Mississippi but for the most part follows a remarkably straight course. It is still in the process of refilling its broad river-cut valley by its own alluvium, depositing silt into settling basins on a massive scale during each flood season. The size of the floodplain is moderate when compared with the volume of the river. The zone of active alluviation is typically 12–30 miles wide and is bordered by steep cliffs (*barrancas*), often capped by a horizon of laterite rock (*canga*). Where these cliffs are being undercut by the river, they produce the *terras caidas,* or "fallen lands," so often referred to by Amazon travellers.

The so-called black-water tributaries of the Amazon—the Xingu, Tapajós, Negro, Tefé, and Trombetas—carry little or no silt, in part because of the bleached sandy character of the country at their headwaters. The Tapajós and Xingu are in reality a bright emerald green, for they lack the strong solution of humic matter that characterizes the Negro. Where such streams enter the main river they are blocked off to form true freshwater lakes and resemble in form, width, and depth marine rias.

The first highland met in ascending the river is on the northern bank, opposite the mouth of the Xingu, and extends for about 150 miles up, as far as Monte Alegre. It is a series of steep, table-topped hills, in part of Devonian and Carboniferous age, interrupted by diabase intrusions. On the southern side, above the Xingu, a line of low bluffs extends in a series of gentle curves with hardly any breaks nearly to Santarém. The line is a considerable distance inland, bordering the floodplain, which is many miles wide. Then the bluffs bend to the southwest and, abutting upon the lower Tapajós, merge into the bluffs that form the terrace margin of that river valley. The next highland on the northern side is Óbidos, a bluff 56 feet above the river, backed by low hills. From Itaquatiara, nearly opposite the Madeira River, to near the mouth of the Negro, the banks are low. Approaching Manaus there are rolling hills; but from the Negro, for 600 miles, only very low land is found, resembling that at the mouth of the river.

On the southern side, from the Tapajós to the Madeira River, the banks are usually low, although two or three hills break the general monotony. From the latter river to the Ucayali, a distance of about 1,500 miles, the forested banks are just out of water and are inundated long before the river attains its maximum flood line. Thence to the Huallaga the elevation of the land is somewhat greater; but not until this river is passed and the Pongo de Manseriche approached does the swelling ground of the Andean foothills raise the country above flood level.

The width of the mouth of the river is usually measured from Cabo Norte to Ponta Tijoca, a distance of 207 statute miles (333 kilometres); but this figure includes the ocean outlet, 40 miles wide, of the Pará River, which should be deducted, because this stream is only the lower reach of the Tocantins. The figure also includes the ocean frontage of Marajó, an island about the size of the kingdom of Denmark, lying in the mouth of the Amazon. Following the coast, a little to the north of Cabo Norte, and for 100 miles along its Guiana margin up the Amazon, is a belt of half-submerged islands and shallow sandbanks. There the tidal phenomenon called the *pororoca,* or bore, occurs, where the soundings are not more than four fathoms (7.3 metres). It commences with a roar, constantly increasing, and advances at the rate of from 10 to 15 miles an hour, with a breaking wall of water from five to 12 feet high. Under such conditions of warfare between the ocean and the river, it is not surprising that the Amazon finds it impossible to build up a delta. Most of the 1,500,000 tons of sediment that the Amazon discharges daily into the sea is carried northward by coastal currents to be deposited along the Guiana coast. (Ed.)

Climate. The climate of Amazonia is warm, rainy, and humid. The length of day and night is equal on the Equator (which runs only slightly north of the Amazon River), and the usually clear nights favour relatively rapid radiation of the heat received from the Sun during the 12-hour day. There is a greater difference between the averages of the noonday and midnight temperatures than between the hottest and coolest months. Hence, night is the winter of the Amazon. To be sure, there are occasional cooler periods, especially during the Southern Hemisphere's winter months when an especially large and powerful air mass from the south polar region pours out over the Amazon Plain and causes a sharp drop in temperature, known locally as a *friagem,* or cold snap. At any time of the year several days of heavy rain will be succeeded by clear sunny days and fresh nights, when the temperature-humidity index is low. In the lower reaches of the river, cooling trade winds blow most of the year.

More important than temperature is the amount of precipitation. Moisture-laden winds blowing from the Atlantic

Ocean across South America are forced to rise when they reach the Andes, where they are cooled and, in cooling, lose their moisture through condensation to feed the great rivers that flow east from the Andes and form part of the Amazon River system. In the lowlands themselves, extensive convectional storms account for much heavy precipitation. There are at least three distinctive climatic types or subregions within Amazonia, based on the rainfall regime: (1) Near the mouth of the Amazon and the western portion of the plain there is an average year-round precipitation of more than 80 inches (2,000 millimetres), fairly well distributed throughout the year. During some years, twice this amount has been recorded, whereas in others there have been prolonged dry seasons. (2) In the rainfall regime that prevails over most of Amazonia there is a season of noticeable slackening of precipitation, although the decrease in rainfall is not severe enough to hinder plant growth. (3) Along the southern edge of the region, the climate of Amazonia grades into that of west central Brazil, with a more distinct dry season during the southern winter.

'hree
limatic
:gions of
.mazonia

The prevailing winds blow between east-northeastward and east-southeastward during the dry season. These winds are moderate in July and August but fresh during the rest of the season, when gusty winds, known locally as *marajós,* sometimes reach great force. This season is the best time for ascending the river. Sailboats descending the river drift with the current.

Normally, the influence of the tide is felt as far as Óbidos, situated about 600 miles above the river mouth. The *pororoca* occurs at times in the Amazon Estuary prior to spring tides; it consists of a wave from five to eight feet high, the crest of which breaks and spreads over the shallow waters of the river and its tributaries.

Soils. The appearance of the vast Amazonian forest vegetation is extremely exotic, leading to the erroneous conclusion that the underlying soil must be very fertile. The luxurious growth, however, stems only from the fragile ecological balance between soil and plants, influenced by light, temperature, and water. Generally, the soil free from flooding is well drained, porous, and of variable structure. Often it is sandy and of low to medium natural fertility because of its lack of phosphate, nitrogen, and potash and relatively high acidity. Small areas are underlain with basaltic and diabasic rocks, which are of high natural fertility. In Brazil these fertile soils are called *terra roxa* (reddish soil) and *terra preta dos Indios* ("black earth of the Indians"). The potentiality of the semiflooded areas is great; their soils do not lack nutrients, but use for agricultural purposes is limited by the periodic inundations.

Plant life. From an airplane, the great treescape of central Amazonia looks like a green carpet with holes or worn spots in the form of oxbow lakes, bright green sloughs, and extensive swamps. There are giants of the forest with white trunks that stand out like clumps of broccoli, flowering trees of many hues, and a great variety of palms. There is in the forests of Amazonia a bewildering diversity of species.

he
›rests of
mazonia

The forest stretches from the swampy mangroves and floating meadows (floating islands composed of intertwined land and water plants) in the east near the Atlantic, to the tree line on the Andes. High in the Andean region, there are large and small meadows (paramos; Spanish *páramos* or *paramillos*) of grass, sometimes with gigantic cacti. The flora in dry areas is distinctly different from that in swampy locations. On lower levels, swamps with arborescent calla-like Araceae (Montrichardia) are common. The Andean forest gradually turns into tropical rain forest, which can be divided into three main types: forest, the ground level of which is not inundated during the rainy season; areas that are regularly inundated; and areas that are occasionally inundated. One can frequently paddle a canoe between the trunks of tree species adapted to such nearly amphibious living. On the Brazilian central plateau, the forest proper is limited by a parklike vegetation of grasses and small to medium trees and bushes, characterized by twisted trunks, thick bark, and leathery leaves. Many of them have big subterranean woody trunks, and roots that reach deep into the groundwater. Where the trees leave sufficient space for a jeep to pass, the vegetation is called *cerrado.* Nearer to the streams, the trees grow taller and denser until they transcend into the forest proper. Pampas (open plain) sometimes suddenly occur.

he three
'pes of
›rest area

The floral province of Amazonia is characterized by the occurrence of rubber-bearing *Hevea* species and *Gnetum* but exhibits many local variations. Four general floral subprovinces may be discerned: (1) the upper Rio Branco, up to the Guianas, with predominantly grasslands and patches of forest; (2) the Jari-Trombetas river area, with varied vegetation of semideciduous, dry forest and *cerrados;* (3) the Río Negro area, which, despite its heavy rainfall, contains much desertlike sclerophilious vegetation (plants with thick leaves, resistant to water loss). This subprovince comprises the Río Negro, the upper part of the Orinoco, the lower and middle Japurá, and the western part of the Trombetas Basin; (4) the rain forest of the Amazon Basin, extending from the base of the Andes to the Atlantic Ocean and up to some parts of the northern boundary of the Amazon Valley.

The variety of trees is enormous; 117 species have been counted in an area of half a square mile. The most frequent varieties are myrtles, laurels, acacias, bignonias, cedrelas, cecropias, rosewoods, Bombacaceae, Brazil nuts, rubber trees, figs, and palms. The only gymnosperms are climbing Gnetums and rare coniferous trees or bushes of the evergreen podocarpus.

Four levels of forest vegetation are distinguishable. The floor level is often covered by a mosaic of small herbs and ferns. The second level is formed by bushes and taller herbs—heliconias, hirtellas, eugenias, calliandras, and others. The third level contains shade-loving trees, such as mauritia, orbygnia, and euterpe. These trees are sometimes covered by orchids, ferns, bromeliads, cacti, and pipers growing on them. It is quite common to find dense groups of palms in the middle of the forest.

The four
growth
levels of
the rain
forest

Higher trees constitute a fourth, or ceiling, level. Generally, these forest giants—such as the silk-cotton, or sumaumeira (*Ceiba*), the para nut, or castanheiro (*Bertholletia*), the sapucaia (*Lecythis*), and the sucupira (*Bowidchia*) with its blue flowers—stand alone and are conspicuous against the skyline.

Beyond the margin of dense forest, the first and second level almost disappear, and one can walk between the trunks of trees and hanging lianas without great difficulty. The diffused lighting prevents the growth of thick underbrush. As the Sun cannot be seen, it is easy to lose one's bearings. Few people have dared to penetrate the forest, and some have been lost on short trips. Even the Indians avoid entering the woods too far from their settlements near the riverbanks. They are afraid of legendary inhabitants of the dense forest and of becoming lost.

Animal life. To give an impression of the complete fauna of the Amazon is as impossible as it is to give one for the flora. More than 8,000 species of insects have been collected and classified. Myriads of mosquitos bother the traveller and may transmit malaria and yellow fever. The leaf-cutting ants (*Atta and Acromyrmex*) in a short time can destroy large parts of a plantation. The extermination, or at least the limitation, of these pests is of great importance to further habitation of the area. The most troublesome insects of all are the ubiquitous, small, black flies, called *piums,* whose bite can itch for days. Fireflies, stinging bees, hornets, wasps, beetles, cockroaches, cicades, centipedes, scorpions, ticks, red bugs, giant spiders, and butterflies are abundant. Sometimes hundreds and thousands of butterflies gather in the afternoon on humid sand near riverbanks or on islets, sipping moisture and often covering several square yards.

The number of fish species is estimated at about 2,000. The Amazon and its tributaries, together with the swamps and oxbow lakes, are in reality a vast sea of freshwater, most of it slowly flowing, in which live millions of fish. One of the best fish, reputed to be the biggest freshwater fish in the world, is the pirarucu, taken with a harpoon, not on rod and reel. It is a welcome source of protein to river dwellers. Thousands of tons of giant catfish, *gamitama, paitsche,* and many others are to be seen in the markets along the river. One of the most dangerous of river fish is

The vast
reserve of
fishes

the small, eight- to 10-inch piranha, ferocious and sharp-toothed, which will often attack any animal that enters the water, or even man; its razor-sharp teeth cut out chunks of flesh, and a school of these fish can do lethal damage in a short time. With the rapid means of transportation afforded by jet airplanes, there also are worldwide markets for tropical fish, such as silver carp, neon tetras, and even piranhas, as well as for aquatic plants.

Electric eels, giant water snakes such as the anaconda, and freshwater stingrays are native to many sectors of the vast river system. The flesh of the stingray is coarse-fibred but very palatable. Crocodiles are hunted for their skins, and turtles and their eggs are considered a delicacy.

Frogs and toads abound. Poisonous snakes are not numerous but are feared. Of the nonpoisonous species, the boa constrictor reaches huge proportions but seldom grows as large as the giant anaconda.

The capybara, largest rodent in the world, looks like a giant guinea pig and may weigh up to 100 pounds; it lives along the Amazon and its tributaries, feeding on the grasses, reeds, and rushes that grow on the banks, and taking refuge in the water when attacked; it is an excellent swimmer and diver; its flesh, though not considered a delicacy, is used for food and is frequently dried for shipment to market; its hide is made into very good leather.

The manatee, often called the sea cow, attains a length of six to eight feet, and a large specimen may weigh up to 2,000 pounds or more. It is hunted for its flesh and its oil, but it is so rare as to be almost extinct. Great efforts are being made at present to capture the animal alive, for shipment to zoos around the world.

The nutria, or tropical otter, is also to be found in less frequented areas and is valuable for its pelt. The tapir, which attains a weight of several hundred pounds, is an expert swimmer and diver and is much sought after for its excellent flesh. The peccary, or wild pig, and many kind of deer are natives of Amazonia.

Water buffalo, brought to the Ilha de Marajó (Island of Marajó) many years ago, have gone wild and now live in the more remote, swampy parts of that ranching area. The Brazilian government has recently introduced domesticated water buffalo and is studying them at an experimental station not too far from Santarém. They seem ideally adapted to such an area and may in time become the basis for a dairying industry, as they are in the brackish inlets and swamps north of Naples, Italy.

One of the most common animals is the small, agile squirrel monkey, in great demand for use in laboratories. There are howler monkeys, woolly monkeys, capuchin monkeys—those that so often accompanied the organ-grinder—and a host of others. The decapitated bodies of monkeys, split and blackened over a charcoal fire, are used extensively for food and are frequently seen for sale in local markets.

In more remote areas is found the sloth, a harmless animal that spends its life hanging from the limbs of trees, the leaves of which are its favourite food. The iguana lizard, often attaining a length of some five feet and looking like a prehistoric monster, climbs trees with the facility of a squirrel and is much sought after because of its delicious flesh.

The people. At the time of the Conquest, the bottom lands of the Amazon and its major tributaries supported relatively dense, sedentary populations who practiced intensive root-crop farming, abetted by fishing and by the hunting of aquatic mammals and reptiles. The *campos,* the higher areas between the rivers and their floodplains, were, and still are in some of the more remote sectors, inhabited by small, widely dispersed nomadic tribes of Indians. These groups rely largely on hunting of animals, large and small, and the gathering of wild fruits, berries, and nuts, and on some small patch agriculture of low yield.

The Amazonian Indians devised means of making the poisonous bitter manioca edible, the end product, *farinha,* being a food staple widely used today over most of tropical America. Amazonian Indians had perfected the use of quinine as a specific against malaria and extracted cocaine from the leaves of the coca tree. They also collected the sap of the Para rubber tree. The Amazonian Indians were skilled rivermen in their dugout canoes and had invented and used the hammock, an ideal bed in which to sleep in the humid tropics. Their ancient arrow poison, curare (*Chondodendron tomentosum*), has been used in the therapy of a host of paralyses and spastic disorders, such as multiple sclerosis, Parkinson's disease, and St. Vitus's Dance.

The early explorers of the Amazon, such as Orellana and Orsua and Aguirre, provisioned themselves from the food supplies of the Indians they met, commandeered their canoes, and took into slavery those that they wanted. The result was a complete breakdown in native life and an appalling decrease in the Indian population. As late as 1906 there were reports of the wholesale capture of wild Indians who were enslaved in order to tap rubber, which was plentiful, with a high price on the world market but which was difficult to get at because rubber trees were sparsely scattered over a huge area, and increasing toil was necessary to locate and work them. Women and children were wantonly butchered, as were any men who became ill or who could not gather their assigned quota. More recently, in the late 20th century, there were further reports of oppressive treatment of Indians by traders and land speculators.

The economy. *Resources.* The resources of the Amazon are many and varied. The first explorers were attracted by gold and diamonds, and industrial diamonds are still mined on the Rio Araguaia. Precious gems are occasionally found there and elsewhere. The most important game animals, used and traded for human nutrition, are: the manatee (or sea cow), which is menaced by extinction; various turtles and their eggs; the tapir; the red deer; the capybara; and, among the Indians, several types of monkeys. Wild forest birds and waterfowl are common game. Fish is also an important food source, especially the huge pirarucu (*Arapaima gigas*). Its flesh is dried in the sun and sold at every marketplace as a durable staple food.

Timber and other forest resources

Excellent timber is furnished by the mahogany (*Swietenia macrophylla*), the Amazonian cedar (*Cedrella odorata*), and many other species. Other trees, such as the cumaru (*Coumoruna odorata*), furnish perfumes and medical ingredients. The louro-inhamuí (*Ocotea barcellensis*) produces a liquid that can be used as a fuel for internal-combustion engines. The big sorva (*Couma macrocarpa*) exudes a drinkable milk sap. The economic kings of trees, however, are the rubber tree (*Hevea brasiliensis*) and the para nut (*Bertholletia excelsa*). The rubber tree has been one of the most important objectives of penetration and exploitation of the forest. It gave rise to a period of great but temporary prosperity, especially to the city of Manaus. Grown in plantations, it continues to make an important contribution to the economy.

Other nutritional sources include the heart of palm (*Euterpe oleracea*), the vanilla orchid, and the guarana (*Paullinia cupana*). Guarana, black pepper, and vanilla have become important agricultural crops. Guarana seeds are used to prepare a stimulating Brazilian soft drink, which may find world markets. Rice is cultivated in some areas. Agriculture and animal husbandry must be conducted carefully and scientifically, however, to preserve the very delicate ecologic balance that sustains the fertility of the soil.

Geological exploration is producing new knowledge of mineral resources. Oil has been found, and tin mining has been carried on for some time. Other valuable minerals are also extracted. (A.R.S.)

Trade and industry. The vast network of rivers is today organized under the control of a hierarchy of traders. Since travel away from the rivers is still extremely difficult, the people scattered in small communities along the riverbanks are dependent for their connections with the outside world on the launch of the trader to whom they are in debt. The largest trading companies are at Belém and Manaus; many smaller traders are controlled by the big companies, each of which occupies a strategic point at a river junction and controls all upstream communications. From many small villages comes a considerable movement of goods downstream, including such products as Brazil nuts, rotenone, cabinet woods, skins, and some

Dominance of trading companie

quantity of rubber. In return, the traders supply foodstuffs, tools, and firearms.

The relative emptiness of the Amazon has intrigued people for many years, and the increasing effectiveness of malaria control, improved diets, and sanitation measures, and the increased ease in transportation are all making the Amazonian region more attractive for human settlement.

The construction of new highways and the immediate development of agriculture along them, the construction of factories, paper mills, and an oil refinery, and the recent development of the tourist industry, all have injected a new dynamism essential to the economic development of the Amazon and its vast hinterland.

he
nportance
' new
ghways

Transportation. The opening of highways between Brasília and Belêm and between Brasília and Pôrto Velho has meant that manufactured goods from southern Brazil arrive with fewer delays and in many cases at freight rates much lower than formerly, when shippers had no competition. New settlements and agricultural developments have appeared along these roads and are themselves generating new passenger and commodity traffic. The road between Manaus and Pôrto Velho is having a profound impact on Amazonia and Manaus. Along the relatively short road (170 miles) from Manaus to Itacoatiara, Japanese colonists have organized agricultural cooperatives and have introduced modern agriculture techniques to produce vegetables, chickens, and eggs for the local markets, and black pepper for export to southern Brazil and the world market. The federal government is allocating millions of cruzeiros to provide lines of communication and supply, mainly by river, for the construction of the Trans-Amazonian and Cuiabá–Santarêm highways. Furthermore, plans are being made and moneys allocated to resettle thousands of families from drought-stricken northeastern Brazil to farms along these new highways.

The highway engineers seek to build the north–south roads as much as possible on the higher, better drained land between the rivers tributary to the Amazon. The Manaus–Pôrto Velho highway follows in general the interfluve area, frequently grass-covered *campos,* between the Purus and Madeira rivers. The Cuiabá–Santarêm highway will follow for the most part the higher interfluve area between the mighty Tapajós and Xingu rivers. The east–west road, however, cuts across the valleys of the north-flowing rivers. The first stretch of the TransAmazonian Highway, from Pôrto Franco in Maranhão to Itaituba in Pará, some 850 miles, was completed in early 1972. The second stretch of some 600 miles, from Itaituba to Humaitá in Amazonas, on the Rio Madeira, was finished almost a year later.

Study and exploration. In early days the river was the only means of access into the forest. Francisco de Orellana was the first to travel on the mainstream from the Peruvian Andes to the Atlantic (1541). Nearly a century later, Pedro Texeira went from Belêm to Quito, Ecuador and the region became more and more known through the explorations of the Portuguese. About 1751 Charles de La Condamine made the first astronomic-geographic survey and brought the deadly Indian arrow poison curare to Europe. At the beginning of the 19th century, the German explorer Alexander von Humboldt, accompanied by the French botanist Aimé Bonpland, mapped the connection between the Amazon and the Orinoco system through the Brazo Casiquiare (Casiquiare River), whose existence was insufficiently known before. The English naturalist Henry Walter Bates spent the years from 1848 to 1859 along the Amazon, collecting thousands of species of animals and writing up his notes of animals, human beings, and natural phenomena in a charmingly objective manner. His book, *The Naturalist on the River Amazons,* originally published in two volumes in 1863, was an immediate success and is one of the great classics on the Amazon River. An official U.S. expedition was sent to Amazonia in the middle of the 19th century; in 1854 in Washington William Lewis Herndon published as a public document the report that he and Lardner Gibbon, lieutenants in the U.S. Navy, had made to Congress under the title of *Exploration of the Valley of the Amazon.* They reported in fascinating detail the possibilities for power navigation in various parts of the river system, the resources available, health problems, social conditions, and possibilities for further settlement and development.

The 20th century has been a period of exploratory and scientific expeditions. In 1913–14, former U.S. president Theodore Roosevelt and Brazilian Col. Cândido Rondon headed an expedition that explored the Rio Roosevelt (originally Río Teodoro), a tributary of the Madeira, and made natural-history collections and observations. An expedition sponsored by Harvard's Institute of Geographical Exploration did important scientific work between the years 1910–24. The American Geographical Society and the American Museum of Natural History have undertaken geographic explorations and sent collecting expeditions into Amazonia over the years. The American Geographical Society compiled data for and published 1-to-1,000,000 scale maps of this vast area. These are only some of the highlights of exploration during the first three decades of the 20th century. Hardly a year passes that enthusiastic scientists and adventure-seeking hunters do not write books on this area.

During World War II the United States government sent dozens of American scientists and technicians into Amazonia to make regional surveys and to study the actual physical and human resources as well as to draw up plans for future development. Since the war, expeditions have been sponsored both by the American Geographical Society and by various United Nations organizations. Brazilian scientists have increasingly concerned themselves with the Amazon region. This is especially true of scientists associated with the Instituto Agronômico do Norte and the Museu Paraense Emílio Goeldi (Goeldi Natural History and Ethnographical Museum), both located in Belêm, the Conselho Nacional de Geografia, and the Instituto Nacional de Pesquisas da Amazônia, established in 1952, in Manaus. (R.E.Cr./A.R.S./Ed.)

ORINOCO RIVER

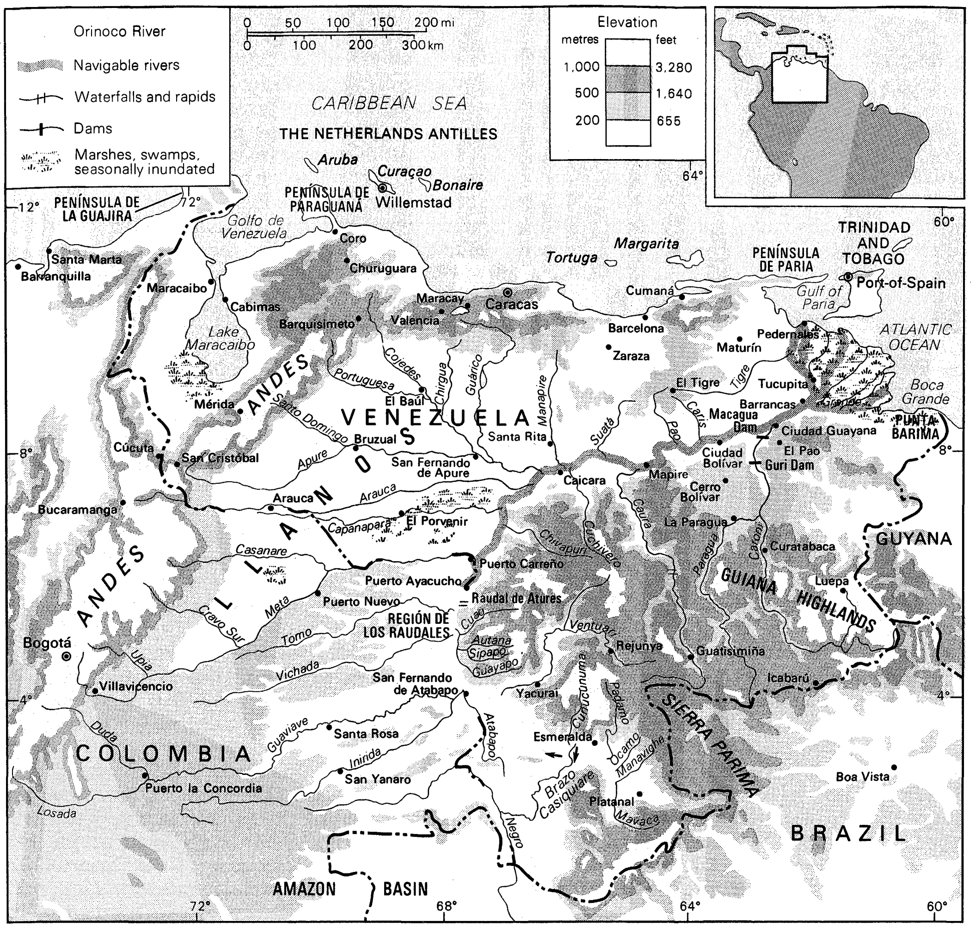
The Orinoco River (Río Orinoco) and its tributaries constitute the northernmost of South America's three major river systems. Bordered by the Andes mountains to the west and the north, the Guiana Highlands to the east, and the Amazon watershed to the south, the river basin covers an area of about 366,000 square miles (948,000 square kilometres). It encompasses approximately four-fifths of Venezuela and one-fourth of Colombia. The Orinoco River itself flows in a giant arc for 1,337 miles (2,151 kilometres) from its source in the Guiana Highlands to its mouth on the Atlantic Ocean. Throughout most of its course it flows through Venezuela, except for a section where it forms part of the frontier between Venezuela and Colombia. The name "Orinoco" is derived from Guarauno words meaning "a place to paddle"—*i.e.,* a navigable place.

For most of its length, the Orinoco flows through impenetrable rain forest or undeveloped grasslands; however, the cities of Ciudad Bolívar and Ciudad Guayana (also known as Santo Tomé de Guayana or Guayana City) are located on its lower course, and this region is fast developing into one of the most industrialized areas of South America. The river forms a waterway used in the exploitation of the vast mineral wealth of the Venezuelan interior.

Physical features. *Physiography.* The western slopes of the Sierra Parima (Parima Mountains), which form part of the boundary between Venezuela and Brazil, are drained by spring-fed streams that give rise to the Orinoco River. The source is placed in Venezuela at 63°21′ W and 2°19′ N, at an elevation of 3,523 feet (1,074 metres). From its headwaters the river flows west-northwest, leaving the mountains to meander through level plains known as the Llanos. The volume of its waters increases as the river receives numerous mountain tributaries, including the Río Mavaca on the left bank and the Manaviche, Ocamo, Padamo, and Cunucunuma rivers on the right.

The Brazo Casiquiare

Below the town of Esmeralda some of the waters of the Orinoco flow south into the Brazo Casiquiare (Casiquiare River, or Channel). This channel, a feature peculiar to the Orinoco River system, is a natural passage that flows for more than 220 miles to the Río Negro linking the Orinoco and Amazon river systems.



The Orinoco River Basin and its drainage network.

After its bifurcation in the Casiquiare, the river bends to the northwest and flows in great meandering curves to its confluence with the Río Ventuari. There the river turns to the west to run between high alluvial banks; its course is marked by numerous extensive sandbars. Near San Fernando de Atabapo, the Atabapo and Guaviare rivers join the Orinoco, marking the end of the Upper Orinoco.

Downstream from San Fernando de Atabapo, the river flows northward and forms part of the border between Venezuela and Colombia. It passes through a transitional zone, the Región de los Raudales (Region of the Rapids), where the Orinoco forces its way through a series of narrow passes among enormous granite boulders. The waters fall in a succession of rapids, ending with the Raudal de Atures (Atures Rapids). In this region, the main tributaries are the Vichada and Tomo rivers from the Colombian plains, and the Guayapo, Sipapo, Autana, and Cuao rivers from the Guiana Highlands.

The Raudal de Atures marks the beginning of the Lower Orinoco Basin, in which the river makes its great bend to the east. In this section, the river flows slowly through the lowest level of the plains and increases to about five miles in width. Along the bend, it receives the largest number of affluents of its entire course, including the Meta, Arauca, and Capanaparo rivers. The Río Apure contributes waters from numerous Andean streams, which form a swampy maze in their lower courses.

From its junction with the Río Apure, the Orinoco meanders eastward over gently sloping plains. Shoals and alluvial islands are abundant; some of the islands are large enough to divide the channel into narrow passages. Tributaries include the Guárico, Manapire, Suatá, Pao, and Caris rivers, which enter on the left bank, and the Río Cuchivero and the Río Caura, which join the main stream on the right. So much sediment is carried by these rivers that they often form islands at their mouths. The Río Caroní, one of the Orinoco's largest tributaries, joins the river on its right bank after passing through the 310-square-mile Guri Reservoir of Guri (Raúl Leoni) Dam, above Ciudad Guayana. Many lagoons, including the Mamo, Amana, and Colorada, are located on the banks of the Orinoco west of its confluence with the Río Caroní and east of Ciudad Bolívar. At the town of Barrancas, the river begins to form its great delta.

The delta region

The delta extends for about 275 miles along the Atlantic coast, from Pedernales on the Gulf of Paria in the north to Punta Barima in the south on the Boca Grande (literally the "Great Mouth"). Scores of islands are connected by innumerable *caños,* or canals, which constitute an intricate network. The main channel of the Orinoco, known in the delta as the Río Grande, flows eastward from Barrancas to discharge into the Boca Grande.

Climate. The climate is tropical, with the seasons marked by differences in rainfall rather than in temperature. The year is divided into two seasons, the rainy and the dry (locally known as winter and summer), the former extending from April to October or November, the latter most marked from November through March or April. Temperature differences, on the other hand, are slight throughout the year; and no month averages more than 69° F (21° C) or less than 64° F (18° C). Whatever the average temperature, there is little difference from month to month. The only marked variation is from day to night, being greater than that from month to month.

Rainfall varies considerably from district to district. The northeast trade winds blow across the coastal districts without losing much precipitation, in some places leaving less than 20 inches (508 millimetres) per year. Areas lying behind topographic barriers also get little rain, while windward slopes are generally well watered. In some areas enough rain falls to support a lush jungle growth and, in others, enough for a true selva (rain forest). The Llanos suffer severely from drought from about January to April, and then suffer equally from flooding of the whole countryside from June to October.

Hydrography. The river basin, as a geomorphological feature, dates from the Quaternary Period (from 2,500,000 years ago to the present). The enormous quantities of material produced by highlands are carried down by torrential rains to the rivers. The rivers, unable to hold the excessive material, overflow or break their banks, producing periodic floods that submerge the lowlands. Under these conditions, drainage presents an unstable and indefinite pattern, marked by the shifting of rivers, lagoons, and swamps over the lower lands. The Orinoco Delta is rapidly extending into the ocean, but the tremendous amounts of sediment that accumulate are accelerating the subsidence occurring in the entire delta region.

he
ainage
attern

Wide fluctuations in the river's flow reflect the seasonal rainfall regime. During the dry season or "low-water" period from October to March, the average river depth is about 49 feet in the lower basin near Ciudad Bolívar. The rise of the Orinoco begins with manifest regularity in April at the beginning of the rainy season. The "high water" period from April to October reaches its maximum in July. The depth of the river at this period is about 165 feet at Ciudad Bolívar. From June to August the lowlands of the basin are flooded and in some places are 65 feet under water. At the end of August the waters gradually recede and reach their lowest point in October.

Plant and animal life. The Llanos are grass-covered plains with isolated stands of palms and chaparro (scrub oak). The tributary streams that cross the plains have deposited alluvial soils, and strips of forests line their banks. Some of this natural tree cover, however, has been reduced by deforestation. The Guiana Highlands are covered with high, dense forest that is interrupted by small patches of savanna. The tropical rain forest of the Upper Orinoco Valley contains hundreds of species of trees. Rain forest also covers the delta region.

More than a thousand species of birds frequent the Orinoco region; among the more spectacular are the scarlet ibis, the bell bird, the umbrella bird, and numerous parrots. The numerous fish include the voracious caribe (piranha) and the *laulao,* a catfish that often weighs more than 200 pounds. The Orinoco crocodile is the longest of its kind in the world, reaching a length of more than 20 feet. The array, or side-necked turtle, which grows to a shell-length of about 30 inches, nests on the sandy islands of the river.

ıdigenous
ɔoples of
ıe basin

The people. Except for the Guajiros of Lake Maracaibo, all of the Venezuelan aboriginal population lives within the Orinoco Basin. The most important indigenous groups include the Guaicas (Waicas), also known as the Guaharibos, and the Maquiritares (Makiritares) of the southern uplands, the Warao of the delta region, and the Guahibos and the Yaruros of the western Llanos. These peoples live in intimate relationship with the rivers of the basin, using them as a source of food as well as for purposes of communication.

The important towns, with the exception of Ciudad Bolívar, are built on high ground to avoid recurrent flooding. Town plans reflect Spanish influence: streets are arranged in a chequered pattern with a central plaza. Most of the original colonial buildings have vanished, although some old homes still remain. In the small villages, thatched huts are the usual dwellings.

The economy. *Navigation and river crossings.* The Orinoco and its tributaries have long served as vast waterways for the indigenous inhabitants of the Venezuelan interior. Especially during the floods of the rainy season, with outboard motors are the only means of communication throughout large areas of the river basin. Large river steamers travel upriver for about 700 miles from the delta to the Raudal de Atures. Dredging has allowed large oceangoing vessels to navigate the Orinoco from its mouth to its confluence with the Río Caroní—a distance of 226 miles—in order to tap the iron-ore deposits of the Guiana Highlands. The Llanos and the Guiana region were connected in 1967 with the completion of the 5,507-foot bridge across the Orinoco at Ciudad Bolívar. In 1961, the Venezuelan government bridged the mouth of the Río Caroní to connect the new industrial town of Puerto Ordaz with the old Orinoco port of San Félix, thereby creating the urban unit of Ciudad Guayana.

Resources. The Guiana Highlands are rich in mineral deposits. Iron ore, which is more than 60 percent pure, is mined at Cerro Bolívar and El Pao.

Other minerals include deposits of manganese, nickel, vanadium (a metallic element used to form alloys), and chrome. There are also deposits of gold and diamonds. Petroleum and natural gas are exploited in the northern and northwestern parts of the basin.

The site of the Saltos del Caroní (Caroní Falls) had one of the largest hydroelectric potentials in South America, amounting to 10,000,000 kilowatts. A vast project to tap this power source includes the Macagua and Guri dams, development of which began in 1959 and 1968, respectively. The falls consequently were submerged.

The Llanos, along the great bend of the Orinoco, have long formed one of Venezuela's major cattle-raising areas. Seasonal subsistence farming is carried out along the river from the confluence with the Apure to Ciudad Bolívar. West of the Apure junction, cotton is grown on a commercial scale. Land reclamation and flood control projects in the delta region are planned in order to open vast agricultural lands.

The
Ciudad
Guayana
complex

Industrial development of the river basin is concentrated at Ciudad Guayana. The semiautonomous government body Corporación Venezolana de Guayana (Venezuelan Guiana Corporation) is conducting a careful development plan based on the resources of the nearby Guiana Highlands and the Río Caroní. The dams at Saltos del Caroní supply power for industry, including a steel mill, an aluminum plant, and a paper factory. Power is also supplied by natural gas, which is piped from the oil fields north of the Orinoco River.

Study and exploration. European exploration of the Orinoco River Basin began in the 16th century. A series of expeditions sponsored by the banking house of Welser of Augsburg penetrated the Llanos southward across the Apure and Meta rivers. From the east, several Spanish expeditions ascended the river from its mouth without much success. In 1531 the Spanish explorer Diego de Ordaz voyaged up the river, and that same year another Spanish explorer, Antonio de Berrio, descended the Casanare and Meta rivers and then descended the Orinoco to its mouth.

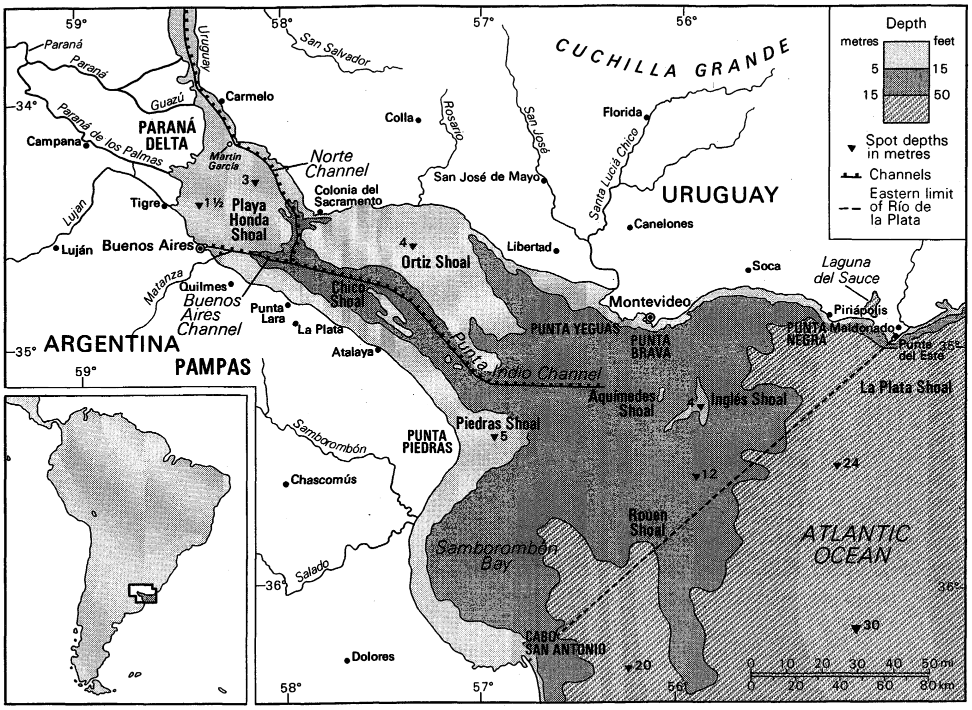
In 1744 Jesuit missionaries discovered the Brazo Casiquiare. Alexander von Humboldt, the German naturalist, travelled over 1,700 miles of the river basin in 1800. By 1860 steamships were navigating the Orinoco. The source of the river remained in dispute, however, until a Venezuelan expedition finally identified it in 1951.

(Me.F.G.)

RÍO DE LA PLATA SYSTEM

The Río de la Plata (literally River of Silver; in English often called the River Plate) is a tapering intrusion of the Atlantic Ocean occurring on the east coast of South America between Uruguay to the north and Argentina to the south. While some geographers regard it as a gulf, others even as a sea, and still others as a river, the majority regard it as the estuary of the Paraná and Uruguay rivers (as well as of the Paraguay River, which drains into the Paraná). As such, the Río de la Plata receives waters draining from the basin of these rivers, which covers northern Argentina, the greater part of Uruguay, Paraguay, eastern Bolivia, and much of southern Brazil and represents an area of 1,600,000 square miles (4,144,000 square kilometres). Montevideo, the capital of Uruguay, is located on the northern shore, and Buenos Aires, the capital of Argentina, is on the southwestern shore.

The head of the Río de la Plata is where the delta of the Paraná and the mouth of the Uruguay meet; the eastern, or Atlantic, extremity lies about 180 miles (290 kilometres) seaward and would be represented by a line on a map drawn from Cabo San Antonio, Argentina, to Punta del Este, Uruguay. The breadth of the estuary increases from the head seaward: it is 31 miles from Punta Lara on the southern (Argentine) shore to the port of Colonia del Sacramento on the northern (Uruguayan) shore, and 136 miles from shore to shore at the Atlantic extremity of the estuary. To those who regard the Río de la Plata as a river, it is the widest in the world,



Río de la Plata.

with a total area of 13,500 square miles (35,000 square kilometres).

Physical features. *Physiography.* The two contributory rivers bring down about 2,000,000,000 cubic feet of silt every year. The muddiness of the water in the Río de la Plata itself is increased by the tides and winds that prevent the settling of the deposit on the bed. The deposit thus forms great shoals, banks, or bars of clay, sand, and organic matter: the Playa Honda Shoal is just off the Paraná Delta; the Ortiz and Chico shoals are further downstream; and the Rouen, the Inglês, the Alemán, and the Arquímedes shoals are still farther out. The depth of the water—varying from six feet (two metres) above the shoals to 65 feet in the intervening channels—is reduced along the south coast by an offshore shoal.

Silt and shoals

The Argentine coast of the estuary is low-lying; its banks are of marine debris and coarse sand, and the coast is subject to flooding in places. The Uruguayan coast stands quite high and consists largely of rocks and dunes. Off the Uruguayan coast there are several small islands, such as Hornos, San Gabriel, López, Lobos, Farallón, and—opposite the mouths of the Uruguay and Paraná Guazú rivers—Martín García.

The volume of water discharged by the Río de la Plata into the Atlantic is estimated at 776,925 cubic feet per second. Although the water of the tributary rivers is so widely distributed over the length and breadth of the estuary that variations in their volume do not affect the level of the water, the level is considerably affected by variations of the tides and, especially, of the winds reaching it. The ocean tides, weak though they may be, flow 120 miles up the Paraná and the Uruguay rivers from their mouths on the estuary. The average tidal range is six inches at Montevideo and two and a half feet at Buenos Aires. The pampero (the southwest wind) and the southeast wind both exert a great influence on the Río de la Plata: the pampero, when it is most powerful, drives the water on to the Uruguayan coast, so that the water level drops on the Argentine side; the southeast wind has the contrary effect, flooding parts of the Argentine shore, and causing the level to drop on the Uruguayan coast.

The Paraná River (Spanish Río Paraná; Portuguese Rio Paraná), together with its tributaries, forms the larger of the two river systems that drain into the Río de la Plata Estuary on the east coast of South America. The Paraná is 2,485 miles (3,998 kilometres) long and extends from the confluence of the Grande and Paranaíba rivers in southern Brazil, running generally southwestward for most of its course, before turning southeastward to drain into the Río de la Plata. Paraná means "father of the waters" in the Guarani language.

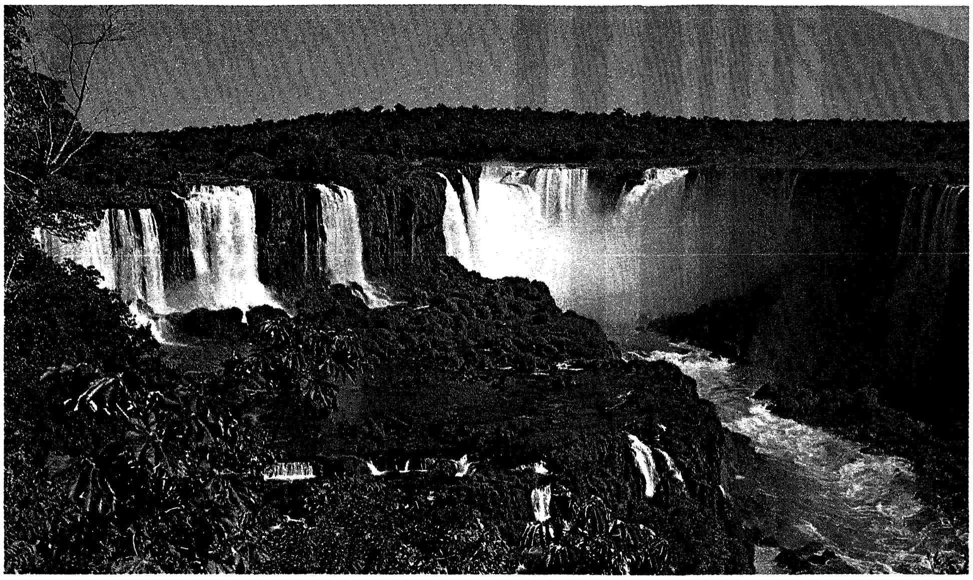
The Grande rises in the Serra da Mantiqueira, part of the mountainous hinterland of Rio de Janeiro, and flows westward for approximately 800 miles; but its numerous waterfalls (such as the Cachoeira do Marimbondo [Marimbondo Falls], with a height of 72 feet [22 metres]) makes it of little use for navigation. The Paranaíba, which also has numerous waterfalls, is formed by many affluents, the northernmost headstream being the São Bartolomeu, which rises in the Serra dos Pirineus near Brasília.

From its origin in the Grande–Paranaíba confluence to its junction, 1,740 miles downstream, with the Paraguay, the river is known as the Alto (Upper) Paraná. This upper course of the river receives many tributaries, some from the right, some from the left. The three most important of these tributaries, namely the Tietê, the Paranapanema, and the Iguaçu, all belong to the left bank, having their sources near the Atlantic coast of Brazil (the first two in São Paulo state, the last-named in Paraná state).

The Alto Paraná and its tributaries

The Alto Paraná flows at first in a southwesterly direction down a deep cleavage in the ancient Brazilian massif, the configuration of which determines its course. Just before it begins to run along the frontier between Brazil to the east and Paraguay to the west, the river has to cut through the Serra de Maracaju (or Cordillera de Mbaracayú), which has the effect of a dam; as the river approaches this natural barrier, it expands its bed into a lake two and a half miles wide and four and a half miles long. Pôrto Guaíra in Brazil stands on the southern shore of this lake. The river's passage through the mountains was, until 1982, marked by the Salto das Sete Quedas (Guaíra Falls), which had eight times the water volume of the Niagara River of North America. The completion of the Itaipu hydroelectric project's first stage in that year submerged the falls in the waters of the reservoir.

From the Iguaçu confluence to its junction with the Paraguay River, the Alto Paraná continues as the frontier between Paraguay and Argentina. So long as it is flanked on the left (Argentinian) bank by the steep edge of the Sierra de Misiones, the river proceeds in a generally southwesterly direction, but twists repeatedly to and fro over a rocky bed studded with outcrops of melaphyre (a dark-coloured rock of porphyritic texture). At Posadas in Argentina, however, where it is about one and a half miles



Iguaçu Falls, where the Río Iguaçu drops about 270 feet over a series of approximately 275 separate waterfalls.

Manley—Shostal/EB Inc.

wide, it turns abruptly westward and begins a more amply meandering course, embracing islands of quite considerable size and punctuated so frequently by rapids and by outcrops of basalt that navigation is difficult. At the Apipé Rapids the river is only about four to six feet deep.

The Río Iguaçu ("Great Water" in the Guaraní language) joins the Alto Paraná at the point where Brazil, Paraguay, and Argentina converge, 82 miles downstream from Pôrto Mendes. Rising in the Serra do Mar near the Brazilian city of Curitiba (for which reason it is sometimes called the Rio Grande de Curitiba), the Iguaçu has a course of 820 miles from east to west, during which some 70 waterfalls reduce its bed's altitude above sea level by a total of about 2,650 feet. While the Ñacunday Falls are 131 feet high, the spectacular Iguaçu Falls, on the frontier between Brazil and Argentina, 14 miles upstream from the Iguaçu–Alto Paraná confluence, have a height of about 270 feet. One of the currents into which the river is divided descends from this height at a single drop, but most of the others make two successive drops of 98 feet each. The cataracts have a mean water volume of 59,400 cubic feet per second. The crescent-shaped edge over which the water pours is more than one and a half miles wide; thereafter, the river passes for several miles through the Garganta del Diablo, or Devil's Gorge—a canyon only 164 feet wide between heights varying from 65 to 328 feet. At its junction with the Alto Paraná, however, the Iguaçu is 820 feet wide and 40 feet deep.

At Paso de Patria, on the right (Paraguayan) bank, at a place appropriately named Confluencia, the Paraná receives its greatest tributary, the Paraguay River, that joins it on the right bank.

ıe
iver of
ckades"

The fifth largest river in South America, the Paraguay River (Spanish Río Paraguay; Portuguese Rio Paraguai) is 1,584 miles (2,550 kilometres) long. The name Paraguay is taken from the Guaraní language and could be translated "river of *paraguas* (coloured, plumed birds)" or "river of cockades," an allusion to the plumed headdresses once worn by the riverine peoples.

The Paraguay rises in Brazil, in the central plateaus of Mato Grosso, at an altitude of 980 feet (300 metres) above sea level. Where it becomes navigable, about 150 miles downstream, near Cáceres in Brazil, after its confluence with the Sepotuba, it is 275 feet wide and 20 feet deep. Another 20 miles downstream, where the Jauru joins it at an altitude of 400 feet above sea level, the Paraguay enters the alluvial Mato Grosso *pantanal* (floodplain); it crosses the plain's western edge over a sandy bed, flowing around the many islands in its course. During its passage through the *pantanal,* it receives such important tributaries as the Cuiabá, the Taquari, and the Miranda.

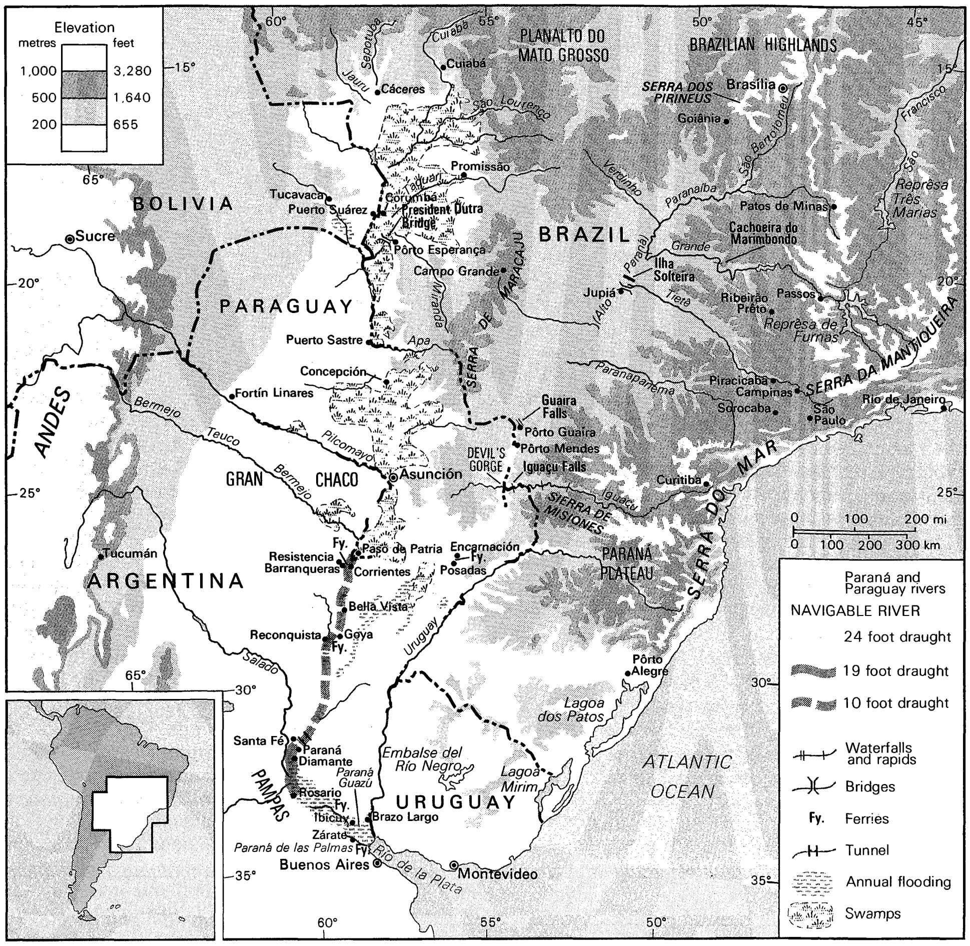
About 470 miles downstream, it flows north to south to form the boundary between Brazil and Paraguay before being joined by a tributary, the Río Apa, which flows in from the east and demarcates part of the Brazilian–Paraguayan frontier. The river then enters Paraguay, having travelled about 950 miles from its source. After flowing for more than 200 miles across Paraguay, it is joined by the Río Pilcomayo at the Argentinian border, near Asunción. It then flows south-southwest along the Argentinian–Paraguayan frontier for about 140 miles, until it is joined on its west bank by the Bermejo. Continuing along the border for another 40 miles, it then empties into the Paraná River at a short distance from the Argentinian city of Corrientes.

Throughout its hydrographic basin, which covers more than 380,000 square miles (985,000 square kilometres), altitudes rarely exceed 650 feet above sea level. Rainfall varies from 40 to 80 inches annually. Over a long distance, the gradient of the river varies from about three-quarters of an inch to an inch per mile. The various streams of the basin have low banks or natural levees, built up when silt is deposited along the slower flowing portions of the river channel during flood stage. When the river recedes, its banks thus remain elevated above the level of the neighbouring plains. During floods a continuous water table, often as much as 15 miles wide, underies the inundated plains, and about 38,600 square miles of surface area are flooded.

The Paraguay River has varying rates of flow between its source and its mouth. Above Corumbá, in Brazil, it has a typically tropical regime—at its highest in February and at its lowest from July to August. Below Corumbá, the high point occurs in July and the low point from December to January.

From its confluence with the Apa for the 630 miles to its mouth, the Paraguay, while still navigable, runs on a shallow, broad bed, with an average width of about 2,000 feet. Its right (Argentinian) bank gradually lowers, whereas its left (Paraguayan) bank becomes elevated, forming cliffs. Along this stretch, floods develop principally on the western bank, spreading over the Argentinian plain for distances of from three to six miles. These lands form part of the Gran Chaco. Downstream from the complex of low hills south of Assunção known as Lomas Valentinas, the floods spill out over both banks, inundating areas between six and nine miles wide. At this point, the riverbed has an average width of about 2,300 feet.

After its juncture with the Paraguay, the combined stream of the Paraná turns southward as it passes Corrientes. It now becomes a typical "plainstype" river, banked by its own alluvial deposits and having an extensive floodplain



The Paraná and Paraguay river basins.

on its righthand side for the next stretch of its course, with tracts up to 24 miles wide liable to inundation. Its permanent bed, about two and a half miles wide at Corrientes, narrows to about 8,000 feet at Bella Vista, to about 7,000 feet at Santa Fé, and to about 6,000 feet at Rosario and is strewn throughout with chains of islands. Santa Fé, on the right bank opposite the port of Paraná, stands where the Paraná River receives its last considerable tributary, the Salado. Between Santa Fé and Rosario, however, the right bank begins to rise as the river skirts the edge of the undulating plain, which flanks it down to the delta, and reaches altitudes ranging from about 30 to 65 feet. The left bank, meanwhile, is always higher than the right but has to sustain the erosive action of the water, which becomes more and more turbid as great masses of earth are constantly falling into it; in the delta the main branch of the river runs along a break in the terrain, with its left bank consisting of a cliff about 75 feet high.

The delta

The delta of the Paraná has its apex as far north as Diamante, upstream from Rosario, where branches of the river begin to turn southeastward. About 11 miles wide at its upper end, the delta is 40 miles wide at its lower, where the separated branches flow into the Río de la Plata, about 200 miles from Diamante. With an area of 5,500 square miles in 1970, the delta—from calculations based on the maps made in 1818, 1912, and 1945—appears to be advancing at the rate of 230 feet per annum, the yearly volume of alluvial deposit being estimated at 165,000,000 tons. Within the delta the river divides again and again into distributary branches, the most important being the two last great channels, the Paraná Guazú and the Paraná de las Palmas. The islands of the delta, alluvial in origin, are low-lying and of varying size. Their shores and the outer fringes of the river have protective embankments covered with trees, but may nevertheless be submerged in times of flooding, when they present the appearance of flooded forests.

The velocity of the current of the Paraná changes frequently during the river's long course. For the Alto Paraná, the rate becomes slower wherever the bed widens (especially when a real lake is formed, as at Itaipu Dam) and very much faster wherever the bed narrows (as in the canyon downstream from Itaipu). Farther downstream it slackens on its way to Posadas, but accelerates thereafter over a series of rapids and races. Downstream from Corrientes it becomes slower again, stabilizing its flow at a mean rate of two and a half miles per hour on the way to the Río de la Plata.

The volume of flow

The volume of the Paraná is, for practical purposes, correlated to the amount it receives from the Paraguay, which supplies about 25 percent of the total. High periods occur normally in the summer (November–February); low periods in August and September. An important factor is that the Alto Paraná and the Paraguay reach their maximum flow at different times. Whereas the mountainous basin of the Alto Paraná is drained so rapidly that water begins to rise at Corrientes in November, reaching its maximum height there in February, the swamps of the upper basin of the Paraguay retain precipitation so much longer that the Paraguay's high water does not reach Corrientes till May, reaching its maximum in June. Thus, levels on the lower Paraná begin to sink in March, then rise again from May, and sink again from July to September. Whenever both the Alto Paraná and the Paraguay reach their highest levels at the same time, the lower Paraná has to carry an exceptionally heavy volume of water—as it did in 1905, when the delta experienced heavy flooding.

The Uruguay River (Spanish Río Uruguay; Portuguese Rio Uruguai) is the second major system, 932 miles (1,500 kilometres) in length, that flows into the Río de la Plata. Like the Upper Paraná and the Paraguay, the Uruguay originates in Brazil, formed by several small streams that rise on the western slopes of the Serra do Mar, at 27°09′ S. From the south it is joined by the Pelotas, the two rivers serving to divide the states of Rio Grande do Sul and Santa Catarina. After flowing west, the Uruguay turns southwest at its juncture with the Peperi Guaçu, the first sizable affluent to join it from the north. For most of its course, the fast-flowing Peperi Guaçu marks the boundary between the Argentine province of Misiones and Brazil. Now the Uruguay serves to divide Brazil and Argentina. A few miles beyond the juncture with the Peperi Guaçu, the river is constricted between rocky walls in the Salto Grande de Misiones, a two-mile stretch of rapids with a total descent of 26 feet (eight metres) in eight miles. At the cataracts, the river narrows suddenly from 1,500 feet to an extreme of 70 feet.

Several small rivers join the Uruguay from the west and are navigable in their lower reaches by canoes and small boats. The principal of these, from north to south, are the Aguaypey, Mirinay, Mocoretá (which divides Entre Ríos and Corrientes), and Gualeguaychú. The important affluents of the Uruguay, however, come from the east. The Ijuí, Ibicuí, and the Quaraí (Guareim) are short rivers but of considerable volume; the last forms part of the boundary between Brazil and Uruguay. At the mouth of the Quaraí, the Uruguay becomes the boundary line between Argentina and Uruguay, and the river flows almost directly south. The Negro River, 434 miles long and the Uruguay's largest tributary, joins the latter only 60 miles from the estuary of the Río de la Plata. The Negro rises on the Brazilian border in Rio Grande do Sul and flows westward through central Uruguay. Sizable river craft can reach Mercedes, 32 miles from its mouth. Like the Upper Paraná, the Uruguay is generally clear and carries little silt except in the seasonal floods. After its juncture with the Negro, the Uruguay broadens sharply to a width of four to six miles and becomes a virtual extension of the Río de la Plata estuary.

Climate. The Río de la Plata has a temperate climate and copious rainfall at all seasons, amounting to 39 inches (990 millimetres) a year. The mean annual temperature is 55° F (13° C), and monthly averages are always over 50° F (10° C). Humidity, at 70 percent, is notably high. The prevailing winds are the north, the pampero, and the southeast. Frosts are frequent in the winter months.

The basin of the Upper Paraná has a hot and humid climate all the year round. The winters are dry, and the summers (lasting from November to April) are rainy, with precipitation averaging 60 inches a year but decreasing toward the west. Rainfall takes the form of drenching downpours often accompanied by hailstorms; the rainwater is drained away rapidly into the mainstream. The middle and lower basins have a subtropical climate of temperate humidity, with less plentiful rainfall (except in the Misiones Region) but without seasonal differences, the mean annual precipitation being 39 inches in the delta as opposed to 66 in Misiones. The climate along the Lower Paraná is very humid indeed and sometimes quite stifling in summer; the moist vapours become still thicker in the delta when the river brings down the torrential waters of the tropical basin.

The predominant climate of the Paraguay Basin is of the hot and humid savanna type, characterized by dry winters from April to September and heavy rains in summer from October to March. More than 80 percent of the annual precipitation occurs in the summer months, with little or no rainfall in June and July. Annual mean temperatures are above 64° F (18° C), the absolute maximum temperature being from 104° to 107° F (40° to 42° C) and the abslute minimum temperature being about 34° F (1° C). October is frequently the warmest month, except in the south, where the warmest month is January and where the contrast between rainy summers and dry winters is more pronounced. The extreme southern portion of the basin has a humid, moderately warm climate in which the heaviest rains occur during the summer and the winters are not completely dry.

Plant and animal life. The Brazilian section of the Alto Paraná forms the boundary between two natural zones: that of the forest to the east, that of the savanna to the west. Wood and maté tea leaves from the forest are shipped on the river, and the treeless savannah, with grasses and bushes, is good for cattle raising. On Argentina's section of the river, forest extends from the Misiones region to the vicinity of Corrientes; some forest trees, outside the zone of the forest proper, still occur in areas of woodland all the way downstream to the delta.

Species of fish

The Paraná has a rich and varied animal life. Among its many edible fish, the pejerrey (a marine fish, silver in colour, with two darker bands on each side), the dorado (a brilliantly coloured river fish with a golden appearance), the surubi (a fish with a long rounded body, flattened at the nose), the pati (a large scaleless fish, which frequents deep and muddy waters), and the pacu (a large river fish with a flat body, almost as high as it is long) are worth mentioning; every year sportsmen visit the stretch of the river upstream from Corrientes to angle for dorado. Reptiles include the iguana lizard, two species of caiman (a crocodilian), the water boa, the rattlesnake, the cross viper, and the yarará (the most prevalent South American representative of the viper family). Frogs and toads are plentiful, as also are freshwater crabs. There are innumerable species of insects and spiders, and the islands are plagued by mosquitos. Herons, cormorants, storks, and game birds are also seen.

The *pantanal* complex

In the Paraguay River system, some of the *pantanal*'s vegetation, called the "*pantanal* complex," is typical of the Mato Grosso plateau regions, while some other is typical of lowlands. Plants that thrive in water and in moist soils as well as those that flourish at moderate temperatures or are adapted to dry regions are found within the complex. As on the banks of the Niger in West Africa, both forest and grassland types occur. The water plants, found on the permanently flooded lands, are typified by the water hyacinth and by the "Victoria regia" water lily. Moisture-loving species, such as the trumpetwood and the guama, flourish over most of the floodplain. On the savanna, after the floods, various grasses such as paspalum and knotroot bristle grass reappear. Vegetation of a more evolved type, which thrives at moderate temperatures, occupies the unflooded highland. It is represented by nut-bearing palms and by various types of laurels. In the forests of the region, the carandá, or copernicia palm (a tropical palm that yields wax), the paratudo, the muriti palm (a large fan palm), and various types of quebracho trees (South American hardwoods that are a source of tannin) predominate. In the Gran Chaco region along the west bank of the river, and in other sections where drought is more pronounced, plants adapted to dry conditions occur. In the lowlands of eastern Paraguay, forest cover and savanna grasslands alternate.

Among the fish of the Paraguay River, some of which supplement the diet of the riverine population, are the dorado (a fish resembling salmon), the piranha (a fish resembling the bluegill that travels in large schools and devours any meat that falls in the water), and the pacu (which resembles bass).

The estuary of the Río de la Plata is very rich in fish; pejerrey (kingfish), corbina (white sea bass), patí (large scaleless river fish), surubí (the largest Argentinian river fish), dorado (a large golden fish resembling a salmon), and pacú (a large river fish almost as high as it is long) are among the edible species.

The people. *The Paraná.* The peoples who live up and down the Paraná, be they fishermen, farmers, herdsmen, boatmen, pilots, shipwrights, or raftmakers, all share certain customs and interests in common arising from their environment, from their racial identity, and from the cultural underdevelopment of the region. Originally, almost the whole length of the river region was populated by Guaraní Indians. The Spaniards, for whom the river was the only practicable way of penetrating the interior of the country, mingled freely with the Guaraní. The cultural, physical, and linguistic characteristics of the present population reflect this mingling.

On the Alto Paraná most of the people are poor fishermen living in huts on the fringe of the forest, and depending on the river both for their food and for their transport: apart from their fishing, they profit from services rendered to passing river traffic.

On the lower Paraná the economy is more diversified, but still primitive. Apart from fishing and providing services, the islanders raise livestock; some crops, such as maize, manioc, tobacco, and fruit, are grown. Settlement of the adjacent country was begun from the banks of the river and the economy remains dependent on the river in the absence of a system of roads.

The most distinctive features of human life and activity on the Paraná can best be observed in the delta, where water transport is the sole means of movement for people and for goods alike. Such local fruit growing as there is contributes little to ecological development, as the fruit trees are not suited to the climate and are frequently damaged by floods; it would seem more rational to proceed with the afforestation of the islands liable to flooding with pine and other softwood species.

The original inhabitants

The Paraguay. Before the arrival of the Spaniards in the 16th century, the region was inhabited primarily by the Paiaguá (a part of the Guaycure tribe) and by the Guaraní. A subgroup of the Guaraní, the Xaraiés, lent their name to the original designation of the *pantanal*—the Lagoon of the Xaraiés. The Paiaguá and the Guató of the *pantanal* engaged in foraging, hunting, and fishing. Others, such as the Tereno and the Guiana (Arawak), practiced a rudimentary type of agriculture, building hillocks above flood level. The southern Guaraní cultivated corn (maize) and cassava (manioc). The inhabitants of the Gran Chaco—the Paiaguá, Mataco, Mascoy, and Zamuco—were nomadic fishers and hunters.

In what is now Paraguay, the Spaniards and Portuguese interbred freely with the indigenous Amerindians. Consequently, the present riverine population of the country is largely mestizo, or mixed, and Guaraní as well as Spanish is the common language. In Brazil, however, miscegenation was less general, so that some groups of indigenous peoples remain, forming isolated nuclei. Others, like the Bororo, the Tereno, the Cadiueu, and the Bacairi, constitute minorities who have adopted Christianity and Brazilian culture and who live on the fringe of the region.

Among the first Europeans to enter the region were the Spanish Jesuits who mapped the river in the 17th century. In the 19th century, the first hydrographic survey of the river was made, followed in the 20th century by a more detailed hydrographic survey.

The Río de la Plata. The shores of the Río de la Plata are the most densely populated areas of Argentina and of Uruguay alike. The great ports, Buenos Aires and Montevideo, where the density is highest, are primarily concerned with exporting meat from the hinterland; the refrigerators and shipyards required for this trade are located in the coastal zone, as are flour mills, factories for vegetable oils, textile industries, metallurgical plants, and petroleum refineries.

For the people living along its shores, the Río de la Plata has always been useful as a waterway. For their essential livelihood they have looked inland to the pampas. Nevertheless, the estuary, as a thoroughfare for trade, is important not only to the people of the coasts but also to the remotest inhabitants of the whole drainage basin.

The economy. The economic usefulness of these river systems is not commensurate with the area that they drain. A principal problem is that of navigation. A large portion of the rivers cannot be used at all or only by very shallow draught vessels. Elsewhere navigation can only be maintained by constant dredging and renovation of port facilities. The other economic uses to which these rivers might lend themselves, such as irrigation or hydroelectric power, are equally difficult to achieve. The swamps of Xarayes and the Chaco long made agriculture a virtual impossibility in these areas. The gradual use of the potential electric power represented by sites such as Itaipu or Foz do Areia, however, has supported the development of the cities and industrial sites and of agriculture as well.

Buenos Aires is one of the principal seaports of the world and the main port of Argentina. Vessels approach it from the main estuary channel by one of two side channels that are clearly marked and dredged. Ocean vessels can travel up the Paraná River as far as Santa Fe or Paraná. Ocean trade can also reach Concepción del Uruguay directly by the Uruguay River.

Commerce farther upstream on these river systems operates under conditions that fluctuate considerably. Passage of vessels depends to a large extent on seasonal variations in depth. Long fleets of barges carry the bulk of the river freight. The current, narrowness, and curves above Corrientes on the Upper Paraná, however, rule out such barge transport. Several rapids on the Upper Paraná can only be passed with the use of winches to pull the vessels. Narrowness of the river, whirlpools, and the increased speed of the current make navigation more dangerous as the mouth of the Iguaçu is approached. A railroad, 38 miles long, to the town of Guairá circumvents the Itaipu site and opens up another 400 miles of navigable river farther up the Paraná. Beyond, stretches of the Paraná and of tributaries are navigable only by launches or canoes.

On the Paraguay River, vessels drawing seven feet are able to reach Corumbá in Brazil at all seasons; smaller ships can reach Cáceres, Brazil. Navigation of the Pilcomayo, Bermejo, and Salado is negligible, because of shifting channels, sandbars, and shallowness.

The value of these river systems as a commercial artery is, therefore, concentrated on the lower reaches. A large volume of ocean shipping reaches Rosario and Concepción del Uruguay, although the major seaports are Buenos Aires and Montevideo. The great bulk of river transport is concentrated within the limits of Asunción on the Paraguay, Corrientes on the Paraná, and Salto on the Uruguay. Local ship lines provide regular passenger and freight service to all navigable parts of these river systems. Rafts, canoes, and motorboats provide irregular although important service on the tributaries and upper reaches of these rivers.

Study and exploration. The Río de la Plata was discovered by Juan Díaz de Solís, chief navigator of Spain, in 1516, as a result of efforts to find a route to the Pacific. The estuary was temporarily named in his memory after his death on its shores at the hands of unfriendly Charrua Indians. Ferdinand Magellan touched at the estuary in 1520 during his circumnavigation of the globe. In 1526, Sebastian Cabot ascended the rivers as far as the present city of Asunción and obtained silver trinkets in barter with the Guaraní Indians. Spanish dreams gave the estuary its permanent name, Río de la Plata, in the hope that it might indeed become a river of silver. The major Spanish expedition that settled near the present location of Buenos Aires in 1536 under Pedro de Mendoza proved a fiasco. After much misfortune the survivors moved upstream to the surroundings of the more docile Guaraní Indians at Asunción. Buenos Aires was not refounded until 1580, and throughout the Spanish colonial era the Río de la Plata remained a backwash of the empire. The estuary was virtually closed to legal commerce until the end of the 18th century. Spain renewed its interest in the area only when Portuguese and English ambitions threatened to expand into the Río de la Plata in the 1760s.

Navigation of the river systems became a problem when the national states of Argentina, Uruguay, Brazil, and Bolivia emerged on its courses. Territorial conflicts and restrictions on navigation caused several wars, culminating in the titanic struggle by Francisco Solano López' Paraguay against Brazil, Uruguay, and Argentina from 1864 to 1870. In the 20th century, similar conflicts sharpened by rumoured oil wealth resulted in the Chaco War between Paraguay and Bolivia.

The development of agricultural wealth, particularly in Argentina, resulted in greater appreciation of the commercial value of these river systems after the mid-19th century. Wheat, beef, wool, and hides entered the river and world trade in increasing quantities from Argentina and Uruguay, while from Brazil and Paraguay came forest and tropical products and maté, or Paraguayan tea. Port construction and dredging made Buenos Aires more valuable as a seaport, and by 1902 similar improvements had been completed at Rosario. Marking of channels, soundings, dredging, and other aids to navigation became a responsibility of all the riparian states. (Ed.)

SÃO FRANCISCO RIVER

One of the greatest rivers in South America, the São Francisco, 1,800 miles (2,900 kilometres) long, is the third largest river system of Brazil and the largest river wholly within the country. It is regarded as "the river of national unity" because it has long served as a line of communication between Brazil's maritime and western regions and between the north and the south. The river is named for the 16th-century Jesuit leader, St. Francis Borgia (São Francisco de Borja y Aragon). It is an important source of hydroelectric power and irrigation for northeastern Brazil. The São Francisco Valley includes a total area of 243,682 square miles (631,133 square kilometres).

ıe upper
urse

Physical features. *Physiography.* The São Francisco rises at about 2,400 feet (730 metres) above sea level on the eastern slope of the Serra da Canastra (Canastra Mountains) in southwestern Minas Gerais state (*estado*), some distance south of the national capital of Brasília. It flows for more than 1,000 miles northward across the State of Bahia to the twin cities of Juazeiro and Petrolina. In this stretch the river receives its main left-bank tributaries—the Paracatu, Urucuia, Corrente, and Grande rivers—and its main right-bank tributaries—the Verde Grande, Paramirim, and Jacaré.

About 100 miles above Petrolina, the river begins a great curve to the northeast and enters a 300-mile-long stretch of rapids and falls. In this section the São Francisco forms the border between the states of Bahia to the south and Pernambuco to the north. The upper rapids are navigable during periods of high water, but below Petrolina the river is impassable. The broken course—during which the São Francisco receives the San Pedro, Ipueira, and Pajeú rivers—culminates in the great Paulo Afonso (Paulo Afonso Falls). At the top of the falls, the river divides suddenly and violently and cuts three successive falls through the granite rocks for a total drop of about 275 feet. Below the falls the river flows about 190 miles to its one mile-wide mouth on the Atlantic Ocean, about 60 miles north of Aracaju. In its lower section it is joined by the Rio Moxotó and forms the border between the states of Sergipe to the south and Alagoas to the north.

Climate. The climate of most of the river basin is hot and dry. The average maximum temperature for the region is 97° F (36° C) and the average minimum 66° F (19° C). The highest temperature recorded is 107° F (42° C). The prevailing winds are from the southeast, east, and northeast. Rainfall is deficient over most of the area and drought is frequent. Average annual precipitation measures 20 to 40 inches (508 to 1,016 millimetres) in the middle basin and 40 to 80 inches at the river mouth; about 8 percent of the basin receives only 10 inches of rain annually. Precipitation occurs during the summer months, from December to March, while the rest of the year—the winter season—is dry.

Plant and animal life. The upper river valley is an area of *caatinga,* or thorn forest. Important plants include the caroa, used for its fibres, and the mamona, which yields castor oil. There are trees, such as the oil, carnauba, and date palms; the *cajú* (cashew), which yields edible fruits; the palma, a spineless cactus with a high water content; and rubber trees. The upper middle valley, extending to the falls zone, is less dry and is covered with grassland (*campos cerrados*) and forests of semideciduous trees. Hardwoods include the jacaranda, Brazilian cedar (*cedro*), and vinhatico; and cochineal cactus, aloes, and vanilla plants also grow there.

The falls zone passes through the dry Brazilian interior, known as the *sertão.* The small amount of rainfall in the area permits the growth of only drought-resistant brush and grasses. The dry forests of the hilly uplands support carnauba and babassu palms and such plants as the cactus, the rock rose, and the rhododendron. Underground water in the region is too saline for irrigation or drinking.

The lower São Francisco flows through a floodplain of fine silt soils. Most of the original tropical semideciduous forest that grew along the river has been cleared for agriculture.

The river's fish are an important source of food. They include robalo, sardines, pocomó, and sarapó. The river mouth supports meru (squalas), manatees (sea cows), and sharks.

The people and economy. Except for the blacks who live along the coast, the people of the São Francisco Basin are of mixed Portuguese and Indian descent. The upper middle valley is an agricultural region in which cotton, rice, and corn (maize) are grown. The region also produces pineapples, barley, potatoes, rye, maté (tea), melons, sugarcane, beans, coffee, castor and cottonseed oils, and rum. Its major urban centre is Pirapora.

Pastoralism and *vazante* agriculture

The dry *sertão* is used largely for the grazing of cattle, goats, sheep, and donkeys. Along the river banks *vazante* agriculture is practiced: during the rainy season, shallow waterbeds (*vassantes*) are enclosed by bars of river sediment, and support the cultivation of cassava (manioc), water beans, and melons. Truck crops are grown on the river banks, and carnauba, caroá, and rubber are collected. The major town is the market centre of Salvador. Most of the lower river valley is dry and suitable to pastoral activities. Along the coast, wetland rice is grown.

Because the river flows through the driest region of Brazil, it is subject to seasonal changes in water level of up to 30 feet (10 metres). All of its tributaries run dry during the winter; above Juàzeiro, the riverbed varies from a narrow channel during drought periods to several miles in width during the rainy season.

The river is navigable for small steamers for more than 1,000 miles from Pirapora, Minas Gerais, to Várzea Redonda, Bahia, upstream of Petrolina. Sandbars at the river's mouth prevent the entry of deep-draft ocean vessels. There is a railway bridge across the river at Pirapora, and another bridge connects the cities of Petrolina and Juázeiro. Although river transport is slow and difficult, the São Francisco is an important link between the mining district of Minas Gerais and roads that radiate north and east from Juàzeiro.

The river's hydroelectric potential

The river's hydroelectric potential is its most important resource. The main centre of power for northeastern Brazil is the Paulo Afonso Cachoeira, site of the Trés Marias Dam. There are also hydroelectric plants providing electricity to the coastal cities of Salvador, Aracaju, and Recife at dams on the upper and lower stretches of the river.

The river basin is also important as a source of irrigation water. The largest storage dam is at Quixadá, Ceará state. There are large reservoirs at Cachoeira and Pedra d'água, and there are plans for a reservoir on the Jaguaribe and a 300-mile-long irrigation canal to join the São Francisco at Petrolina to branches in eastern Pernambuco, Paraiba, and eastern Ceará states.

The river basin contains deposits of agate, gold, iron, diamonds, opals, antimony, galena (the principal ore of lead), mercury, copper, arsenic, manganese, cobalt, and pyrites. There are also deposits of salt, sulfur, alum, marble, limestone, and clay. (C.E.Ca.)

BIBLIOGRAPHY

General works: PIERRE BIROT, *Les Régions naturelles du globe* (1970), contains two chapters on the Andes and on the old shields that constitute a modern approach to the structure of South America; JEAN DORST, *South America and Central America* (1967), a natural history of the continent, region by region; E.J. FITTKAU *et al.* (eds.), *Biogeography and Ecology in South America,* 2 vol. (1968–69), a series of valuable contributions on various aspects of natural history and relations between man and environment in South America; ARTHUR S. MORRIS, *South America* (1979), a general regional geography of the continent, showing interrelationships between landscape and economy.

Geological history: HEINRICH GERTH, *Der Geologische Bau der Südamerikanische Kordillere* (1955), a controversial book from which our modern views on the geology of the Andes originated; and *Geologie Südamerikas,* 2 vol. (1932–35), a comprehensive book on the geology of the continent, now somewhat outdated; WILLIAM F. JENKS (ed.), *Handbook of South American Geology* (1956), a country-by-country survey of the geological history of the continent and an analysis of the main formations; HANS STILLE, *Einführung in den Bau Amerikas* (1940), a classic book on the geology and structure of the Western Hemisphere; GEORGE W. STOSE (ed.), *Geologic Map of South America* (1950); E.F. SUSZCZYNSKI, "La Géologie et la tectonique de la Plateforme amazonienne," *Geol. Rdsch.,* 59:1232–53 (1970), a modern view on the structure of the oldest parts of South America; L.G. WEEKS, "Paleogeography of South America," *Bull. Am. Assoc. Petrol. Geol.,* 31:1194–1241 (1947).

Physical geography: "The Soil Resources of Latin America," *World Soil Resources Rept. 23* (1965), a survey of soils and

land use; HENRY W. BATES, *A Naturalist on the River Amazon,* 2 vol. (1863), a classic book on the discovery of Amazonia by one of its first modern explorers; ISAIAH BOWMAN, *The Andes of Southern Peru* (1916, reprinted 1968), a comprehensive view of geological structure and natural history; D.L. BRAMAO and LEMOS PETEZVAL, "Soil Map of South America," *Proc. Int. Congr. Soil. Sci.* (1960); GILBERT J. BUTLAND, *Latin America: A Regional Geography* (1960), a valuable contribution; FRED A. CARLSON, *Geography of Latin America,* 3rd ed. (1952), describes the structure of the continent and the natural regions; PRESTON E. JAMES, *Latin America,* 4th ed. (1969), a broad survey; CARL TROLL (ed.), *Geo-ecology of the Mountainous Regions of the Tropical Americas* (1968), a series of contributions to ecological problems in relation to relief and climate in the Andes.

Vegetation and animal life: MARSTON BATES, *The Land and Wildlife of South America* (1964), a book of interest to the general reader by a leading naturalist; ANGEL CABRERA and JOSÉ YEPES, *Historia natural Ediar. Namíferos sudamericanos* (1940), the only comprehensive book on South American neotropical mammals and their life; ANGEL L. CABRERA and ABRAHAM WILLINK, *Biogeografía de América Latina* (1973), a detailed analysis by biological zones of the animal and vegetable life of the continent; CARL H. EIGENMANN and WILLIAM R. ALLEN, *Fishes of Western South America,* 2 vol. (1942), a key to neotropical fishes and their characteristics; THOMAS H. GOODSPEED, *Plant Hunters in the Andes,* 2nd ed. rev. (1961), a work that contains firsthand data on high mountain vegetation; STEPHEN HADEN-GUEST, JOHN K. WRIGHT, and EILEEN M. TELLAFF, *A World Geography of Forest Resources* (1956), an inventory of world forestry containing a chapter on South America; ROBERT C. MURPHY, *Oceanic Birds of South America,* 2 vol. (1936), oceanography of the South American seas and life histories of all of the marine species; PAUL W. RICHARDS, *The Tropical Rain Forest* (1952), a classic book on the rain forest and its characteristics throughout the world; RODOLPHE MEYER DE SCHAUENSEE, *A Guide to the Birds of South America* (1970), a key to the bird life of the continent, with the description and distribution of every species; G.G. SIMPSON, "History of the Fauna of Latin America," *Am. Scient.,* 38:361–389 (1950), a geological history of neotropical mammals and their distribution in South America; FRANZ VERDOORN (ed.), *Plants and Plant Science in Latin America* (1945), a work that contains contributions on vegetation, agriculture, and land use.

Natural resources: AMÍLCAR O. HERRERA, *Los recursos minerales de América latina,* rev. ed. (1965), a survey of the resources of the continent; CARL A. LAMEY, *Metallic and Industrial Mineral Deposits* (1966), a world survey of mineral resources, with much data on South America; ULRICH PETERSEN, "Metallogenic Provinces in South America," *Geol. Rdsch.,* 59:834–897 (1970), information on the geological origin and localizations of the main sites where ores have been found; the COMMISSION FOR INLAND FISHERIES OF LATIN AMERICA, *The Inland Waters of Latin America* (1979), a detailed scientific inventory.

Human resources: An excellent general survey of ethnic origins, geographic distribution of the various groups, and racial relations, may be found in MAGNUS MORNER, *Race Mixture in the History of Latin America* (1967); and in MARVIN HARRIS, *Patterns of Race in the Americas* (1964, reprinted 1974); and a similar summary and historical analysis in NICOLÁS SÁNCHEZ-ALBORNOZ, *La Población de América latina* (1973; Eng. trans., *The Population of Latin America: a History,* 1974). Two major contributions on the South American Indians are JULIAN H. STEWARD and LOUIS C. FARON, *Native Peoples of South America* (1959); and CLARK WISSLER, *The American Indian,* 3rd ed. (1950). Steward's work is a general summary of the monumental *Handbook of South American Indians,* 7 vol. (1946–59). A typology of the ethnic origins and their distribution in the various South American subcultures is provided by CHARLES WAGLEY and MARVIN HARRIS, "A Typology of Latin American Subcultures," *Am. Anthrop.,* 57:428–451 (1955). On indigenous languages, in addition to chapters in the previously mentioned books, one of the classics is CESTMIR LOUKOTKA, *Classification of South American Indian Languages* (1968). RAMIRO CARDONA G. (ed.), *América Latina: Distribución espacial de la población* (1975), provides detailed papers on aspects of location of human populations in the region. ROBERT M. LEVINE, *Race and Ethnic Relations in Latin America and the Caribbean: An Historical Dictionary and Bibliography* (1980), includes approximately 1,200 entries in the dictionary and more than 1,340 books and articles in the bibliography. The most complete statistics and interpretations on mass European immigration before 1930 is found in WALTER F. WILLCOX (ed.), *International Migrations,* 2 vol. (1929–31, reprinted 1969). A more recent contribution is WILFRID D. BORRIE *et al., The Cultural Integration of Immigrants* (UNESCO, 1959). In addition, the International Labour Office's publication, *Indigenous Peoples: Living and Working Conditions of Aboriginal Populations in Independent Countries* (1953), gives data and documents on the situation of the Indians in South America.

Political geography: J.P. COLE, *Latin America: An Economic and Social Geography* (1965), is a comprehensive treatise. For a perceptive interpretation of the Spanish in South America, see SALVADOR DE MADARIAGA, *El auge del Imperio Español en América* (1945; Eng. trans., *The Rise of The Spanish American Empire,* 1965, reprinted 1977). An important anthology, including sources and interpretation, covering the period from conquest to modern times, is LEWIS HANKE (ed.), *History of Latin American Civilization* (1967). An excellent history covering the same period is HUBERT HERRING, *A History of Latin America from the Beginnings to the Present,* 3rd ed. rev. (1968). A general textbook, covering various aspects of culture and society in the region, from colonial times to our days, is HELEN M. BAILEY and ABRAHAM P. NASATIR, *Latin America: The Development of Its Civilization,* 3rd ed. (1973). An integrated overview of the social, economic, and political modernization of the region is given in GINO GERMANI, "Stages of Modernization in Latin America," in *Studies in Comparative International Development* (1969–70). On the varieties of Spanish, see PEDRO HENRIQUEZ URENA, *Sobre el Problema del Andalucismo Dialectal de América* (1932). See also J.H. PARRY, *The Discovery of South America* (1979), an account relying heavily on contemporary chronicles; and JORGE I. DOMINGUEZ, *Insurrection or Loyalty: The Breakdown of the Spanish American Empire* (1980).

Resource development and commerce: J.P. COLE (*op. cit.*), is a comprehensive study of the population, physical background, major historical events, institutions and main aspects of economic development, together with more detailed analyses for individual countries. See also the UNITED NATIONS, ECONOMIC COMMISSION FOR LATIN AMERICA (UN-ECLA), *Economic Survey of Latin America* (annual), a report covering major developments in the economic field and economic trends for each country—some issues also deal with specific aspects such as energy, transportation, or small industry; and the *Economic Bulletin for Latin America* (published twice yearly), which includes special articles as well as informative and methodological notes and updates the statistical data in the annual surveys. ALCIRA LEISERSON, *Notes on the Process of Industrialization in Argentina, Chile, and Peru* (1966), focuses primarily on the main political characteristics of these countries as they undergo industrialization, and analyzes the relationship between populism and the rise of the new technocratic elites. The report of the INTER-AMERICAN DEVELOPMENT BANK, *The Process for Industrialization in Latin America* (1969), refers particularly to regional integration, technological innovation, external markets for regional manufactures, external financing, and the role of government and private enterprise in industrial development. BÉLA KÁDÁR, *Problems of Economic Growth in Latin America* (1980), is an overview of the development experience with case studies drawn from five countries. The *Report of the Latin American Symposium on Industrial Development* (UN, 1966), contains an evaluation of the process of industrialization, its past evolution, present characteristics, and future prospects. The PAN AMERICAN UNION, *Industrialization in Latin America: Priority Problems* (1967), analyzes the relationship between industrialization, import substitution, and employment. LEHMAN B. FLETCHER and WILLIAM C. MERRILL, *Latin American Agricultural Development and Policies* (1968), contains papers dealing with policies for economic growth in the agricultural sectors of eight countries. JOHN A. HOPKINS, *The Latin American Farmer* (USDA, 1969), is a thorough study of farm conditions in Latin America. ROBERT T. BROWN, *Transport and the Economic Integration of South America* (1966), is a very complete study of the various modes of transportation and their relation to economic integration; see also CHARLES J. STOKES, *Transportation and Economic Development in Latin America* (1968); PAN AMERICAN UNION, *General Problems of Transportation in Latin America* (1962); and the UN-ECLA, *Transport in Latin America,* 3 vol. (1965). DONALD W. BAERRESEN, MARTIN CARNOY, and JOSEPH GRUNWALD, *Latin America Trade Patterns* (1965), is a general historical survey of trade between the Latin American countries and a detailed analysis of more recent developments in trade among LAFTA members. Also of interest are UN-ECLA, *Regional Integration and the Trade of Latin America* (1968), a study of various possible forms of economic integration and of obstacles to achieving their objective; and its *Statistical Bulletin for Latin America,* a monthly publication giving statistical information on selected economic indicators; PAN AMERICAN UNION, *Latin America's Foreign Trade: Problems and Policies* (1966); the UNITED NATIONS, FOOD AND AGRICULTURAL ORGANIZATION, *Production Yearbook* and *Trade Yearbook* (both annual); and the INTERNATIONAL MONETARY FUND and INTERNATIONAL BANK FOR RECONSTRUCTION AND DEVELOPMENT, *Direction of Trade* (monthly).

Demography: An important series of studies of demographic trends may be found in the special issue of the *Milbank Memorial Fund Quarterly,* vol. 43, no. 4, pt. 2 (1965), on *Components of Population Change in Latin America.* A short

general presentation by a major Latin American demographer is the article by CARMEN A. MIRO, "The Population of Latin America," in CLAUDIO VELIZ (ed.), *Latin America and the Caribbean* (1968). Another general article is KINGSLEY DAVIS, "The Place of Latin America in World Demographic History," *Milbank* (*op. cit.*), 42:19–47 (1964). Many demographic studies appear in the *Economic Bulletin for Latin America,* a journal published in English (twice yearly) by the United Nations Economic Commission for Latin America, Santiago de Chile. On fertility, see the special issue of the *Milbank* (*op. cit.*), vol. 46, no. 3 (1968), on *Current Research on Fertility and Family Planning in Latin America;* and O. ANDREW COLLVER, *Birth Rates in Latin America* (1965). On urbanization, internal migration, and living conditions in the cities, see PHILIP M. HAUSER (ed.), *Urbanization in Latin America* (1961); and GLENN H. BEYER (ed.), *The Urban Explosion in Latin America* (1967). Current demographic statistics appear in the *Statistical Abstract of Latin America* (annual). DOUGLAS BUTTERWORTH and JOHN K. CHANCE, *Latin American Urbanization* (1981), a review of rural–urban migration of the lower classes.

South Asian Arts

South Asia, consisting of the huge subcontinent of India, includes Pakistan, Bangladesh, Sri Lanka (formerly Ceylon), as well as the nation of India itself. In spite of differences in physical appearance, complexion, stature, and other ethnological features, the people of the entire region of South Asia are unified by a common cultural and ethical outlook; a wealth of ancient textual literature in Sanskrit, Prākrit, and regional languages is a major unifying factor. Music and dance, ritual customs, modes of worship, and literary ideals are similar throughout the subcontinent, even though the region has been divided into kaleidoscopic political patterns through the centuries.

The close interrelationship of the various peoples of South Asia may be traced in their epics, as in the *Rāmāyaṇa* and the *Mahābhārata.* Kinship between the gods and heroes of regions far distant from each other is evident, and the place-names themselves often evoke common sources. Moreover, there have been continual attempts to impose a political unity over the region. In the 3rd century BC, for example, the emperor Aśoka had almost all of this region under his sway; in the 11th century AD, Rājendra I Cōḻa conquered almost the whole of India and a good portion of Southeast Asia; and the great Mughal Akbar again achieved this in the 16th century. Though the expansion and attenuation of boundary lines, the bringing together or pulling apart politically of whole regions, have characterized all of South Asian history, the culture has remained essentially one.

The geography of the region encouraged a common adoration of mountains and rivers. The great Himalayas, which form the northern boundary, are the loftiest of mountains and are conceived to be the embodiment of nobility, the abode of immaculate snow, and the symbol of a cultural ideal. Similarly, the great rivers such as the Brahmaputra and the Indus are regarded as the mothers of their respective regions, assuring prosperity through their perennial supply of water.

The association of lakes and springs with water sprites and sylvan fairies, called *nāga*s and *yakṣa*s, is common throughout the region. Karkoṭa, the name of an early dynasty, itself signifies *nāga* worship in Kashmir. Sculptures of *nāga*s and *yakṣa*s found in widespread sites suggest a common spirit of adoration, as do sculptures, paintings, temples, and religious texts that for centuries were preserved within an oral tradition without losing their immaculate intonation. The same classical dance is seen in sculpture in Gandhāra in Pakistan, in Bhārhut in the north, and in Amarāvatī in the south.

The relation of the various arts to each other is very close in South Asia, where proficiency in several arts is necessary for specialization in any one. Thus, it is believed that without a good knowledge of dance there can be no proficiency in sculpture, for dance, like painting or sculpture, is a depiction of all the world. For its rhythmic movements and exposition of emotion, dance also requires musical accompaniments; hence, knowledge of musical rhythm is essential. For the stirring of emotion either in music or in dance, knowledge of literature and rhetoric is believed to be necessary; the flavour (*rasa*) to be expressed in music, dance, sculpture, or painting requires a literary background. Thus all the arts are closely linked together.

The arts were cultivated in South Asia not only as a noble pastime but also in a spirit of dedication, as an offering to the Almighty. Passages in literature refer to princes studying works of art for possible defects. One inscription that mentions the name of the *sūtra-dhāra* ("architect") of the 8th-century Mallikārjuna temple at Pattadakal epitomizes the accomplishments and ideals, in both theory and practice, of the artist.

Artists traditionally have enjoyed a high position in South Asian societies. Poets, musicians, and dancers held honoured seats in the royal court. An inscription mentions the appreciation bestowed by Rājendra Cōḻa on a talented dancer, and the architect of the temple at Tiruvoṟṟiyūr, who was also patronized by Rājendra, was eulogized for his encyclopaedic knowledge of architecture and art. Nonetheless, the folk arts were closely linked with the elite arts. Tribal group dances, for example, shared common elements with classical art, dance, and music. Among the artistic traditions of the Indian subcontinent, sculpture in the round (*citra*) is considered the highest artistic expression of form, and sculpture in relief (*ardhacitra*) is next in importance. Painting (*citrābhāsa,* literally "the semblance of sculpture") ranks third. Feeling for volume was so great that the effect of chiaroscuro (*i.e.,* use of light and shade to indicate modelling) was considered very important in painting; a passage from a drama of the 5th-century poet Kālidāsa describes how the eye tumbles over the heights and depths suggested in the modelling of a painting. A classical text on art, *Citrasūtra* enumerates noteworthy factors in paintings: the line sketch, firmly and gracefully drawn, is considered the highest element by the masters; shading and depiction of modelling are valued by others; the decorative element appeals to feminine taste; and the splendour of colour appeals to common taste. The use of a minimum of drawing to produce the maximum effect in suggesting form is considered most admirable.

Portraits play an important role in the visual arts of South Asia, and there are many literary references to the effective depiction of portraits both in painting and in sculpture. A 6th-century text, the *Viṣṇudharmottara,* classifies portraiture into natural, lyrical, sophisticated, and mixed, and men and women are classified into types by varieties of hair—long and fine, curling to right, wavy, straight and flowing, curled and abundant; similarly, eyes may be bow-shaped, of the hue of the blue lotus, fishlike, lotus-petal-like, or globular. Artistic stances are enumerated, and principles of foreshortening are explained. Paintings or sculptures were believed to take after their creators, even as a poem reflects the poet.

Although South Asia has continually been subjected to strong outside influences, it has always incorporated them into native forms, resulting not in imitation but in a new synthesis. This may be seen even in the art of the Gandhāra region of Pakistan, which in the 4th century BC was immersed in Greco-Roman tradition. In the sculpture of this period Indian themes and modes have softened the Western style. Foreign influence is evident after the invasion of the Kushāns in the 1st century AD, but the native element predominated and overwhelmed the foreign influence. During the Mughal period, from the 16th century, when Muslims from Central Asia reigned in South Asia, the blend of Iranian and Indian elements produced a predominantly Indian school that spread throughout the region, making it a unified cultural area under imperial rule. The influence of Islāmic art was enhanced by the second Mughal emperor, Humāyūn, who imported painters from the court of the Shāh of Persia and began a tradition that blended Indian and Persian elements to produce an efflorescence of painting and architecture.

Art in all these regions has reflected a system of government, a set of moral and ethical attitudes, and social patterns. The king's desire to serve the people and to take care of them almost as his offspring is evident as early as the 3rd century BC. The ideal of the king as the unrivalled bowman, as the unifier, as the tall and stately noble spirit, as the sacrificer for the welfare of the subjects, and as the hero of his people (who conceive of him as on a stately elephant) is comprehensively illustrated in a magnificent series of coins from the Gupta Empire of North India of the 4th–6th centuries. The concepts of righteous conquest and righteous warfare are illustrated in sculpture. The long series of sculptures illustrating the history of the South Indian Pallava dynasty of the 4th–9th centuries gives an excellent picture of the various activities of government—war and conquests, symbolic horse sacrifices, the king's council, diplomatic receptions, peace negotiations, the building of temples, appreciation of the fine arts (including dance and music), the coronation of kings, and so on, all clearly demonstrating what an orderly government meant to the people. Similarly, moral and ethical attitudes are illustrated in sculptures that lay stress on *dharma*—customs or laws governing duty. The doctrine of *ahiṃsā,* or noninjury to others, often is conceived in the symbol of a deer, and the ideal of a holy place has been conceived as a place where the deer roams freely. The joy in giving and of renunciation is clearly indicated in art. Sculptures illustrate simple and effective stories, as from the *Pañcatantra,* one of the oldest books of fables in the world. The spirit of devotion, faith, and respect for moral and ethical standards pervading the social structure in South Asia through the centuries is continuously represented in South Asian painting and sculpture. (C.S.)

The article is divided into the following major sections:

Literature

The peoples of South Asia have had a continuous literature from the first appearance in the Punjab of a branch of the Indo-European-speaking peoples who also settled all of Europe and Iran. In India this branch of Indo-Aryans, as they are usually called, met earlier inhabitants with different languages and no doubt a different culture—possibly a culture akin to that of the Indus Valley civilization, which had a script, and perhaps a literature of its own, of which nothing is known. Certain to have been settled in India were peoples who spoke languages of Dravidian origin, as well as other languages, called Munda, now preserved only by aboriginal tribes, which show affinities with the languages of Southeast Asia.

The earliest literature is of a sacred character and dates from *c.* 1400 BC in the form of the Rigveda. This work stands at the beginning of the literature of the Veda, or canonical Hindu sacred writings, which as a whole is roughly contemporary with the settlement of the Indo-Aryan peoples in the Punjab and farther east, in the mesopotamia of the Ganges and Yamunā rivers. The language of the Rigveda, which is a compilation of hymns to the high gods of the Aryan religion, is complex and archaic. It was simplified and codified in the course of the centuries from 1000 to 500 BC, which saw the development of prose commentaries called the *Brāhmaṇas, Āraṇyakas,* and *Upaniṣads.* While there must have been a long tradition of grammarians, the final codification of the language is ascribed to Pāṇini (5th or 6th century BC), whose grammar has remained normative for the correct language ever since. This language is called Sanskrit (Tongue Perfected). Sanskrit has had a scarcely interrupted literature from *c.* 600 BC until today, but its greatest efflorescence was in the classical period, from the 1st to 7th centuries AD. Because it was identified with the Brahminical religion of the Vedas, reform movements such as Buddhism and Jainism disdained the use of Sanskrit and adopted literary languages—amalgams of different dialects of the parent language—of their own, Pāli in Buddhism and Ardhamāgadhī in Jainism. These languages, usually called Prākrits—that is, derivative as well as more "natural" languages—produced a vast and, again, mostly sacred literature. In a further development of these dialects, the early beginnings can be seen of modern Indo-Aryan languages of northern India: Bengali (also the language of Bangladesh), Hindi (the official language of the Republic of India since 1947), Rajasthani, Punjabi, Gujarati, Marathi, Kashmiri, Oriya, Assamese, and Sindhi, each of which produced a literature of its own. Their names are derived from the regions in which they are spoken, regions with uncertain boundaries, where the different dialects fused at the borders. They all retained a close family resemblance that made bilingualism easy and a fact of Indian literary life.

Far more marked was the difference between Indo-Aryan speech and the languages of the Dravidian family, which are structurally wholly different, though in time a measure of convergence took place. Among them, the oldest recorded is Tamil, now the language of Tamil Nadu (Madras) state and of northern Ceylon (now called Sri Lanka), whose literature goes back to the early centuries of the Christian Era. Later to be put to literary uses were the cognate Telugu (Andhra Pradesh), Kannada (or Kanarese, Mysore state), and Malayalam (Kerala state) languages.

Influence
of Sanskri

In spite of this linguistic differentiation, the literatures composed in all of these languages reflect, in different degrees, the monumental influence of Sanskrit literature, Sanskrit being the universal Indian language of culture. This influence was one of both substance and form: in substance it provided the basic themes of literary enterprise, notably through the epics, the *Mahābhārata* and *Rāmāyaṇa,* the Hindu popular texts of the *Purāṇas,* especially the *Bhāgavata,* and the mythological repertory that came with Sanskritic Hinduism; in form, Sanskrit belles lettres bequeathed models of literary composition, and Sanskrit poetics provided the aesthetic theory underlying the models. The impact of Islām created a new language, Urdu (from Persian: Camp), based on Hindi; Urdu was the lingua franca of the army. Urdu was used later for literature and at present is the mother tongue of most Indian Muslims and their brethren in Pakistan. Its influence, however, does not compare with that of Sanskrit.

Comparable to the impact of Sanskrit, but far more alien, is that of English, which began to assert itself in the 18th century. The language brought with it new literary forms that were gradually adapted to the old ones, producing new genres—without necessarily giving up the older ones—in the local languages and giving rise to an interesting literature in the English language. Once more, a universal cultural language to a large extent unified aims in the scattered languages; English still plays this role, though it appears to be slowly declining.

SANSKRIT, PĀLI, AND PRĀKRIT LITERATURES: 1400 BC–AD 1200

Sanskrit: formative period (1400–400 BC). The oldest document in the literature of South Asia is the Rigveda, or Veda of the Stanzas (*c.* 1400 BC), the fundamental text of Brahminical Hinduism. Not literary but religious-magical in its purposes, it is mostly a compilation of hymns, dedicated to a number of gods of the Vedic religion. They have the regular structure of an invocation: the attention of the god is evoked; a brief account of some of his feats is given, to hold his attention; and an exhortation for his help concludes the hymn. The poets, of whom little is known, appear to have come at the close of a priestly poetical tradition, rivalling one another in allusions to obscure exploits, in language often opaque and at times intended to mystify. Nevertheless, the Rigvedic hymns include lines of great beauty. They may occur in a riddling verse, such as "When the ancient Dawns first dawned, the great Syllable was born in the footsteps of the Cow," alluding to the birth of speech at the beginning of creation. Or they may occur in poetry addressed to a deity whose beauty inspires the poet to well-turned lines. To the Dawns, for example: "They approach equally in the east, spreading themselves equally from the same place./ The Goddesses waking from the seat of order, like herds of kine set loose, the Dawns are active"; or to the goddess of the night: "Night coming on, the goddess shines/ In many places with her eyes:/ All-glorious she has decked herself./ Immortal goddess, far and wide/ She fills the valleys and the heights:/ Darkness with light she overcomes."

Nonsacred verses are very rare in the Rigveda, but, when they occur, they can be quite powerful, as in a hymn of a gambler, who is speaking:

It pains the gambler when he sees a woman,
Another's wife, and their well-ordered household:
He yokes these brown steeds early in the morning,
And, when the fire is low, sinks down an outcast.

"Play not with dice, but cultivate thy cornfield;
Rejoice in thy goods, deeming them abundant:
There are thy cows, there is thy wife, O gambler."
This counsel Savitri the kindly gives me.

fluence
the
gveda on
nskrit
etry

Although not literary in purpose, the Rigveda had a decisive influence on the form of Sanskrit poetry: except for narrative verse, the basic unit of all subsequent poems (no matter how many verses they consist of) is the single stanza that contains one complete thought.

The second Veda (*c.* 1200 BC), the Yajurveda (Veda of the Yajus [Formulas]), contains sacred formulas recited by a group of priests at the great Vedic sacrifices; and the third (*c.* 1100 BC), the Sāmaveda (Veda of the Chants), is in essence an anthology of the Rigveda. More literary interest attaches to the fourth Veda (1200 BC), the Atharvaveda (an *atharvan* was a special priest), which contains hymns, incantations, and many magic charms.

The succeeding literature (*c.* 1000–700 BC), the *Brāhmaṇas* ("Disquisitions About the Ritual"), continues not the poetry but the liturgical concerns of the Rigveda. They were written in a dry, expository prose, so that only their narrative portions have any literary interest. Much the same is true of the next layer of Vedic texts (800–600 BC), the *Āraṇyakas* ("Books Studied in the Forest").

ie
panisads

But the picture changes in the *Upaniṣads* (*c.* 1000–500 BC; "Collections of Esoteric Equations"). These prose texts at times convey the actual mode of teaching of a revered sage, in a style that can be strikingly intimate:

"Bring me a fruit of that *nyagrodha* (banyan) tree."
"Here it is, venerable Sir."
"Break it."
"It is broken, venerable Sir."
"What do you see there?"
"These seeds, exceedingly small, venerable Sir."
"Break one of these, my son."
"It is broken, venerable Sir."
"What do you see there?"
"Nothing at all, venerable Sir."
The father said: "That subtle essence, my dear, which you do not perceive there—from that very essence this great *nyagrodha* arises. Believe me, my dear."

While the older *Upaniṣads* are in prose, the later ones, dating from around 500 BC, mark a shift back to verse. They are the oldest examples of didactic verse, a genre that later gained enormous popularity.

The contribution of late-Vedic texts to later literature is preeminently that of the development of an expository prose style and the evolution of a sacred language, which, in order to be effective, must be completely correct. Thus, the Vedic religion evolved a science of phonetics and, later, of grammar, which was summed up in the 5th or 6th century BC by the grammarian Pāṇini in *Aṣṭādhyāyī* ("Eight Chapters"), a book that was to become basic to Sanskrit education. This language, Sanskrit, remained the language *par excellence* for later literature and was used for literary purposes until the 13th century and, epigonically, until today.

Sanskrit: epic and didactic literature (400 BC–AD 1000). After the formative period of the Vedic age, literature moved in several different directions. The close of the Vedic period was one of great cultural renewal, with the founding of the new monastic religions of Buddhism and Jainism (6th century BC) and the more slowly emerging rearticulation of Brahminism into Hinduism. Neither the earliest Buddhists nor the Jains availed themselves of Sanskrit in their preachings, apparently viewing the language as the preserve of a Brahmin elite. Sanskrit continued in derivative works of Vedic inspiration and above all in the *Mahābhārata* and the *Rāmāyaṇa.*

Mahābhārata. From references in Vedic literature it appears that side by side with the ritual texts there flourished a more secular literature carried on by bards. Originally charioteers to noblemen and thus witnesses of their feats, they chronicled the martial history of the families to which they were attached. From these beginnings, part chronicle, part panegyric, developed the epic style.

Like most Sanskrit poetry, the *Mahābhārata* consists of couplets, two successive lines with the same metre. Generally, one metre is used throughout the poem, though for stylistic effects other metres may be interspersed. The epic metre, or *śloka,* is a very fluid one that lends itself excellently to improvisation. The *Mahābhārata* is the longest poem in history, with about 100,000 couplets, more than seven times the size of Homer's *Iliad* and *Odyssey* combined. Its characters go back to around 1000 BC, but in its present form the epic could not have been composed before 400 BC. From that time until AD 400, it underwent continuous elaboration, by insertions of episodes (one of which is related in the religious poem called the *Bhagavadgītā*), accounts of separate adventures of the heroes, tales generated by their ancestors, and so on; and in the end it became a storehouse of general Hindu lore, with lengthy didactic books inserted.

The main narrative of the *Mahābhārata* recounts the growing up of two sets of cousins, both of whom aspire to a throne, the title to which is clouded. The protagonists, the Pāṇḍavas, stake their possessions in a dice game with the antagonists, the Kauravas, who are in effective control of the realm; they lose, and must live for 13 years in exile. This the five brothers do, along with the wife they hold in common. Upon their return from exile, they are refused their promised share of the kingdom, and, though parleys are held, war is inevitable. All of the Indian dynasties and tribes take sides in a war that lasts for 18 days, which only seven warriors, among them the Pāṇḍavas, survive. Noteworthy is the picture of gloom and doom that the *Mahābhārata* draws: there is little extolling of the heroic virtues of prowess and gallantry; rather, the wastefulness and bloodshed of war are pointed up, prefiguring a later concern with *ahiṃsā,* or nonviolence.

This summary does no justice to an extremely complex story with hundreds of participants, but it sketches the general outline of epic events. The main story has an unmistakable epic and heroic tone, and some of the events and encounters are completely comparable to those in epics of other peoples. But narrative and stylistic unity are disrupted by the inserted quasi-related and unrelated secondary episodes, each of which has a style of its own, ranging from light badinage to sonorous morality tales. It was in these episodes that the *Mahābhārata* lived on and greatly influenced succeeding literature; the story of Śakuntalā, for example, which the great 5th-century classical poet Kālidāsa embroidered, the slaying of Śiśupāla, the battle of the hero Arjuna with the mountain man, the story of Nala, and so on. But the most celebrated episode surely is the *Bhagavadgītā*.

The *Bhagavadgītā*

The influence of the *Bhagavadgītā* ("Song of the Lord") has mainly been on the development of Hindu religion and philosophy. Still, it is open to doubt whether it would have exerted this influence were it not for its poetry. Like most of the *Mahābhārata,* the style is simple and direct, not given to embellishment; nevertheless, the poem often reaches the height of expressiveness, as in its evocation of the theophany of Krishna as Vishnu, in the 11th of its 18 chapters. It led to imitations such as the *Īśvaragītā,* ("Song of the Lord [Śiva]"), also in the *Mahābhārata,* in which the god Śiva (Shiva) is celebrated.

Rāmāyaṇa. While the unity of the *Mahābhārata* has been disrupted by interpolations, the unity of the second epic, the *Rāmāyaṇa,* has been remarkably preserved. It is less an epic than a romance, recounting the story of prince Rāma and his wife Sītā. The first book, a later addition, tells of the youth of the prince, who later, by the trickery of one of his father's wives, is excluded from the throne to which he is heir. He goes into voluntary exile in the forests with his wife and his brother Lakṣmaṇa. There a demon, Rāvaṇa, abducts Sītā to his island kingdom of Laṅkā. In the course of his quest for her, Rāma allies himself with a monkey nation, whose general, Hanumān, later revered as a god, discovers Sītā on Laṅkā. A monumental battle ensues. "As the sky can only be likened to the ocean and the ocean to the sky, so the battle of Rāma and Rāvaṇa can only be likened to the battle of Rāvaṇa and Rāma." After his victory, Rāma is restored to the throne, but (in what appears to be a later addition) the populace accuses Sītā of misbehaviour, probable adultery, while in Laṅkā. Rāma thus abandons her to a hermitage (the sage of the hermitage, Vālmīki, is credited with the authorship of the *Rāmāyaṇa*), where she gives birth to their twin sons. Ultimately, Rāma takes Sītā and his sons back. In the later additions, the first and probably the last books, King Rāma is accepted as an incarnation of the god Vishnu, rather than merely a perfect man and hero.

It is the main story of the romance that has made an indelible impression on Indian culture, morally as well as literarily. Rāma is the perfect, just king; Sītā, the model of an Indian wife; Lakṣmaṇa (the brother), the paragon of fraternal love; and the monkey Hanumān, the epitome of a servitor's loyalty. It was translated into and adapted in many modern Indian languages, and (like parts of the *Mahābhārata*) it found its way into Java. Vālmīki himself was hailed by later classical poets as the first true poet (*kavi*), and indeed much of his work has a poetic freshness and literary intention that is largely absent from the *Mahābhārata.* Vālmīki's great tools are metaphor and simile, as is also true of later literature. He delights in description of pastoral scenes, in lamentations and grand martial spectacles, and in the idyll of the hermitage, which depicts a serene sage leading a life of quiet meditation and living on simple forest fare in a tranquil woodland close to a sacred river. And the entire work is suffused with a confident, unwavering morality, for which the heroes of the *Mahābhārata* are still searching.

Harivaṃśa and Purāṇas. The role of the *Mahābhārata* as the storehouse of Hindu lore was supplemented by the *Harivaṃśa* ("Genealogy of Hari"—that is, the god Vishnu), which deals with the ancestry and exploits of Krishna, the Pāṇḍavas' friend and adviser in the epic but now wholly deified and identified with the great god Vishnu. Then, from perhaps the 4th century, the literature of the *Purāṇas* took over. Encyclopaedic works, often of considerable length, the *Purāṇas* deal with the mythology of time and space and of deities, with sagas of great heroic dynasties, and with legends of saints and ascetics; their interest is largely religious. Aesthetically, the most important of them is the *Bhāgavata-Purāṇa* (9th or 10th century), which celebrates the blessed lord (*bhagavat*) Vishnu in his many theophanies but is particularly evocative in its celebration of Vishnu's incarnation as Krishna and the playful story of his youth. The influence of the *Bhāgavata-Purāṇa,* particularly the 10th book, on Indian religion, art, and literature has been monumental. In the opinion of one scholar, this book constitutes the greatest poem ever written; and so it is in the popular estimation of the Hindus. It was adapted in many Indian languages and provided themes and scenes for the flourishing miniature styles of the Middle Ages.

Pāli and Prākrit literature (c. 200 BC–AD 200). No more than the Vedic literature do the literatures of early Buddhism and Jainism have a literary intention. Their texts, written in dialects other than Sanskrit, articulate the teachings of the religious founders and their successors. Because they were transmitted orally for a considerable time before they were written down in the form they would retain, they underwent the inevitable censorship of the centuries, both negative in the form of documents dropped out of use and positive in the form of newer documents added. The dates given here are only approximations of the time of the documentary fixation of the dates.

Buddhist texts. The earliest records of Buddhism are not textual but inscriptional, in the famous edicts of the Mauryan emperor Aśoka, who reigned *c.* 269–232 BC. Among these inscriptions on stone, the so-called 13th rock edict—in which Aśoka, after the massacre of the Kaliṅgas (modern Orissa), abjures war—is the most moving document of any dynastic history. The inscriptions were written in a variety of Prākrits; that is, Indo-Aryan languages closely cognate to, but considerably later than, the earliest stabilized Sanskrit.

Pāli, the language of textual literature

The vehicle of the extant textual literature is the Pāli language, which is held to be a western Indian dialect on a substratum of several central and eastern ones. It was the language in use by the Theravāda school of Buddhism; but, since that school became the dominant one among many in early Buddhism, the Pāli language is often identified with the Buddha's own speech. Most of the canonical literature is exclusively of religious interest, but interspersed in it are works of considerable literary interest.

Foremost perhaps are discourses put into the Buddha's mouth—for example, his sermon "In the Deer Park"—and no doubt deriving from fairly accurate memories. With their straightforward, lively, and incisive style, homely similes, and simple humour, they are excellent examples of the homiletics of early Buddhist preaching. Incorporated in the canon, too, are more general works of literature. The *Dhammapada* ("Verses on the Buddhist Doctrine") is a fine example of the moralistic, aphoristic strain in Indian literature, in which virtue is extolled and vice condemned. It has remained a work of considerable diffusion in all Buddhist countries, and, as in the case of the *Bhagavadgītā* in Hinduism, much of its popularity is due to its literary style. The *Suttanipāta* collection of the Buddhist canon, composed in a more formal style, contains 55 narrative and didactic poems, in the form of dialogues and ballads; they are composed in a metre akin to the Sanskrit *śloka.* Of great interest are the *Theragāthā* and the *Therīgāthā* ("Hymns of the Senior Monks" and "Hymns of the Senior Nuns"), which give at times a vivid insight into the ambience in which a conversion to Buddhism took place: a monk celebrates his newfound freedom in an idyll of the hermit's life; and a nun reminisces over the pains of deserting her home and child, yet without regrets, since she has won the freedom of Buddhism. The prosodic variety of Buddhist lyrics is great; about 30 different metres can be distinguished. Pāli poems, with their new metres (often based on a musical phrase), stylistic features, figures of speech, and choice diction, foreshadow classical *kāvya* literature in Sanskrit, whose extant specimens date from a later period.

The *Jātakas*

Of great importance is a huge volume called *Jātakas* ("Birth Stories"), recounting some 500 episodes supposedly having occurred in the Buddha's earlier lives. Only those parts in archaic verse are canonical; the prose portion was written later (*c.* 3rd century AD), probably in

Ceylon. The *Jātakas* consist of fairy tales, animal stories and fables (the future Buddha may be incarnate in an animal), ballads, and anecdotes. Though their setting is often imaginary, they provide significant material for the historian of society and culture. These mostly short tales abound in moving, delicate, often rustic touches that have made them the delight of the Buddhist world. Their themes are illustrated in bas-reliefs of Buddhist shrines (or *stūpas*) at Bhārhut and Sānchi and monumentally on the great *stūpa* of Java, the Borobuḍur.

Of considerable literary as well as historical interest is the Pāli text *Milinda-pañha* ("The Questions of Milinda"). Milinda is identical to the Greek Menander, the name of a Bactrian Indo-Greek king (*c.* 140–110 BC) who was skeptical of the verities of Buddhism and was enlightened by the teaching of an elder, Nāgasena. The extensive Buddhist erudition that the sage displays is artfully presented in the form of simile and parable, and the work has contributed importantly to the edification of audiences in the countries where Buddhism came to be established. The style, in spite of the repetitions so typical of Buddhist doctrinal texts, is lively and presents the reader with an invaluable picture of contemporary Indian life.

Jaina texts. Less interest attaches to Jaina canonical works, which were written in an adapted and stabilized literary dialect called Ardhamāgadhī (Semi-Māgadhī, Māgadhī being the dialect of the ancient kingdom of Magadha, in present day Bihār). The belletristic contribution of Jaina literature is discussed below.

harac-
ristics of
e *kāvya*
yle

Classical Sanskrit kāvya (200–1200). Prepared for by the systematization of the Sanskrit language by Pāṇini, the development of the great epics, notably the *Rāmāyaṇa,* and the refinements of prosody represented by the Pāli lyrics, there arose, in the first centuries AD, a Sanskrit literary style that governed canons of taste for a millennium and remained influential far later through modern Indian languages and their literatures. The style, called *kāvya,* is characterized by an extremely self-conscious effort on the part of the writer to compose poetry pleasing to both the ear and the mind. It evolved an elaborate poetics of figures of speech, among which the metaphor and simile, in their many manifestations, predominate; a careful use of language, governed by the stated norms of grammar; an ever-increasing tendency to use compound nouns instead of drawing on the quite plentiful possibilities of Sanskrit inflection; a sometimes ostentatious display of erudition in the arts and sciences; an adroitness in the use of varied and complicated, if appropriate, metres—all applied to traditional themes such as the epic had provided and to the rendering of emotions, most often the love between men and women.

The style finds its classical expression in the so-called *mahākāvya* ("great poem"), most akin to the epyllion ("miniature epic") art form of the Alexandrian poets (a school of Greek poets, *c.* 3rd–1st centuries BC); the strophic lyric (a lyric based on a rhythmic system of two or more lines repeated as a unit); and the Sanskrit theatre. It can also be extended to narrative literature, especially the prose novel. The great masters in the *Kāvya* form (which was also exported to Java) were Aśvaghoṣa, Kālidāsa, Bāṇa, Daṇḍin, Māgha, Bhavabhūti, and Bhāravi.

The earliest surviving *kāvya* literature was written by a Buddhist, Aśvaghoṣa, said to have been a contemporary of the Kuṣāṇa (Kushān) king Kaniṣka (1st century AD). Aśvaghoṣa's work also marks a shift away from the Pāli of the Theravāda branch of Buddhism back to the more and more accepted Sanskrit of the Mahāyāna branch. Two works are extant, both in the style of *mahākāvya:* the *Buddhacarita* ("Life of the Buddha") and the *Saundarānanda* ("Of Sundarī and Nanda"). Compared with later examples, they are fairly simple in style but reveal typical propensities of writers in this genre: a great predilection for descriptions of nature scenes, for grand spectacles, amorous episodes, and aphoristic observations. The resources of the Sanskrit language are fully exploited; stylistic embellishments (*alaṅkāra*) of simile and metaphor, alliteration, assonance, and the like are employed, often quite felicitously. The original *Buddhacarita,* rediscovered in 1892, had been known from Tibetan and Chinese translations. The Sanskrit text is fragmentary, breaking off in the 14th canto (major division of the poem) with the enlightenment of the Buddha, while the other versions take the story through the Buddha's Nirvāṇa. Though intended to instruct the reader to turn away from the sensuous life and follow the Buddha's path, the work is at its best in descriptions of that very life. This is even more apparent in the *Saundarānanda,* which recounts a well-known story of how the Buddha converted his half-brother Nanda, who was deeply in love with his wife, Sundarī, and with the good life, to the monastic life of austerity. In his mastery of the intricacies of prosody and the subtleties of grammar and vocabulary, Aśvaghoṣa shows himself the complete forerunner of the Hindu *mahākāvya* authors.

The mahākāvya. In its classical form, a *mahākāvya* consists of a variable number of comparatively short cantos, each composed in a metre appropriate to its particular subject matter. The subject matter of the *mahākāvya* itself is taken from the epic, which is not, however, followed slavishly. Most *mahākāvya*s display such set pieces as descriptions of cities, oceans, mountains, the seasons, the rising of the sun and moon, games, festivals, weddings, embassies, councils, war, and triumph. It is typical of the genre that, while each strophe, or stanza, is intended to be part of a narrative sequence, it more often stands by itself, a discrete unit conveying one idea or developing one image. In this, the tendency of the Rigvedic stanza (see above *Sanskrit: formative period [1200–400 BC]*) continues in the classical literature. Although the lines of the classical stanza are long enough to convey their meaning quite explicitly, it is the pride of the poet to suggest rather than to express. Sometimes this is done by simple collocation of words: for example, in the first line of Kālidāsa's *Meghadūta* a *yakṣa* (a mischievous elf-like creature) is afflicted by a curse, "the more painful because it spelt separation from his beloved"; the next word notes that he had been negligent in his duties; taken together, the two words, though syntactically unrelated, suggest that it was his amour that made him neglect his duties. Another common suggestive device is the double meaning, or play on words. These double meanings often add a certain graceful playfulness to the poetry, reminding one that the poem was written first of all to give pleasure to the man of taste.

Traditionally there are six model *mahākāvya*s, three by Kālidāsa and one each by Bhāravi, Māgha, and Śriharṣa, to which sometimes the *Bhaṭṭikāvya* is added.

The works
of Kālidāsa

Nothing is known with certainty of the life of Kālidāsa, the greatest of Sanskrit poets, but there is substantial agreement that at one time he lived in Ujjayinī (Ujjain, in the present state of Madhya Pradesh), the capital of Avanti and an important centre of Sanskrit culture in a commercially busy area. His name, which means Servitor of Kālī, indicates that he was a follower of that goddess, whom he was to celebrate as Pārvatī, the daughter of the mountain, in the *Kumārasaṃbhava.* Probably he lived during the reign of Candra Gupta II Vikramāditya (*c.* 380–*c.* 415), and there are reports that he died, by the hand of an envious courtesan, while a guest of King Kumāradāsa of Ceylon.

Compared with those of others, Kālidāsa's style might be called simple, but it is a very studied, very felicitous simplicity, hiding the actual complexity of his constructions. In two of his *mahākāvya*s, Kālidāsa draws on epic lore. The first, and probably earlier one, is the *Kumārasaṃbhava* ("Birth of the War God"), which describes the courting of the ascetic Śiva, who is meditating in the mountains, by Pārvatī, the daughter of the Himalayas; the destruction of the god of love (after his arrow has struck Śiva) by the fire from Śiva's third eye; and the wedding and lovemaking of Śiva and Pārvatī, which results in the conception of the war god. The original is in eight cantos, but a sequel was added by an imitator. The second *mahākāvya,* the *Raghuvaṃśa* ("Dynasty of Raghu"), deals with themes from the *Rāmāyaṇa:* it describes the vicissitudes of the Solar dynasty of the ancient Indian barons, culminating in the *Rāmāyaṇa* story of Rāma and Sītā. The *Raghuvaṃśa* is famous for its beautiful descriptions and incidental narratives, which give the poem a somewhat episodic character; among them are a description of the six seasons (spring, summer, rainy, autumn, winter, and dewy) and the story of a young hermit who went to the river to fill a water jar for his parents and was killed by a stray arrow.

Unique in Sanskrit love poetry is Kālidāsa's *Meghadūta,* in which the poet tries to go beyond the strophic unity of the short lyric (see below), which normally characterizes love poems, by stringing the stanzas into a narrative. This innovation did not take hold, though the poem inspired imitations along precisely the same story line. The *Meghadūta* is the lament of an exiled *yakṣa* who is pining for his beloved on a lonely mountain peak. When, at the beginning of the monsoon, a cloud perches on the peak, he asks it to deliver a message to his love in the Himalayan city of Alakā. Most of the poem, composed in an extremely graceful metre, consists of a description of the landmarks, cities, and the like on the cloud's route to Alakā. It must be considered among the finest poems, if not the finest poem, written in Sanskrit. Kālidāsa also wrote for the theatre (see below) and was no doubt the most versatile author of Sanskrit literature; his works became well-nigh canonical models.

Bhāravi (6th century) probably hailed from the south during the reign of the Pallava dynasty. He took up a *Mahābhārata* theme in his *Kirātārjunīya* ("Arjuna and the Mountain Man"), recounting the Pāṇḍava prince Arjuna's encounter and ensuing combat with a wild mountaineer who in the end proves to be the god Śiva. Bhāravi's language and style are more difficult than Kālidāsa's, but the poem is highly regarded in Indian literary tradition.

Māgha, who wrote in the 8th century, was a conscious rival of Bhāravi, whom he attempted to surpass in every respect. His *Śiśupālavadha* ("The Slaying of King Śiśupāla") is based on an episode of the *Mahābhārata* in which the rival King Śiśupāla insults the hero-god Krishna, who beheads him in the ensuing duel. Māgha is a master of technique in the strict Sanskrit sense of luscious descriptions; intricate syntax; compounds that, depending on how they are split, deliver quite different meanings; and the full register of stylistic embellishments.

To some critics, the preoccupation with technique, the triumph of form over substance, appears to have spelled the doom of the *mahākāvya.* A curious but entirely Sanskritic phenomenon, for example, is the *Bhaṭṭikāvya,* a poem by Bhaṭṭi (probably 6th or 7th century). It again deals with the story of Rāma and Sītā, but at the same time it illustrates in stanza after stanza, in exactly the proper sequence, the principal rules of Sanskrit grammar and poetics. Less artificial is the *Naiṣadhacarita* ("The Life of Nala, King of Niṣadha"), written by the 12th-century poet Śrīharṣa and based on the story of Nala and Damayantī in the *Mahābhārata.* An example of another kind of excess indulged in by *mahākāvya* writers is the *Rāmacarita* ("Deeds of Rāma"), by the 12th-century poet Sandhyākāra, which celebrates simultaneously the hero-god Rāma and the poet's own king, Rāmapāla of Bengal. Many other works were written in this style, and, even today, one may encounter a *mahākāvya* treatment of a great man such as Mahatma Gandhi or Jawaharlal Nehru.

Jayadeva's *Gītagovinda*

Difficult to classify is the work of the 12th-century Bengal poet Jayadeva, who wrote the *Gītagovinda* ("Cowherd Song"). The basic structure of this long poem, in which the poet recounts the youthful loves of the cowherd hero and god Krishna, largely based on the story of the *Bhāgavata-Purāṇa,* is that of the *mahākāvya.* Generously interspersed between cantos, however, are erotic-religious lyrics of extremely musical assonances, which were, and still are, sung. Jayadeva's work, rather lacking in the grammatical rigidity of the other *mahākāvya* writers, has been extremely popular and affords a fine example of the devotional lyric (see below).

The short lyric. It is in the short, one-stanza lyric that Sanskrit poetry is revealed most intimately in its real aims. As noted, almost all of high Sanskrit poetry is strophic in fact; in the lyric it is so in intention. It is eminently a genre of the poetic moment, making an aesthetic observation and placing it within the Sanskritic universe of discourse. It may be an observation of anything: a fish glintingly jumping from a pond, aboriginal tribesmen engaged in a bloody rite, love in all its manifestations, a glimpse of God perceived or remembered. But in the monumental lyric collections that have been preserved, and in the many stray verses still circulating among educated Hindus in India as so-called *subhāṣita*s ("well-turned" couplets), the more common topics are praise of the god of one's devotion and the vagaries of love.

In the short lyric it is hard to make a distinction that depends on the language in which it is composed; for, although the language may be different, the subject matter and forms are the same. Many love lyrics, especially when they describe feelings experienced by women, are composed not in Sanskrit but, instead, in one of the Prākrits, or Middle Indo-Aryan languages, among which the dialect called Māhārāṣṭrī is particularly popular. The collection of 700 poems in this language, compiled by Hāla under the name of *Sattasaī* ("The Seven Hundred"), tends to be simpler in imagery and in the emotion portrayed than their Sanskrit counterparts, but essential differences are difficult to pinpoint.

The devotional lyric, a short verse expressing the author's devotion to a god, is linked with both the hymnal poetry of the Rigveda—though far less determined by a desire for compelling magic—and the temple worship of Hinduism. Though by no means always, there is often a particularism about them: the deity is invoked as it appears in a specific iconic stance or in a local temple or in a manifestation especially pleasing to the poet. The number of such verses is countless; every major religious and philosophic leader is held to have added to their stock. Some are especially famous: the *Sūryāṣṭaka* ("Eight Strophes for the Sun"), by Mayūra; the collections attributed to the philosopher Śaṅkara, the *Saundaryalaharī* ("The Wavy River of the Beautiful Sky"); and the *Kṛṣṇakarṇāmṛta* ("The Elixir of Hearing of Krishna"), by Bilvamaṅgala, among others. These *stotra* ("lyrics of praise") quite often were set to music, and people continue to sing them today—without necessarily comprehending the full intention of the Sanskrit, much as hymns in Latin were traditionally sung by Roman Catholic believers.

The love lyric

The entire erotic experience, from budding love to the aftermath of consummation, is represented brilliantly in lyric poetry. But among the many themes inspired by love, poets have been most attracted to the lament of separated lovers. It is mostly the sufferings of the woman that are portrayed, but the grief of the man is also depicted—in Kālidāsa's *Meghadūta,* for example. The love lyrics consist of single verses, many of which seek to suggest the mood of *śṛṅgāra* (physical love). While often extremely erotic, they are very rarely obscene. Sanskrit norm banned all coarse expressions for sexual play; and, although much probably escapes the modern reader, blunt allusions to genital organs are rare and, where allusions occur, extremely veiled. Bodily parts with less overt sexual connotations, such as breasts and buttocks, are frankly mentioned and described—in fact, celebrated. In allusions to sexual intercourse the terminology of the *Kāmasūtra* of Vātsyāyana is frequently invoked, as though this ancient textbook of Indian erudition was a protection against possible opprobrium—not unlike Latin terms resorted to in the West for actions that most know by shorter, more colloquial names.

The erotic and the devotional lyric merge freely, and at times it is impossible to make out whether the free sexual imagery employed is to be taken literally or as an allegory of the human soul courting the love of its god. The task—not a very pressing one—is made more difficult by the fact that some *bhakti* (devotion) religions have developed the poetics of love poetry into a kind of theology, a phenomenon quite characteristic of Bengal Krishnaism (see below *Indo-Aryan literatures: 12th–18th century*).

Authors of *subhāṣita*s often collected them themselves, the favourite form being that of the *śataka* ("century" of verses), in which 100 short lyrics on a common theme were strung together. Mention has been made of Hāla's *Sattasaī* ("The Seven Hundred," consisting of lyrics in the Māhārāṣṭrī dialect). Four well-known Sanskrit collections, of the 7th century, are the famous "century" of Amaru, king of Kashmir, and the three "centuries" by the poet

Bhartṛhari; one of the latter's collections is devoted to love, another to worldly wisdom—a very popular theme in epigrammatic verse—and the third to dispassion. Of the same type but in a different vein is *Caurapañcāśikā* ("Fifty Poems on Secret Love"), in which the 12th-century poet Bilhaṇa fondly recalls the pleasure of his clandestine amours with a local princess.

The theatre. Of all the literary arts, the Indians esteemed the play most highly, and it is in this form that most of the other arts were wedded together. Its origins are obscure, but there is reason to assume that the play developed out of recitations of well-known epic stories by professional reciters. It is an extremely rich genre with a number of outstanding playwrights.

The style is extremely varied. Although it might be called a Sanskrit play, Sanskrit is by no means the only language used, for the less educated characters, including all women, speak Prākrits of different degrees of niceness. The action is carried by prose, but at the least provocation—indeed, at any of the poetic moments characteristic of the strophic lyric—the author reverts to verse, sometimes in mid-sentence. Two principal types of play are distinguished: the *nāṭaka,* which is based on epic material, and the *prakaraṇa,* which is of the author's invention, though often borrowed from narrative literature.

:incipal
pes of
·ama

Characteristic of the Sanskrit theatre are elements of sacrality. The play begins and ends with a benediction, many of which consist of subject matter taken from sacred texts. It is also expressed in numerous taboos: the play must have a happy outcome in which harmony, interrupted by the drama of the play, is restored; improper scenes, such as eating, dressing and undressing, and sexual intercourse, are not to be portrayed; no violence among the higher characters is permitted; war, which often occurs, should simply be reported on, often by lower characters, not in any way staged.

Fragments of Buddhist plays prior to the flowering of Hindu theatre have survived, but no complete plays earlier than 13 ascribed to the playwright Bhāsa. There is considerable controversy over the authenticity of the Bhāsa plays, but at least some of them must be authentic, perhaps dating back to the 3rd century. The plays are based on the epic and on the *Bṛhat-kathā* narrative cycle (see below); among the latter, the *Svapnavāsavadattā* ("The Dream of Vāsavadattā") is the most famous. Of considerable interest also is the *Daridra-Cārudatta* ("The Poverty of Cārudatta"), which became the basis for the play *Mṛcchakaṭika* ("Little Clay Cart") of Śūdraka (see below).

It must be assumed that there was an efflorescence of poetry and theatre in the city of Ujjayinī, one of the capitals of the Gupta Empire, in the 5th century, for a number of authors can be placed there during this reign; among these were Viśākhadatta, Śūdraka, Śyāmilaka, the writer of one of the best farces, and Kālidāsa, who at the beginning of the development of the genre produced some of the greatest plays in the tradition.

ılidāsa's
ıys

Three plays by Kālidāsa remain, one of which is the *Mālavikāgnimitra* ("Agnimitra and Mālavikā"), a harem play of amorous intrigue at a royal court. The other two are based on old themes. *Vikramorvaśī* ("Urvaśī Won by Valour") is based on a story as old as the Rigveda, that of the nymph Urvaśī, who is loved by King Purūravas, whom she marries on the condition that she shall never see him nude. The accident happens, and the nymph returns to heaven, leaving her husband crazed with longing, until a final reunion. But the Indian tradition holds the *Abhijñānaśakuntalā* ("Śakuntalā and the Token of Recognition") to be the greatest of all Sanskrit plays. It recounts a *Mahābhārata* story—rather freely to be sure—of a hermit girl secretly married to a visiting king, who leaves with her a keepsake that will serve her as a token of recognition. She gives birth to a son, Bharata, and goes to the King's court; on the way she loses the ring in a river, where a fish swallows it. The King fails to recognize her and rejects her, and her mother, a nymph, carries her to heaven. When the ring is recovered by a fisherman and the King's memory is restored, he searches for Śakuntalā but does not find her. In the end he meets a boy who proves to be his son and is restored to him.

Kālidāsa's great forte is the portrayal of emotions—ordinary enough in themselves (budding love, love consummated, rejection, despair, a father's love for his son)—but Kālidāsa applies to them a mastery of expression and image that makes the play a work of perennial beauty.

Next to nothing is known of Śūdraka except that he must have hailed from Ujjayinī. His is the most charming of all *prakaraṇa* plays (those that are not based on epic material): the *Mṛcchakaṭikā* ("Little Clay Cart"), the story of an impoverished merchant and a courtesan who love each other but are thwarted by a powerful rival who tries to kill the woman and place the blame on the hero, Cārudatta. The play offers a fascinating view of the different layers of urban society. Viśākhadatta, the author of a rare semihistorical play called *Mudrārākṣasa* ("Minister Rākṣasa and his Signet Ring"), apparently was a courtier at the Gupta court. His play is a dramatization of the Machiavellian political principles expounded in the book *Arthaśāstra,* by Kauṭilya, who appears as the hero of the play.

To the 7th-century king Harṣa of Kanauj are attributed three charming plays: *Ratnāvalī* and *Priyadarśikā,* both of which are of the harem type; and *Nāgānanda* ("The Joy of the Serpents"), inspired by Buddhism and illustrating the generosity of the snake deity Jīmūtavāhana.

The
plays of
Bhavabhūti

Ranked by Indian tradition close to Kālidāsa himself, Bhavabhūti (early 8th century) was the author of three plays, two of which are based on the *Rāmāyaṇa* story. The *Mahāvīracarita* ("The Exploits of the Great Hero") treats of Rāma's battle with Rāvaṇa and the *Uttararāmacarita* ("The Later Deeds of Rāma") treats of the life of Rāma after he has abandoned Sītā. Bhavabhūti lacks the elegance and grace of Kālidāsa but is more pensive—even brooding—than his predecessor. His style is also very forceful. His *prakaraṇa Mālatī-Mādhava* ("Mālatī and Mādhava") is a complex love intrigue intermingled with sorcery and Tantric practices, including a human sacrifice and much violence.

This list by no means concludes that of the playwrights in the Sanskrit tradition. The writing of plays, mostly derivative from the great models, has continued until the present day.

Apart from the more seriously intended plays described above, the Sanskrit theatre also has a rich repertory of farces, which are usually in one act. Most interesting of these are the *bhāṇa*s, which may be monologues in which an actor addresses imaginary persons and is answered by them, as he paints a picture of town life full of personal and social satire. Among the best in this little-studied genre is Śyāmilaka's 5th-century *Pādataḍitaka* ("The Courtesan's Kick").

Narrative literature. Sanskrit narrative literature is extremely rich, so rich in fact that at one time it was believed that all folktales originally came from India. Many indeed have, and they have found a place in *The Arabian Nights' Entertainment,* Boccaccio's *Decameron,* and other such works down to the fairy tales of Hans Christian Andersen and the fables of Jean de La Fontaine. Certain collections of animal tales, some of which go back to the Buddhist *Jātaka* stories, had incredible histories. The most famous is the *Pañca-tantra* ("The Five Chapters"), which, within a framework of a lesson in the art of politics addressed to young princes, presents a number of animal characters who in their actions both admonish and exhort the reader to a life certain to lead to worldly success. A shorter version, partly drawn from the *Pañca-tantra,* is the *Hitopadeśa* ("Good Advice"). The *Pañca-tantra* found its way to the West through translations into Persian, Arabic, Syrian, Hebrew, and Latin, until most of the medieval literatures possessed their own versions of it. No less extensive were its migrations to Southeast and East Asia.

The *Pañca-*
tantra

The principal work of the novelistic and picaresque tale is the *Bṛhat-kathā* ("Great Story") of Guṇāḍhya, written in Prākrit and now lost, save for Sanskrit retellings. The most important among these Sanskrit versions is the *Kathāsaritsāgara* ("Ocean of Rivers of Stories") of Somadeva (11th century), which includes so many subsidiary tales that the main story line is frequently lost. Perhaps more faithful to the original—in any case far less distracting—is the *Bṛhatkathāślokasaṃgraha* ("Summary in Verse of the

Great Story"), by Budhasvāmin (probably 7th century), one of the most charming of Sanskrit texts. Other collections of tales include the *Vetāla-pañcaviṃśatikā* ("Twenty-five Tales of a Ghost"), *Śūkasaptati* ("The Seventy Stories of a Parrot"), and the *Siṃhāsana-dvātrim-sātikā* ("Thirty-two Stories of a Royal Throne").

Related to the *Bṛhat-kathā* cycle, though the exact relationship is unclear, is the Jain Prākrit text of the *Vāsudevahiṇḍī,* "The Roamings of Vāsudeva" (before 6th century), describing the acquisition of numerous wives by Krishna Vāsudeva.

Though the tales are often artless, sometimes they are elaborately embroidered in the Sanskrit *kāvya* style. A fine example is the *Daśakumāracarita* ("Tales of Ten Princes"), by Daṇḍin (6th/7th century), in which, within the framework of a boxing story, the picaresque adventures of 10 disinherited princes are described in prose. While tending overly to description, the work remains eminently readable for the modern reader, a quality that cannot be attributed to the prose novels of the 7th-century writer Bāṇa: the *Harṣacarita,* "The Life of Harṣa" (king of Kanauj and the author of three plays, discussed above), which is important for its information on culture and society; and the *Kādambarī* (the name of the heroine), which describes the affairs of two sets of lovers through a series of incarnations, in which they are constantly harassed by a cruel fate. (J.A.B.v.B.)

DRAVIDIAN LITERATURE: 1ST–19TH CENTURY

Of the four literary Dravidian languages, Tamil has been recorded earliest, followed by Kannada, Telugu, and Malayalam. Tamil literature has a classical tradition of its own, while the literatures of the other languages have been influenced by Sanskrit models.

Early Tamil literature (1st–10th century). *Caṅkam literature.* Early classical Tamil literature is represented by eight anthologies of lyrics, 10 long poems, and a grammar called the *Tolkāppiyam* ("Old Composition"). According to a fanciful Tamil tradition, this literature was produced by poets of three "academies," or *caṅkams,* that in the hoary past were centred in the southern Indian city of Madurai and supposedly lasted 4,400, 3,700, and 1,850 years, respectively. The *Tolkāppiyam* was ascribed to the second *caṅkam,* the eight anthologies and 10 long poems to the third; according to tradition, nothing is extant from the first *caṅkam.* The early literature, itself known as *caṅkam,* comprises 2,381 poems, ranging from four to nearly 800 lines each and assigned to 473 poets who are known by name or epithet; about 100 poems are anonymous. Though the literature does not go back as far as native tradition would have it, it is generally ascribed to the first three centuries of the Christian Era and represents the oldest non-Sanskrit literature to be found on the South Asian subcontinent.

The eight anthologies and their contents, excluding opening invocations that were added later, are as follows: *akam* anthologies consisting of (1) *Kuṟuntokai,* 400 love poems, (2) *Naṟṟiṇai,* 400 love poems, (3) *Akanāṉūṟu,* 400 love poems, (4) *Aiṅkuṟunūṟu,* 500 love poems, each 100 (assigned to a different poet) dealing with one of five phases of love, (5) *Kalittokai,* 150 love poems in a metre called *kali;* and *puṟam* anthologies consisting of (6) *Puṟanāṉūṟu,* 400 poems, (7) *Patiṟṟuppattu* ("The Ten Tens"), 100 poems on kings (the first and last decades are missing), and (8) *Paripāṭal,* a collection of 70 religious poems. *Paripāṭal* and *Kalittokai* appear to be the latest of the anthologies; *Kuṟuntokai* and *Puṟanāṉūṟu* probably contain the earliest compositions. The remarkable work of grammar and rhetoric, *Tolkāppiyam,* is the crucial text for an understanding of early Tamil language and literature. Divided into three sections (each consisting of *cūttirams*, or aphorisms)—sounds, words, and meaning—the *Tolkāppiyam* details, in the third, the canons of *caṅkam* poetic traditions.

Characteristics of *caṅkam* poetry

In the *Tolkāppiyam* and the anthologies, poems are classified by theme into *akam* ("interior") and *puṟam* ("exterior"), the former highly structured love poems, the latter heroic poems on war, death, personal virtues, the ferocity and glory of kings, and the poverty of poets. Both the *akam* and the *puṟam* had well-defined *tiṇai*s (genres) that paralleled one another: *e.g.,* the *kuṟiñci* genre, in love poetry, which dealt with the lovers' clandestine union on a hillside by night; and the *veṭci* genre, in heroic poetry, which dealt with the first onset of war, by nocturnal cattle stealing. Both *kuṟiñci* and *veṭci* are names of flowers that grow on the hillside, here symbolic of the poetic genre, the mood, and the theme. By such pairings across *akam* and *puṟam,* love and war become part of the same universe and metaphors for one another; the same poets—for example, Paraṇar and Kapilar—wrote great poems in both genres. The basic technique depended on a taxonomy of Tamil nature and culture, of culturally defined time, space, nature, and human nature. For example, matched in metaphor with five phases of *akam* love (union; infidelity; anxious waiting; patient waiting; and the lover or lovers eloping or journeying for wealth, knowledge, and so on) are six seasons, six parts (dawn, forenoon, noon, afternoon, evening, and night) of the day, and five landscapes (hill, seashore, forest, pasture, and wasteland, named after characteristic flowers—*kuṟiñci, neytal, mullai,* and *marutam*—and the evergreen tree, *pālai*) and their contents (including gods, foods, birds, beasts, drums, occupations, lutes, musical styles, flowers, and kinds of running or standing water). Each landscape becomes a repertoire of images—anything in it, bird or drum, tribal name or dance, may evoke a specific feeling. A favourite poetic device is *uḷḷuṟai* (*i.e.,* metonymy, a figure of speech consisting of the description of one thing used to evoke that of another with which it is associated). Thus, the natural scene implicitly evokes the human scene; for example, bees making honey out of *kuṟiñci* flowers evokes the lovers' union. Not only is the poet's language Tamil, but the landscapes, the personae, and the appropriate moods and situations formulate the realities of the Tamil world into a code of symbols. For some five or six generations, the *caṅkam* poets spoke this common language of symbols, creating a body of lyrical poetry probably unequalled in passion, maturity, and delicacy by anything in any Indian literature.

Eighteen Ethical Works. The *Patiṟeṇ-kīṟkkaṇakku* ("Eighteen Ethical Works"), usually dated as post-*caṅkam* (4th–7th centuries), are all affected by Jainism and Buddhism. Of these the *Tirukkuṟaḷ* ("Sacred Couplets"), ascribed to Tiruvaḷḷuvar, is the most celebrated. Its 1,330 hemistichs (half lines of verse) are probably the final distillation of different periods. There are many parallels in the work to the Sanskrit *Kāma-sūtra,* the treatise on erotic love, to *Manu-smṛti,* an ancient treatise on special obligation and religious law, and to *Artha-śāstra,* Kauṭilya's treatise on politics. The *Kuṟaḷ* has three sections: *aṟam,* or virtue (Sanskrit *dharma*); *poruḷ,* government and society (Sanskrit *artha*); and *kāmam,* love (Sanskrit *kāma*). There is no special treatment of *mokṣa,* or salvation, though *aṟam* seems to include it. In the *aṟam* (virtue) section, the *Kuṟaḷ* sums up a world-affirming wisdom, the wisdom of human sympathy, expanding from wife, children, and friends to clan, village, and country. In the *poruḷ* (government and society) section, the aphorisms project a vision of an ideal state, based on educated human nature, and relate the good citizen to the good man. Prostitution, disease, drink, and gambling are listed, with foreign enemies, as dangers to the state. In the *kāmam* (love) section, the *Kuṟaḷ* follows the *caṅkam's* love—eros, or sexual love—yet anticipates agape, the perfecting of love through many lives, which appears in religious poetry of the next age.

Epics. The age of the Pallavas (300?–900), a warrior dynasty of Hindu kings, is known for its epics, beginning with *Cilappatikāram* ("The Jewelled Anklet") and *Maṇimēkalai* ("The Girdle of Gems") and including an incomplete narrative, *Peruṅkatai* ("The Great Story"), the *Cīvakacintāmaṇi* ("The Amulet of Cīvakaṉ") by Tiruttakkatēvar, and *Cūḷāmaṇi* ("The Crest Jewel") by Tōlāmoḻittēvar. The last three works depict Jaina kings and their ideals of the good life, nonviolence, and the attainment of salvation through self-sacrifice. They are also characterized by excellent descriptions of city and country and by a mixture of supernatural and natural elements. In their episodic methods of narration and set descriptions of erotic, heroic, and religious themes, these Jaina epics became both models and sources for later epic works.

'he
'ilappati-
āram

The *Cilappatikāram,* by Iḷaṅkō Aṭikaḷ, is in three books, set in the capitals of the three Tamil kingdoms: Pukār (the Cōḻa capital), Maturai (*i.e.,* Madurai, the Pāṇṭiya [Pāṇḍya] capital), and Vañci (the Cēra capital). The story is not about kings but about Kōvalaṉ, a young Pukār merchant, telling of his marriage to the virtuous Kaṇṇaki, his love for the courtesan Mātavi, and his consequent ruin and exile in Maturai, where he dies, unjustly executed when he tries to sell his wife's anklet to a wicked goldsmith who had stolen the Queen's similar anklet and charged Kōvalaṉ with the theft. Kaṇṇaki, the widow, comes running to the city and shows the King her other anklet, breaks it to prove it is not the Queen's—Kaṇṇaki's contains rubies, and the Queen's contains pearls—and thus proves Kōvalaṉ's innocence. Kaṇṇaki tears off one breast and throws it at the kingdom of Maturai, which goes up in flames. Such is the power of a faithful wife. The third book deals with the Cēra king's victorious expedition to the north to bring Himalayan stone for an image of Kaṇṇaki, now become a goddess of chastity (*pattiṉi*).

The *Cilappatikāram* is a fine synthesis of mood poetry in the ancient Tamil *caṅkam* tradition and the rhetoric of Sanskrit poetry—even the title is a blend of Tamil and Sanskrit—including in the epic frame *akam* lyrics, the dialogues of *Kalittokai* (poems of unrequited or mismatched love), chorus folk song, descriptions of city and village, lovingly technical accounts of dance and music, and strikingly dramatic scenes of love and tragic death. One of the great achievements of Tamil genius, the *Cilappatikāram* is a detailed poetic witness to Tamil culture, its varied religions, town plans and city types, the commingling of Greek, Arab, and Tamil peoples, and the arts of dance and music.

Maṇimēkalai (the heroine's name, "Girdle of Gems"), the second, "twin," epic (the last part of which is missing), by Cātaṉār, continues the story of the *Cilappatikāram;* the heroine is Mātavi's daughter, Maṇimēkalai, a dancer and courtesan like her mother. Maṇimēkalai is torn between her passion for a princely lover and her spiritual yearnings, the first encouraged by her grandmother, the second by her mother. She flees the attentions of the prince, and, while he pursues her, she attains magical powers: she changes forms; survives prison, lecherous villains, and other dangers; converts the Queen; and finally goes to Pukār, which is being destroyed by oceanic erosion, worships Kaṇṇaki, and arrives in Vañcī to work in famine relief and to perform "penance." Unlike the *Cilappatikāram,* the *Maṇimēkalai* is partisan to Buddhism. It is known for its poetry and its lively discussions of religion and philosophy.

Bhakti poetry. From the 6th century onward, a movement with religious origins made itself heard in literature. The movement was that of *bhakti,* or intense personal devotion to the two principal gods of Hinduism, Śiva and Vishnu. The earliest *bhakti* poets were the followers of Śiva, the Nāyaṉārs (Śiva Devotees), whose first representative was the poetess Kāraikkāl Ammaiyār, who called herself a *pēy,* or ghostly minion of Śiva, and sang ecstatically of his dances. Tirumūlar was a mystic and reformer in the so-called Siddhānta (Perfected Man) school of Śaivism, which rejected caste and asceticism, and believed that the body is the true temple of Śiva. There were 12 early Nāyaṉār saints. Similar poets, in the tradition of devotion to the god Vishnu, also belonged to this early period. Called Āḻvārs (Immersed Ones), they had as their first representatives Poykai, Pūtaṉ, and Pēyār, who composed "centuries" (groups of 100) of linked verses (*antāti*), in which the final line of a verse is the beginning line of the next and the final line of the last verse is the beginning of the first, so that a "garland" is formed. To these Āḻvārs, God is the light of lights, lit in the heart.

ıe
āyaṉārs

The most important Nāyaṉārs were Appar and Campantar, in the 7th century, and Cuntarar, in the 8th. Appar, a self-mortifying Jain ascetic before he became a Śaiva saint, sings of his conversion to a religion of love, surprised by the Lord stealing into his heart. After him, the term *tēvāram* ("private worship") came to mean "hymn." Campantar, too, wrote these personal, "bone-melting" songs for the common man. Cuntarar, however, who sees a vision of 63 Tamil saints—rich, poor, male, female, of every caste and trade, unified even with bird and beast in the love of God—epitomizes *bhakti.* To him and other Bhaktas, every act is worship, every word God's name. Unlike the ascetics, they return man to the world of men, bringing hope, joy, and beauty into religion and making worship an act of music. Their songs have become part of temple ritual. Further, in *bhakti,* erotic love (as seen in *akam*) in all its phases became a metaphor for man's love for God, the lover.

In the 9th century, Māṇikkavācakar, in his great, moving collection of hymns in *Tiruvācakam,* sees Śiva as lover, lord, master, and guru; the poet sings richly and intimately of all sensory joys merging in God. Minister and scholar, he had a child's love for God.

Āṇṭāḷ (8th century), a Vaiṣṇava poetess, is literally lovesick for Krishna. Periyāḻvār, her father, sings of Krishna in the aspect of a divine child, originating a new genre of celebrant poetry. Kulacēkarar, a Cēra prince, sings of both Rāma and Krishna, identifying himself with several roles in the holy legends: a *gopī* in love with Krishna or his mother, Devakī, who misses nursing him, or the exiled Rāma's father, Daśaratha. Tiruppāṇāḻvār, an untouchable poet (*pāṇaṉ*), sang 10 songs about the god in Śrīraṅgam, his eyes, mouth, chest, navel, his clothes, and feet. To these Bhaktas, God is not only love but beauty. His creation is his jewel; in separation he longs for union, as man longs for him. Tirumaṅkaiyāḻvār, religious philosopher, probably guru (personal religious teacher and spiritual guide in Hinduism) to the Pallava kings, and poet of more than 1,000 verses, was apparently responsible for the building of many Vaiṣṇava temples. The last of the Āḻvārs, Nammāḻvār (Our Āḻvār), writing in the 9th century, expresses poignantly both the pain and ecstasy of being in love with God, revivifying mythology into revelation.

Period of the Tamil Cōḻa Empire (10th–13th century). The next period, the time of the Tamil Cōḻa Empire (10th–13th centuries), saw an awakening of neighbouring literatures: Kannada, Telugu, and Malayalam. The first extant Kannada work is the 9th-century *Kavirājamārga* ("The Royal Road of Poets"), a work of rhetoric rather indebted to Sanskrit rhetoricians, containing the first descriptions of the Kannada country, people, and dialects, with references to earlier works. From the 10th century on, *campū* narratives (part prose, part verse) became popular both in Kannada and in Telugu, as did renderings of the Sanskrit epics *Rāmāyaṇa* and *Mahābhārata* and Jaina legends and biography.

In Kannada, this period was dominated by the "three gems" of Jaina literature, Pampa, Ponna, and Ranna, as well as by Nāgavarma I, a 10th-century Kannada grammarian. Pampa was the *ādikavi* ("first of poets"), having attained that stature with two great epics: *Vikramārjuna Vijaya* and *Ādipurāṇa.* The former is a rendering of the *Mahābhārata,* with the hero, Arjuna, identified with the poet's royal patron, Arikēsarī. This felicitous epic is known for its succinct, powerful characterizations, its rich descriptions of Kannada country and court, its moving sentiments, and its harmonious blend of Sanskrit and Kannada. While the *Vikramārjuna* is a secular work, Pampa's *Ādipurāṇa* tells the story of the Jaina hero-saint Purudēva, his previous lives, his life from birth to marriage to holy death, as well as the lives of his sons, Bharata and Bāhubali.

Telugu had its *ādikavi* ("first of poets"), in the Brahmin Nannaya Bhaṭṭa (1100–60), who, in *campū* style, wrote three books of a version of the *Mahābhārata,* later finished by Tikkana (13th century) and by Errāpraggaḍa. Like other regional versions of the *Mahābhārata,* the Telugu version is not a literal translation but an interpretation, with many local elements and differences of emphasis; for example, Nannaya emphasizes the importance of Vedic religion. Such works have made the Sanskrit epics and *Purāṇas* part of a live and growing tradition, both oral and literary, in the regional language.

The work
of Kampaṉ

This period also saw the eminence of Kampaṉ's Tamil version of the *Rāmāyaṇa* (12th century). In him there is a climactic blend of earlier *caṅkam* poetry, Tamil epics, the Āḻvārs' fervour of personal *bhakti* (devotion) toward Rāma, folk motifs, and Sanskrit stories, metres, and poetic devices. Instead of a just king and a perfect man, Rāma is an incarnation of Vishnu and an intense object of devotion, dwarfing the Vedic gods; Kampaṉ called his work *Irāmāvatāram* ("Rāma's Incarnation"); yet the emphasis is not on Vishnu but on *dharma* ("the law"), localized and Tamilized. More like Sanskrit than *caṅkam* poets, Kampaṉ revels in elaborate metaphor, hyperbole, and fanciful descriptions of virtue and nature. The work is long, consisting of about 40,000 lines; the *Yuttakāṇṭam* ("War Canto") alone, with 14 battles, equals the *Iliad* in length. The poem is also justly known for its variety of style, its exploitation of the resources of Tamil and Sanskrit both in form and content, its humour, and its handling of the narrative, dramatic, and lyric modes.

Kampaṉ's popularity extended not only into all of Tamil country but apparently into the north, influencing some episodes of Tulsī's Hindi version of the *Rāmāyaṇa,* and into northern Kerala, where 32 plays based on Kampaṉ are enacted ritually with marionettes in Śiva temples.

Pre-15th-century Tamil influence on early Malayalam, the language of Kerala, was strong and led to the literature of *pāṭṭu* ("song"), in which only Dravidian, or Tamil, phonemes may occur and Tamil-like second-syllable rhymes are kept. The best known *pāṭṭu* is *Rāmacaritam* (*c.* 12th–13th century; "Deeds of Rāma"), probably the earliest Malayalam work written in a mixture of Tamil and Malayalam. Other *pāṭṭu*s in Tamilized Malayalam, written by a family of poets (14th–15th centuries) from Niraṇam in central Travancore, appear in *Kaṇṇassan Pāṭṭukaḷ,* in which Tamil conventions of metre and phonology are loosened and more Sanskrit is allowed. Similar in style is a version of the *Rāmāyaṇa* by Rāma Paṇikkar, an abridged *Bhagavadgītā* by his uncle Mādhava Paṇikkar, and a condensed *Mahābhārata* and the 10th book of the *Bhāgavata-Purāṇa* by another uncle, Śaṅkara Paṇikkar.

As strong as Tamil influence was, the predominant influence on Malayalam was Sanskrit, in language as well as literary form. The influence on language led early to a mixture of Sanskrit and Malayalam in a literary dialect called *maṇipravāḷa* (meaning "necklace of diamonds and coral"). The author of the *Līlātilakam,* a 14th-century treatise on grammar and poetics, describes both the Tamilizing and Sanskritizing trends and genres and insists on harmonious blendings. Many kinds of poems were composed in *maṇipravāḷa* styles: *kūḍyāṭṭam*s (dramatic presentations using Sanskrit *śloka*s, or epic metres, for hero and heroine, *maṇipravāḷa* for the clown, and Malayalam for explanations intended for the laity); didactic works such as the 11th-century *Vaiśikatantram* ("Advice to a Courtesan by Her Mother"); 13th- and 14th-century *campūs* (narratives combining prose and verse) on dancers, such as *Unniyati Caritam* by Dāmōdara Cākkiyār; and several short poems in praise of women and kings. *Maṇipravāḷa* poems like these are essentially artifical expressions of courtly high-caste poets, preoccupied with eroticism and harlots. The *Candrōtsavam* (*c.* 1500; "Moon Festival") is a satire on the voluptuary *maṇipravāḷa* tradition, jostling together all the famed courtesans of the period.

Coexisting with the Tamilized and Sanskritized Malayalam poems produced by scholars was a live *pacca* ("pure, fresh") Malayalam tradition represented mostly by folk songs and ballads—for example, *Vaḍukkan Pāṭṭukaḷ* (hero ballads of the northern Malabar Coast); songs sung during weddings, deaths, or festivals; and work songs. All three styles—the indigenous folk style, the Tamil, and the Sanskrit—began to converge and influence each other by the 15th century in works such as *Kṛṣṇa Pāṭṭu* ("Song of Krishna") and *Gāthā* ("Song"). Such grafting reached its full flowering in the 16th-century poet Eḻuttaccan (Father [or Leader] of Letters), who popularized the *kiḷippāṭṭu* ("parrot song"), a genre in which the narrator is a parrot, a bee, a swan, and so on. His outstanding works are *Adhyātma Rāmāyaṇam, Bhāratam,* and *Bhāgavatam,* all based on Sanskrit originals yet powerfully re-created with masterly language craft.

While Vaiṣṇava works were proliferating in Malayalam, Śaiva movements swept the other three languages, Tamil, Kannada, and Telugu. In Tamil, the hymns of the Nāyaṉārs were arranged and anthologized for scriptural and recitative use by the 11th century. Another such consolidation of sacred materials was Cēkkiḻār's 12th-century *Toṇṭar Purāṇam,* or *Periyapurāṇam,* narrating in epic style the lives of the 63 great Śaiva saints and creating a tradition for all Śaivas, even in the Kannada and Telugu areas. The theology of the Siddhānta (Perfected Man) school of Śaivism was elaborated in Meykaṇṭār's *Civañāṉapōtam* (13th century).

The new Kannada genre of the 13th century

By the 12th century, a new Kannada genre, the *vacana* ("saying" or "prose poem"), had come into being with the Vīraśaiva saints. In the language of the people, the saints expressed their radical views on religion and society, rejected both Brahminical ritualism and Jaina ascetic world negation, called all men to the *anubhāva* ("experience") of God, and broke the bonds of caste, creed, and sexual difference. Five important poet-saints were Dāsimayya; Basava, a self-searching social reformer and a minister of the Jaina king Bijjaḷa; Allama Prabhu, the elder and metaphysical master of them all; Mahādēviyakka, a woman saint singing love poems to Śiva; and Cannabasava, a brilliant theologian of the movement, who elaborated the theory of "six stages" of mystic ascent for the devotee. The many-facetted lyrics written by the poet-saints were in the spoken dialects of Middle Kannada, yet they drew on archetypal human images as well as ancient pan-Indian symbology for their intense and searing expressions of *bhakti.* Inspired by these lyrics, Harihara, in the late 12th century, wrote some 120 *ragaḷe* (blank verse) biographies of the Śaiva saints, including the Vīraśaiva (or Liṅgāyat) and the earlier Tamil Nāyaṉārs. In the early 13th century, his disciple and nephew, Rāghavāṅka, wrote, in *ṣaṭpadi*s (six-line stanzas), of the lives of saints, in well-structured works such as *Sōmanātha Carite* and *Siddharāma Caritra;* his most mature work is *Hariścandrakāvya,* an unequalled reworking of an ancient Job-like story of Hariścandra, who suffered every ordeal for his love of truth. The Vīraśaiva saints' lives and the *vacana* ("saying" or "prose poem") literature were codified in a masterpiece called *Śūnya Sampādane* ("The Achievement of Nothing"), consisting of dialogues interweaving the saints' *vacana*s, with the poet Allama Prabhu as the central figure.

Contemporary with the 13th-century Vīraśaiva saints were Telugu Śaiva poets such as Pālkuriki Sōmanātha, who composed the *Basavapurāṇam* employing popular metres and idiomatic Telugu. His *Paṇḍitārādhya Caritra* is a life of the Śaiva devotee Paṇḍitārādhya as well as a book of general knowledge including social customs, arts, crafts, and particularly music. His *Vṛṣādhipa Śatakam* consists of verses in Tamil, Kannada, Marathi, Sanskrit, and Telugu. This work was probably the first of the genre of *śataka*s ("centuries" of verses) literature, particularly popular in Telugu but also written in the other three languages as well as in Sanskrit (see above *Sanskrit: formative period* [*1200–400 BC*]).

Also of the 13th century is Āṇḍayya's *Kabbigara Kāva* ("The Poet's Defender"), in Kannada, a linguistic tour de force, eschewing unmodified Sanskrit forms; and Mallikārjuna's *Sūktisudhārṇava,* an excellent Kannada anthology of lyrics and passages. From 1240 to 1326, poets of Telugu produced over 100 verse renderings of the Sanskrit epic *Rāmāyaṇa* and many more in prose, the earliest being *Raṅganātha Rāmāyaṇa,* assigned to Gōna Buddhā Reḍḍi.

14th–19th century. The next age, from the 14th to the 16th century, is the great age of the Vijayanagar Empire. In this period, Kannada and Telugu were under the aegis of one dynasty and were also hospitable to the influence of neighbouring Muslim Bahmanī kingdoms. Śrīnātha was a 15th-century poet honoured in many courts for his scholarship, poetry, and polemics. He rendered Sanskrit poems and wrote *Haravilāsam* (Four Śaiva Tales); *Krīḍābhirāmam,* a charming, often vulgar account of social life in Warangal; and *Palanāṭi Vīra Caritra,* a popular ballad on a fratricidal war. Many erotic *cāṭu*s, or stray epigrams, are also attributed to him. Bammera Pōtana, a great Śaiva devotee in life and poetry, unschooled yet a scholar, is widely known for his *Bhāgavatam,* a masterpiece that is said to excel the original Sanskrit *Bhāgavata-Purāṇa.* Tāḷḷapāka Annāmācārya, son of a great family of scholars, fathered an exciting new genre of devotional song, all addressed to the god Śrī Veṅkaṭeśvara of Tirupati (a form of Vishnu). His *Saṅkīrtana Lakṣaṇam* is a collection of 32,000 songs in Sanskrit and Telugu, which made a significant contribution to Karnatic (southern Indian) musical technique.

The 16th century was an age of patronage by Vijayanagar kings, beginning with Kṛṣṇa Dēva Rāya, himself a poet versed in Sanskrit, Kannada, and Telugu. The *rāyala yugam* ("age of kings") was known for its courtly *prabandhas*, virtuoso poetic narratives by and for pandits (learned men). Among the most famous court poets were Piṅgaḷi Sūranna, whose verse novel, *Kalāpurṇōdayam* (1550)—a story full of surprises, magic, and changes of identity—is justly celebrated for its artistry; and Tenāli Rāmakṛṣṇa, known for his clownish pranks and humour, whose writings are the centre of a very popular cycle of tales in all four Dravidian languages.

During the 16th century and for the next few centuries, Telugu poets also flourished outside the Telugu country, especially in Tanjore (Thanjavūr) and Madurai, in Tamil country, and Pudukkoṭṭa and Mysore, in Kannada country. Their most important contribution was to native Kannada and Telugu dance drama on mythological themes, called *yakṣagāna.* The form is comparable to *kathākali* in the Malayalam area and to *terukkūttu* ("street drama") and *kuṟavañci* ("gypsy drama") in the Tamil area. The earliest Telugu *yakṣagāna* text is *Sugrīva Vijayam* (*c.* 1570), by Kandukuru Rudra Kavi; the earliest in Kannada is probably Śāntavīra Dēśika's *Saundarēśvara* (1678). The most celebrated of Kannada *yakṣagāna* dramatists is the versatile Pārti Subba, who flourished around 1800 and is known for his moving *Rāmāyaṇa* episodes and songs.

ontribu-
on of
oets to
ative
ance
rama

The 15th and 16th centuries produced some of the most popular classics in Kannada. Of these the greatest is Gadugu's *Kumāra Vyāsa,* or *Nāraṇappa*'s, 10 cantos of the *Mahābhārata;* recited in assemblies as well as in households, these are a continual delight, abounding in humour, passion, and memorable poetry. In *Prabhuliṅgalīle,* Cāmarasa made poetry out of the life of the poet-saint Allama. The *Jaimini Bhārata* and the many versions of *Rāmāyaṇa* episodes (especially Sītā's abandonment in the forest) written by the distinguished Śaiva epic poet Lakṣmīśa are known for their melodious verses and moving scenes. Ratnākaravarṇi's *Bharateśa Vaibhava* is a great Jaina story, tersely told in a Kannada song metre and celebrated for its depiction of many *rasas* ("moods"), especially the erotic.

Kannada Vaiṣṇava *dāsa*s ("servants [of God]") wrote in a song genre called *pada,* parallel and often indebted to the Vīraśaiva *vacana*s ("sayings" or "prose poems"). Purandaradāsa, a rich 16th-century merchant turned mendicant, saint, and poet, composed *bhakti* (devotional) songs on Viṭṭhala (a manifestation of the god Vishnu), criticizing divisions of caste and class and calling on the mercy of God. His *pada*s and *kīrtanas* ("lauds") are also landmarks in Karnatic music. Karnatic music. Kanakadāsa, his contemporary and a shepherd by birth, wrote *pada*s and longer popular works. *Dāsa* songs are part of the repertory of all South Indian musicians.

The folk *tripadi* ("three-line verse") of Sarvajña (1700?) is a household word for wit and wisdom, like the *Kuṟaḷ* in Tamil (see above *Eighteen Ethical Works*) and the "century" of four-line verses in Telugu by Vēmana (15th century). The moral, social, satiric, and wise proverb-like aphorisms of Vēmana and Sarvajña are widely quoted by pundit and layman alike. Equally popular in the Malayalam region is the 18th-century folk poet of *tuḷḷal*s (a song-dance form), Kuñcan Nampiyār, unparalleled for his wit and exuberance, his satiric sketches of caste types, his versions of Sanskrit *Purāṇa* narratives projected on the backdrop of Kerala, and his humorous renderings even of mythic characters.

The 17th and 18th centuries also saw Tamil court poetry—*Purāṇas,* translations from the Sanskrit, and praise poems, known more for their learning and imitative character than for their genius. This was also a period of many schisms and the founding of monasteries in Śaivism and Vaiṣṇavism, which led to many sectarian and polemic works. Muslims and Christians also wrote epics in the Hindu *Purāṇa* style; for example, Umaṟu-p-pulavar's 17th-century *Cīṟā-p-purāṇam,* on the life of the prophet Muḥammad, and Father Beschi's *Tēmpāvaṇi,* on the life of St. Joseph, with echoes from both Kampaṉ and the 16th-century Italian poet Torquato Tasso.

Probably the most impressive Tamil poetry of this period is that of Arunakiriv's learned and melodious *Tiruppukaḻ* (praise of Munikaṉ) and of the Cittars, eclectic mystics known for their radical, fierce folk songs and common-speech style. Tāyumāṉavar (18th century) and Paṭṭiṉattār (and later, in the 19th century, Rāmaliṅkar) are poets of unconditioned love, self-search, and rejection of corrupt society.

The 17th and 18th centuries are also periods of datable folk expression, which include many *tiruviḷaiyāṭal* ("stories of God's sport") *purāṇas;* temple tales (about miracles that took place in the temple); *kuṟavañci* (*i.e.,* "gypsy," a kind of musical dance drama); *paḷḷu*s (plays about village agricultural life); realistic *noṇṭi-nāṭakam*s ("dramas of the lame"), in which a Hindu temple god cures lameness; *kummi* songs sung by young girls, clapping as they dance round and round; and *ammāṉai* ballads. Noteworthy historical ballads are *Kaṭṭa Pommaṉ,* about a chieftain who revolted against the British, and *Tēciṅku-rācaṉ Katai,* about the prince of Gingi and his Muslim friend. Malayalam *āṭṭakkatha,* the literature associated with *kathākali,* the complex traditional dance drama, was also written during this period. Royal poets such as Kōṭṭayattu Tampurān, in the 17th century, and Kārttika Tiruṇal, in the 18th, wrote *āṭṭakkathās.* (A.K.R.)

INDO-ARYAN LITERATURES: 12TH–18TH CENTURY

It is difficult to pinpoint the time when the Indo-Aryan dialects first became identifiable as languages. Around the 10th century AD, Sanskrit was still the language of high culture and serious literature, as well as the language of ritual. The spoken language, however, had continued to develop, and at the turn of the millennium there began to appear, at different times during the subsequent two or three centuries, the languages now known as the regional languages of the subcontinent: Hindi, Bengali, Kashmiri, Punjabi, Rajasthani, Marathi, Gujarati, Oriya, Sindhi (which did not develop an appreciable literature), and Assamese; Urdu did not develop until much later (see below *Islāmic literatures: 11th–19th century*).

Regional
languages

The literatures in their early stages show three characteristics: first, a debt to Sanskrit that can be seen in their use of Sanskrit lexicon and imagery, in their use of myth and story preserved in that refined language, and frequently in their conformity to ideals and values put forward in Sanskrit texts of poetics and philosophy; second, a less obvious debt to their immediate Apabhraṃśa past (dialects that are immediate predecessors of the modern Indo-Aryan vernaculars); third, regional peculiarities.

The narratives in the early stages of the development of the languages are most often mythological tales drawn from the epics and *Purāṇas* of classical Hindu tradition (see above *Sanskrit, Pāli, and Prākrit literatures: 1400 BC–AD 1200*), though in later times, in the 17th and 18th centuries, secular romances and heroic tales were also treated in narrative poems. Although the themes of the narratives are based on *Purāṇa* tales, often they include materials peculiar to the area in which the narrative was written.

In addition to themes, regional literatures frequently borrowed forms from the Sanskrit; for example, the *Rāmāyaṇa* appears in a 16th-century Hindi version by Tulsīdās, called the *Rāmcaritmānas* ("Lake of Rāma's Deeds"), which has the same form, though a different emphasis, as the Sanskrit poem. The stylized conventions and imagery of Sanskrit court poetry also appear, though here, too, with different emphasis; for example, in the work of the 15th-century Maithili (Eastern Hindi) lyric poet Vidyāpati. Even the somewhat abstruse rhetorical speculations of the Sanskritic poetic schools of analysis were used as formulas for the production of 17th-century Hindi court poetry; the

Rasikapriyā ("Beloved of the Connoisseur") of Keśavadāsa is a good example of this kind of tour de force.

There are other characteristics common to the regional literatures, some of which come not from Sanskrit but most likely from the Apabhraṃśa. There are two poetic forms, for example, that are found in many northern Indian languages: the *bārah-māsā* ("twelve months"), in which 12 beauties of a girl or 12 attributes of a deity might be extolled by relating them to the characteristics of each month of the year; and the *caütīs* ("thirty-four"), in which the 34 consonants of the northern Indian Devanāgarī alphabet are used as the initial letters of a poem of 34 lines or stanzas, describing 34 joys of love, 34 attributes, and so on.

Finally, there are common characteristics that may have come either through Apabhraṃśa or through the transmission of stories and texts from one language to another. The stories of Gopi-candra, the cult hero of the Nātha Yogī sect, a school of mendicant *sannyāsin*s, were known from Bengal to the Punjab even in the early period. And the story of the Rājput heroine Padmāvatī, originally a romance, was beautifully recorded, with a Ṣūfī (mystic) twist, by the 16th-century Muslim Hindi poet Malik Muḥammad Jāyasī and later by the 17th-century Bengali Muslim poet Ālāol. From the late 13th through the 17th century, bhakti (devotional) poetry took hold in one region after another in northern and eastern India. Beginning with the *Jñāneśvarī*, a Marathi verse commentary on the *Bhagavadgītā* written by Jñāneśvara (Jñānadeva) in the late 13th century, the devotional movement spread through Mahārāshtra, in the works of the poet-saints Nāmdev and Tukārām; through Rājasthān, where it is represented by the works of Mīrā Bāī; through northern India, in the poetry of Tulsīdās, Sūrdās, Kabīr, and others; through Mithilā, in the work of the great poet Vidyāpati; and into Bengal, where Caṇḍīdās and others sang of their love of God. Because of the *bhakti* movement, beautiful lyric poetry and passionate devotional song were created; and in some cases, as in Bengal, serious philosophical works and biographies were written for the first time in a regional language rather than in Sanskrit. The languages and their literatures gained strength as mediums of self-expression as well as exposition. And, although there is much Sanskrit imagery and expression in the poetry and song, as well as similarities to Sanskrit textual models, its basic character is not Sanskritic: true to the nature of any spoken, everyday language, it is more vital than polished, more vivid than refined.

Spread of *bhakti* poetry

One more historical generality can be stated regarding regional Indian literature before considering the characteristics peculiar to the several "Indian literatures." In all of the early literatures, writing was lyrical, narrative, or didactic, entirely in verse, and all in some way related to religion or love or both. In the 16th century, prose texts, such as the Assamese histories known as the *buranji* texts, began to appear.

Hindi. What is commonly spoken of as Hindu is actually a range of languages, from Maithili in the east to Rajasthani in the west. The first major work in Hindi is the 12th-century epic poem *Pṛthvīrāj Rāsau,* by Chand Bardaī of Lahore, which recounts the feats of Pṛthvīrāj, the last Hindu king of Delhi before the Islāmic invasions. The work evolved from the bardic tradition maintained at the courts of the Rājputs. Noteworthy also is the poetry of the Persian poet Amīr Khosrow, who wrote in the Awadhi dialect. Most of the literature in Hindi is religious in inspiration; in the late 15th and early 16th centuries, the reform-minded Kabīr, for example, wrote sturdy short poems in which he sought to reconcile Islām and Hinduism.

The most celebrated author in Hindi is Tulsīdās of Rājāpur (died 1623), a Brahmin who renounced the world early in life and spent his days in Benares (Vārānasi) as a religious devotee. He wrote much, mostly in Awadhi, and focussed Hinduism on the worship of Rāma. His most important work is the *Rāmcaritmānas* ("Sacred Lake of the Acts of Rāma"), which is based on the Sanskrit *Rāmāyaṇa.* More than any other work it has become a Hindu sacred text for the Hindi-speaking area and annually has been staged in the popular Rām Līlā festival.

Outstanding among the followers of Vallabha, philosopher and *bhakti* ("devotion") advocate of the Middle Ages, is the blind poet Sūrdās (died 1563), who composed countless *bhajan*s (chants) in praise of Krishna and Rādhā, which are collected in the *Sūrsāgar* ("Ocean of Sūrdās"). While many of the *bhakti* poets were of modest origin, an exception was Mīrā Bāī, a princess of Jodhpur, who wrote her famous lyrics both in Hindi and Gujarati; the quality of her poetry, still very popular, is not as high, however, as that of Sūrdās. Significant also is the religious epic *Padmāvatī* by Jāyasī, a Muslim from former Oudh state. Written in Awadhi (*c.* 1540), the epic is composed according to the conventions of Sanskrit poetics.

The 18th century saw the beginning of a gradual transformation from the older forms of religious lyric and epic to new literary forms influenced by Western models that began to be known. The new trends reached their pinnacle in the work of Prem Chand (died 1936), whose novels (especially *Godān*) and short stories depict common rural life; and in the work of Harishchandra of Benares (died 1885), honoured as Bhāratendu (Moon of India), who wrote in the Braj Bhasa dialect.

Bengali. While developments in Bengali literature began somewhat earlier, they followed the same general course as those in Hindi. The oldest documents are Buddhist didactic texts, called *caryā-pada*s ("lines on proper practice"), which have been dated to the 10th and 11th centuries and are the oldest testimony to literature in any Indo-Aryan language.

Bengali poetry, including poetry by Bengalis in other dialects, is largely written in three distinct genres. It is certain that well before the 15th century there existed texts in a typically Bengali genre called *maṅgal-kāvya* ("poetry of an auspicious happening"), which consists of eulogies of gods and goddesses; such poetry is likely to have had a considerable history in oral transmission before it was committed to writing. A good example of an orally transmitted *maṅgal* poem is the *Caṇḍī-māngal* ("Poem of the Goddess Caṇḍī"), by Mukundarāma, which was put into written form in the latter part of the 15th century. *Maṅgal* poetry remained a favourite genre well into the 18th century, when Bhārat-candra wrote the *Annadā-maṅgal* ("Maṅgal of the Goddess Annadā [the Giver of Food]"), a witty and sophisticated poem that bears little resemblance to its more rustic forebears. Despite this popularity, it is the devotional lyrics to the divine pair Krishna and Rādhā that are still known and sung today in Bengal, and these lyrics are the gems of medieval Bengali literature.

The three genres of Bengali poetry

Poems of the second genre, the *mahākāvya* ("great poem," but not to be confused with the Sanskrit *mahākāvya* genre), are based mainly on the Sanskrit models of the *Mahābhārata, Rāmāyaṇa,* and *Purāṇas.* Kṛttibās Ojhā (late 14th century) stands at the beginning of this literature; he wrote a version of the *Rāmāyaṇa* that often differs from the Sanskrit original, for he includes many local legends and places the setting in Bengal. Kavīndrā (died 1525) wrote on the *Mahābhārata* theme, as did Kāsiram Dās in the 17th century.

The third genre, *padāvalī* ("string of verse") songs, is also found elsewhere; inspired by the religious *bhakti* movement, the songs resemble the devotional poetry of the Nāyaṉārs and Āḻvārs in Tamil. It was such poetry that established Bengali as a significant literary language. The earliest work in what may be considered a distinctively Bengali style is the *Śrīkṛṣṇa-kīrtana* ("Praise of the Lord Krishna"), a long *padāvalī* poem by Caṇḍīdās, which is dated to the early 15th century. In it the poet praises the virtues and celebrates the loves of Krishna, a theme that had remained popular in Bengal ever since its first glorification by the Bengali Sanskrit poet Jayadeva, who composed his *Gītagovinda* ("The Cowherd Song") in the 12th century. *Padāvalī* songs describe and glorify all phases of Krishna's love for the cowherds' wives (especially Rādhā, who later became a goddess), and it is love poetry before it is religious poetry. After the great Bengali mystic and saint Caitanya (died 1533), love *is* religion, and the erotic is inspirited with religious fervour. The great flowering of this poetry occurred in the 16th and 17th centuries.

Religious edification took the forms not only of *maṅgal*s and *padāvalī*s but also of biography (more like hagiography) and philosophy. Important in that style is the long hagiography *Caitanya-caritāmṛta* ("Elixir of the Life of Caitanya"), by the 16th-century author Kṛṣṇadās.

While most of the literature is Hindu in theme and inspiration, there arose a secular Bengali literature among Bengali Muslims. One of the outstanding Muslim poets is Ālāol, author of the *Padmāvatī* (*c.* 1648), which was written after the poem of the same name by the Hindi poet Jāyasī.

Assamese. The earliest text in a language that is incontestably Assamese is the *Prahlāda-caritra* of Mena Sarasvati (13th century); in a heavily Sanskritized style it tells the story, from the *Viṣṇu-Purāṇa,* of how the mythical king Prahlāda's faith and devotion to Vishnu saved him from destruction and restored the moral order. The first great Assamese poet was Kavirāja Mādhava Kandalī (14th century), who translated the Sanskrit *Rāmāyaṇa* and wrote *Devajit,* a narrative on the god Krishna. In Assamese, too, the *bhakti* movement brought with it a great literary upsurge; the most famous Assamese poet of the period was Śaṅkaradeva (died 1568), whose 27 works of poetry and devotion are alive today and who inspired such poets as Mādhava-devi to write lyrics of great beauty. Peculiar to Assamese literature are the *buranji*s, chronicles written in a prose tradition brought to Assam by the Ahoms of Burma. These date in Assamese from the 16th century, while in the Ahom language they are much earlier.

Oriya. *Mādaḷā-pāñji* ("The Drum Chronicle") texts in Oriya, the chronicles of the great temple of Jagannātha in Puri, date from the 12th century. They are in prose, and as such they represent the earliest prose in a regional Indo-Aryan language, although they cannot be said to be literary texts. The 14th century was productive for Oriya literature. Dating from this period are the anonymous *Kalasa-cautīśa,* which tells in 34 verses the story of the marriage of the god Śiva and the mountain goddess Pārvatī, and the famous *Caṇḍī-purāṇa* of Saraladāsa. But the *bhakti* period was once again the most stimulating one; the best known medieval Oriya poet is Jagannātha Dās (whose name means Servant of Jagannātha), a 16th-century disciple of the Bengali Vaiṣṇava saint Caitanya, who spent the better part of his life in Puri. Among the many works of Jagannātha Dās is a version of the Sanskrit *Bhāgavata-Purāṇa* that is still popular in Orissa today.

Marathi. With Bengali, Marathi is the oldest of the regional literatures in Indo-Aryan, dating from about AD 1000. In the 13th century, two Brahminical sects arose, the Mahānubhāva and the Varakari Panth, both of which put forth vast quantities of literature. The latter sect was perhaps the more productive, for it became associated with *bhakti,* when that movement stirred Mahārāshtra in the early 14th century, and particularly with the popular cult of Viṭṭhoba at Pandharpur. It was out of this tradition that the great names of early Marathi literature came: Jñāneśvara, in the 13th century; Nāmdev, his younger contemporary, some of whose devotional songs are included in the holy book of the Sikhs, the *Ādi Granth;* and the 16th-century writer Eknāth, whose most famous work is a Marathi version of the 11th book of the *Bhāgavata-Purāṇa.* Among the *bhakti* poets of Mahārāshtra the most famous is Tukārām, who wrote in the 16th century. A unique contribution of Marathi is the tradition of *povāḍā*s, heroic stories popular among a martial people. There is no way of dating the earliest of these; but the literary tradition is particularly vital at the time of Śivajī, the great military leader of Mahārāshtra (born 1630), who led his armies against the might of the Mughal emperor Aurangzeb.

adition
*povāḍā*s

Gujarati. The oldest examples of Gujarati date from the writings of the 12th-century Jaina scholar and saint Hemacandra. The language had fully developed by the late 12th century. There are works extant from the middle of the 14th century, didactic texts written in prose by Jaina monks; one such is the *Bālāvabodha* ("Instructions to the Young"), by Taruṇa-prabha. A non-Jaina text from the same period is the *Vasanta-vilāsa* ("The Joys of Spring"). The two Gujarati *bhakti* poets, both of the 15th century, are Narasiṃha Mahatā (or Mehtā) and Bhālaṇa (or Puruṣottama Mahārāja); the latter cast the 10th book of the *Bhāgavata-Purāṇa* into short lyrics. By far the most famous of the *bhakti* poets is the woman saint Mīrā Bāī, who lived in the first half of the 16th century. Mīrā, though married, thought of Krishna as her true husband, and the lyrics telling of her relationship with her god and lover are among the warmest and most movingly personal in any Indian literature. One of the best known of the non-*bhakti* Gujarati poets is Premānanda Bhaṭṭa (16th century), who wrote narrative poems based on *Purāṇa*-like tales; although his themes were conventional, his characters were real and vital, and he infused new life into the literature of his language.

Punjabi. Punjabi developed a literature later than most of the other regional languages of the subcontinent; and some of the early writings, such as those of the first Sikh Gurū, Nānak (late 15th and early 16th centuries), are in Old Hindi rather than true Punjabi. The first work identifiable as Punjabi is the *Janam-sākhī,* a 16th-century biography of Gurū Nānak by Bala. In 1604, Arjun, the fifth Gurū of the Sikhs, collected the poems of Nānak and others into what is certainly the most famous book to originate in the Punjab (though its language is not entirely Punjabi), the *Adi Granth* ("First Book"). Writing that is not merely incidentally Punjabi began in the 17th century and is almost entirely by Muslims. Between 1616 and 1666, a writer named ʿAbdullāh, for example, composed a major work called *Bāra Anva* ("Twelve Topics"), which is a treatise on Islām in 9,000 couplets. Muslim Ṣūfīs, such as Bullhē Shāh (died 1758), also contributed many devotional lyrics, and Ṣūfī Islām can be said to have been the main stimulus to Punjabi literature in the medieval period. There are also many romances in the language (as in Rajasthani) which, being oral literature, are undatable.

Kashmiri. The hitherto commonly accepted period of Old Kashmiri is 1200–1500; but in fact the earliest example of the language is found in 94 four-line stanzas embedded in the Sanskrit philosophical work *Mahānaya-prakāśa* ("Illumination of the Highest Attainment"), which some scholars now date as late as the 15th century. As is true for Gujarati, the most famous poets of Kashmiri in this period are women. Lallā (14th century) wrote poems about the god Śiva; and Hubb Khātun (16th century) and especially Arani-mal (18th century) are famous for their hauntingly beautiful love lyrics. Despite these outstanding poets in Kashmiri, the great literary language of Kashmir in the medieval period was Persian, which was encouraged by many rulers of the country, such as Zayn-ul-ʿĀbidīn, in whose 15th-century court were many scholars and poets writing in both the Kashmiri and Persian languages.

(E.C.D.)

ISLĀMIC LITERATURES: 11TH–19TH CENTURY

The adventure of Islām in India began in the 8th century with the conquest of Sind (the extreme western province), but it was only in the 11th and 12th centuries that Muslim literary and cultural traditions reached the Indian heartland. Then, in the 13th century, refugee noblemen, soldiers, and men of letters from Iran and Central Asia came pouring into India. Although the causes changed, the attraction of India as a place of refuge and gracious patronage did not decline for several subsequent centuries. At the same time Muslim soldier-adventurers continued with their conquests, joining hands with their non-Muslim Indian counterparts in many instances, establishing minor or major kingdoms all over the subcontinent. The political map of India remained very much in flux—except for a brief period during the reign of Akbar—until Queen Victoria declared herself empress of India in 1858. The period of Muslim influence thus extends over 800 years.

At the time of the spread of Muslim power and culture in India, Sanskrit was the chief language of Hindu cultural, learned, and religious expression, while Buddhism and Jainism had lent their prestige and patronage to various Prākrits. The progress of and developments in these literatures remained unaffected by the advent of Islām in India. The emergence of the new Indo-Aryan languages out of the Prākrit and Apabhraṃśa stages of Sanskrit, however, was furthered by the newcomers, who preferred these regional languages over Sanskrit and encouraged the development

Effect of
Islām
on
Sanskrit
and
Prākrit
literatures

of popular regional literatures. The conversion to Islām of a large number of indigenous people enhanced these developments. Thus, the vehicles of literary expression used by those professing Islām in India were regional dialects and languages, both Indo-Aryan and Indo-Iranian, such as Braj, Awadhi, Sindhi, Baluchi, Urdu, Dakhini, and Bengali, as well as the foreign Arabic, Turkish, and Persian spoken by the Muslim immigrants and conquerors. Of these, only Persian and Urdu require detailed consideration; the others will be discussed only briefly.

Arabic. Arabic was the language of the conquerors of Sind. But it enjoyed more permanent prestige as the language of the Qurʾān, the sacred book of Islām; as such it was extensively used for religious scholarship during the medieval period. Even as late as the 18th century, Shāh Walī Allāh, the greatest theologian to have lived in India, wrote his most important treatises in Arabic. Arabic was also used early for historiography and for making Indian scientific books available to the Middle East in translation. One does not find, however, much in the way of significant Arabic belles lettres in India.

Turkish. Although the earliest Muslim conquerors in northern India were Turks, their language was Persian. It was only during the reigns of Bābur and his son Humāyūn (1526–56) that Turkish flourished for a while as a medium of learned expression. Bābur himself was the foremost contributor. Although his memoirs are better known, he also left a volume of verses of considerable merit.

Regional languages. The literatures of the Indo-Iranian languages of Baluchi and Pashto are exclusively creations of Muslim writers. In the Indo-Aryan languages of Kashmiri, Sindhi, and Punjabi, Muslims were the most influential contributors; the names of Lallā (14th century) for Kashmiri, Shāh ʿAbd-ul-Laṭīf (17th–18th century) for Sindhi, and Wāris̄ Shāh (18th century) for Punjabi exemplify that fact. Muslim chieftains gave impetus to the growth of Bengali literature through their patronage of writers and through their efforts to have Sanskrit classics translated into Bengali. There are also many famous Muslim names during the medieval period of Bengali literature, such as Dawlat Qāz̄ī and Ālāol in the 17th century. In the heartland of northern India, notable contributions were made by Muslims to Hindi literatures in the Braj and Awadhi dialects. Malik Muḥammad Jāyasī, Raḥīm, and Manjhan (all 16th century) and ʿUs̄man (17th century) are some of the important names. In the 16th, 17th, and 18th centuries in India there was a tremendous production of mystic (Ṣūfī and *bhakti*) poetry in all of the important dialects and languages. It was a period of great mystic, syncretic movements, and the Muslim contribution in the form of love narratives and lyrics was considerable. Quite often metres, motifs, and assorted rhetorical features of Persian *mas̄navīs* and *ghazals* (see below *Urdu*) were used in a new medium. Moreover, interaction and assimilation took place between the Muslim Ṣūfī traditions, thought, and practices and the Indian *bhakti* schools. Muslim *bhakti* poets either expressed Ṣūfī ideas, which were close to monotheistic orthodoxy as well as to the doctrines of Indian saints Kabīr and Nānak, in the Indian dialects through narrative poems modelled on Persian *mas̄navīs* or chose the path of ecstasy and became devotees of Krishna (which was still close to the more orthodox forms of Ṣūfism). None of them followed the devotional style of Tulsīdās, their contemporary and a devotee of Rāma.

It was, however, in Persian and Urdu that Muslim men of letters made the greatest contributions—contributions that led in the former case to the establishment of an "Indian" school of Persian poetry and influenced profoundly the development of poetry in Afghanistan and Tadzhikistan and, in the latter case, led to the emergence of a unique pan-Indian language and literature in Urdu.

Persian. Maḥmūd of Ghazna, with whom the chain of Muslim conquests in northern India began, was also the patron of Ferdowsī, one of the greatest of Persian poets. The later conquerors admired literature no less. Since the language of all of them was Persian, the growth of Persian literature in India kept pace with its conquest by the Muslims.

Masʿūd Saʿd Salmān (born 1046 in Lahore), who later became the governor of Jullundhur, was the first noteworthy person of Indian origin to have written poetry in Persian. The first truly great poet was Amīr Khosrow, who wrote in the 13th and 14th centuries. Of Turkish descent, born in the district of Etah in northern India, Khosrow was connected with royal courts all his life, even after 1272, when he became a disciple of the great mystic Niẓām-ud-Dīn Awliyā. He wrote five books of poems, or *dīvāns*, composed of *ghazals* (see below *Urdu*), panegyrics and several *mas̄navīs*—altogether some 200,000 couplets. In poetry, his innovative spirit displayed itself in *waṣf-nigārī*—that is, descriptions of natural objects in short poems, which Khosrow incorporated within longer ones. His keenness of observation is also evident in his use of local fauna and flora as poetic images. Khosrow's distinction lies not so much in the fact that he is an innovator, however, as in the fact that he is equally superb in narrative poetry, panegyrics, and lyrics. The range of his popularity and influence can best be gauged by the fact that, in northern Indian folk literature, one comes across numerous songs and riddles consistently attributed to Amīr Khosrow.

The poems of Amīr Khosrow

In the centuries that followed Khosrow, until the end of the Islāmic period, India contributed to Persian literature in two ways: first, through the production of dictionaries that helped to standardize the language and consolidate its vocabulary; second, through the development of the so-called Indian style of Persian poetry.

It is generally agreed that this Indian style, *sabk-e hindī,* did not originate within the geographic confines of India, though it reached its most sublime form there at the hands of poets who either were born in India or spent their most productive years at various Indian courts. Some of the characteristics of the style are (in the words of one modern critic) the emphasis on

> parallel statement . . . ; on complex conceit like that of the seventeenth century English "metaphysical" poets, arising out of economy of expression and telescoping into a single image a variety of emotional states; on "cerebral" artifice in pushing familiar images to unfamiliar and unexpected lengths; and on the creation of a synthetic poetic diction in which a whole phrase constitutes a single image.

The keen observation of daily life that is also characteristic of Indian Persian poetry could have been inspired by the traditions of classical Sanskrit poetry, with which these poets must have been familiar through the extensive translations done during the reign of the Mughals.

The century (1556–1657) of the reigns of Akbar, Jahāngīr, and Shāh Jahān was the most glorious period for Persian poetry in India, though, except for Fayẕī, all of the important poets were immigrants from Persia who found relief from religious and political persecution as well as generous patronage at the hands of the great Mughals and the lesser kings of southern India. The great men of letters of that period were ʿUrfī, Ṭālib Āmulī, Naẓīrī, Ẓuhūrī, Kalīm, and Ṣāʾib.

The greatest poet of the Indian style, however, was ʿAbdul Qādir Bēdil, born in 1644 in Patna, of Uzbek descent. He came early under the influence of the Ṣūfīs, refused to be attached to any court, and travelled widely throughout India during his long life. Bēdil's 16 books of poetry contain nearly 147,000 verses and include several *mas̄navīs*. Though ignored by the Iranians, Bēdil's poetry had an impact on Tadzhik and Uzbek literatures, and its influence is still evident in Afghanistan. A poet of great virtuosity and philosophic bent, he was well acquainted with Indian religions and philosophy. His anti-feudal views and his critical and skeptical attitude toward all kinds of dogma make his poetry relevant even today. His style is difficult, his metaphors and syntax quite complex (though the language itself is quite simple); and yet, as a modern critic puts it, "the intensity of his subjective assessment is so acute and factual, and his metaphysical experience so intense, that genuine poetry emerges in all its splendour."

The works of ʿAbdul Qādir Bēdil

Urdu. Earlier varieties of Urdu, variously known as Gujari, Hindawi, and Dakhani, show more affinity with eastern Punjabi and Haryani than with Khari Boli, which provides the grammatical structure of standard modern Urdu. The reasons for putting together the literary products of these dialects, forming a continuous tradition with those in Urdu, are as follows: first, they share a common milieu, consisting of Ṣūfī and Muslim court culture, increasingly dominated by the life and values of the urban elite; second, they display wholesale acceptance of Perso-Arabic literary traditions, including genres, metres, and rhetoric; third, they show an increasing acceptance of Perso-Arabic grammatical devices and vocabulary; and fourth, they tend to prefer Perso-Arabic forms over indigenous forms for learned usage.

Apart from themes and metaphysics, the influence of Ṣūfī hospices and royal courts can be seen in two practices that were essential to the development of Urdu poetry (and also unique to the Urdu milieu in the medieval period) and that still exist in modified forms. First, Urdu poets generally chose an *ustād,* or master, just as a Ṣūfī novice chose a *murshid,* or preceptor, and one's poetic genealogy was always a matter of much pride. Second, poets read poetry in private or semiprivate gatherings, called *mushāʿirah,* which displayed hierarchies, status consciousness, and rivalries reminiscent of royal courts.

Urdu literature began to develop in the 16th century, in and around the courts of the Quṭb Shāhī and ʿĀdil Shāhī, kings of Golconda and Bījāpur in the Deccan (central India). In the later part of the 17th century, Aurangābād became the centre of Urdu literary activities. There was much movement of the literati and the elite between Delhi and Aurangābād, and it needed only the genius of Walī Aurangābādī, in the early 18th century, to bridge the linguistic gap between Delhi and the Deccan and to persuade the poets of Delhi to take writing in Urdu seriously. In the 18th century, with the migration of poets from Delhi, Lucknow became another important centre of Urdu poetry, though Delhi never lost its prominence.

The first three centuries are dominated by poetry. Urdu prose truly began only in the 19th century, with translations of Persian *dāstāns,* books prepared at the Delhi College and the Fort William College at Calcutta, and later with the writers of the Aligarh movement.

'he 'orks of ḷazīr .kbarā-ādī

To focus on essential matters, the discussion that follows forgoes a chronological account of the poetry, concentrating instead on characteristics of particular genres and the achievements of the most significant of their practitioners up to 1857. There is one poet, however, who cannot be described as a practitioner of the classical Perso-Arabic traditions adopted by his fellow poets. Naẓīr Akbarābādī, who wrote in the late 18th and early 19th centuries, was a poet of consummate skill who chose to display it in short poems (in various forms) written in the language of popular speech as well as of literature. His themes show similar eclecticism. In his voluminous body of work, there are poems on such diverse topics as popular festivals, the seasons, the vanities of life, erotic pleasures and pursuits, dancing bears, and niggardly merchants. He is a master of the telling detail that immediately brings any event to life. Generally ignored by elitist poets and literary chroniclers of his time, Naẓīr has gained increasing respect and recognition as the first and best poet of the people.

Qasīdahs. *Qasīdah*s are poems written with a "purpose"—the purpose being worldly gain, in the case of poems praising kings and noblemen, or benefit in the afterworld, in the case of poems praising God, the prophet Muḥammad, and other holy personages. These panegyrics are generally overly long and are written in a highly ornate and hyperbolic style, the poets vying to display their prowess by using as many rhymes and discovering as many associative themes as possible. Because of their style and language they are of special interest to lexicographers. Not much scholarly work has been done on the *qasīdah*s written in the Deccan, but in northern India a number of poets are regarded highly for their achievements in this genre: in the 18th century, Sawdā and Inshāʾ, and in the 19th, Ẕawq and Ghālib.

Ḥaju and shahr-āshūb. Less ornate, if not less elaborate, and more edifying are the *ḥaju* (derogatory verses, personal and otherwise) and the *shahr-āshūb* (poems lamenting the decline or destruction of a city). They provide useful information about the mores and morals of the period from the 18th to mid-19th century and truly depict the problems facing the society at large. The poems are not formally restricted to any particular metre or stanza pattern. Sawdā again is one of the more famous names.

Urdu lamentatory elegies

Marsiyeh. *Marsiyeh* means "elegy," but in Urdu literature it generally means an elegy on the travails of the family and kinsmen of Ḥusayn (grandson of Muḥammad) and their martyrdom in the field of Karbalā, Iraq. These elegies and other lamentatory verses were read at public gatherings, especially during the month of Muḥarram. Although a large number of *marsiyeh*s were written in the Deccan and at Delhi, it was in Lucknow, with the patronage of Shīʿite elite and royalty, that *marsiyeh*s gained the tenor and magnitude of epic poetry. The two great masters of that 19th-century period were Mīr Anīs and Mīrzā Dabīr, who together established *musaddas* (a six-line stanza with an *aaaa bb* rhyme scheme) as the preferred form for *marsiyeh*s and added several new topics and details to the ranks of associated themes, thus carrying the form beyond a simple lament. An interesting aspect of these elegies is that, although the scene and personae are Arab, there is no attempt at verisimilitude: Arab gallants and maidens speak and gesture like the elites of Lucknow. Perhaps this added to the pathos and effectiveness of the poems at public readings.

Masnavī. *Masnavī* was the preferred genre for all descriptive and narrative purposes, for it allows the most freedom (only the lines of each couplet must rhyme). In the Deccan, all major poets wrote at least one long *masnavī,* but lack of knowledge of the dialect has prevented their full appreciation. Thus, the more famous *masnavī*s are by later poets of Delhi and Lucknow, such as Mīr, Mīr Ḥasan, Dayā Shankar Nasīm, and Mīrzā Shawq. The topics of descriptive *masnavī*s range from mundane events of life, hunting trips of kings, and the vagaries of nature's seasons to autobiographical discourses. Narrative *masnavī*s are considerably longer, running into hundreds of couplets. In the Deccan several poets wrote abridged versions of Persian *masnavī*s, but many others wrote original compositions utilizing Indian romances as well as the better known Persian and Arabic ones. Apart from the names of the protagonists in the *masnavī*s inspired by Persian and Arabic poems, all else is always local; the landscape, cityscape, processions, customs and rituals, social values and taboos, even the physical characteristics of the people are totally Indian, though dominantly Muslim and feudal. Despite their length, these narratives gained much popularity and, at least in northern India, were often read in public places, in much the same way as storytellers told stories. The *masnavī* form was also used by some of the Hindi Ṣūfī poets.

Characteristics of the *ghazal*

Ghazal. For the most part, the history of Urdu poetry in India is the story of Urdu *ghazal,* which has been the favourite of both poets and their audiences in every period. A short lyric, with prosodic requirements of both metre and rhyme, *ghazal* demands great skill and thought from the poet, for its couplet must be a complete semantic entity and fully express a whole, well-integrated poetic experience. Favourite themes are erotic love, Ṣūfī love, and metaphysics. Naturally, Urdu poets began by closely imitating, often even plagiarizing, Persian masters, but later on they spoke in a more authentic voice. They continued, however, to employ a vocabulary of love that owed almost everything to Persian and shared very little with the traditions of lyrical poetry in other Indian languages. For example, with few exceptions, the lover is always masculine; expression of love is never made by a woman. Unique, too, is the use of masculine grammatical forms and imagery for the beloved, even when, in every other way, the poem is clearly celebrating heterosexual love. This peculiarity, as well as other traditions borrowed from Persian masters, gives a *ghazal* couplet a tremendously wide range of interpretations. It is amazing indeed what a master poet can condense into one terse couplet.

The two greatest *ghazal* writers in Urdu are Mīr Taqī Mīr, in the 18th century, and Mīrzā Asadullāh Khān Ghālib, in the 19th. They are in some ways diametrical opposites. The first prefers either very long metres or very short, employs a simple, non-Persianized language, and restricts himself to affairs of the heart. The other writes in metres of moderate length, uses a highly Persianized vocabulary, and ranges wide in ideas. Mīr speaks of passion and pathos; Ghālib betrays a skeptic's mind and leaves nothing unquestioned, not even his feelings. But both have left indelible marks on the ideas and emotions of succeeding generations. Ghālib wrote poetry in Persian as well as Urdu and also published a couple of volumes of letters in Urdu that helped usher in modern prose. In many ways he bridges the gap separating the medieval sensibility from the modern. The contemporary mind, however, is also moved by the authentic passion of Mīr, idolizing him for the sublimity of his concept of love and for his personal integrity. The poems of Ghālib and Mīr represent the best of the Urdu *ghazal;* and the Urdu *ghazal,* as an anonymous wit has remarked, is the Muslims' greatest gift to India, after the Tāj Mahal. (C.M.N.)

SINHALESE LITERATURE: 10TH CENTURY AD TO 19TH CENTURY

The island nation of Ceylon (now called Sri Lanka), formally a part of South Asia, has been little noticed by the subcontinent, apart from the fact that according to an uncertain tradition it is celebrated in the *Rāmāyaṇa* as the island called Laṅkā. Buddhist sway was introduced there early, during the reign of Aśoka Maurya (*c.* 269–232 BC); and, while on the subcontinent Buddhism prospered, declined, and finally disappeared, in Ceylon it has continued until today. Although there are obvious borrowings in Ceylon from subcontinental literature, notably Sanskrit, and there was rather precarious communication with India through the island's Hindu community of Tamils, Ceylon never became culturally continuous with the mainland. The language itself, although of Indo-Aryan stock, is strongly mixed with a substratum of Dravidian. Also, it was Ceylon's fate early to fall victim to European colonialism, first to the Portuguese, then to the Dutch, and finally to the British, before it regained nationhood in 1948.

While there are inscriptions that antedate the Christian Era, no texts appear to survive from before the 10th century AD. The first texts that emerged were aids in Sinhalese—glossaries, paraphrases, and the like—to the study of the Pāli texts of Buddhism. More interesting are Sinhalese renderings of the life and virtues of the Buddha. Important in this genre, hagiographical rather than literary, is the *Amāvatura* ("Flood of the Ambrosia"), by Guruḷugōmī, which in 18 chapters purports to narrate the life of the Buddha, with specific emphasis on one of his nine virtues—his capacity to tame recalcitrant people or forces. In a similar vein is the literature of devotion and counsel, in which Buddhist virtues are celebrated.

Historical records

Exceptional in the context of the South Asian subcontinent is the early and persistent interest in historical records. Such interest had begun in Pāli with the *Dīpavaṃsa* ("Chronicle of the Island") and had continued with the *Mahāvaṃsa* ("Great Chronicle") and *Cūlavaṃsa* ("Lesser Chronicle"), but it had a life of its own in Sinhalese. The most important, and possibly the oldest, of such chronicles is the *Thūpavaṃsaya* ("Chronicle of the Great Stupa"), by Pārakrama Paṇḍita. Subsequent chronicles, or genealogies of places, comprise the history of all of the major Buddhist monuments. Several chronicles were also inspired by the Tooth Relic, received from Kaliṅga in the 4th century by King Kīrtiśrīmēghavarṇa. Such chronicling included that of the kings who protected the relic.

All of this literature was mostly in prose, but poetry as a literary form no doubt antedated it, as evidenced by early inscriptions. Much poetry was occasioned by Pāli *Jātakas* (stories of the Buddha's previous births) and other Buddhist stories, though Hindu stories were not lacking; for example, a version of the Sanskrit *Mahābhārata* (received through a Tamil source) was cast in the style of a *Jātaka* in the *Mahāpadaraṅga-Jatakaya.*

Likewise of Hindu Indian origin was a genre that took off from the Sanskrit poet Kālidāsa's "Meghadūta" (see above *Classical Sanskrit kāvya* [*200–1200*]), in which an exiled lover sends a message to his beloved by way of a monsoon cloud, thus giving the poet the opportunity to dwell on the description of landmarks in a poetic travelogue. This genre, so-called *saṃdeśa* literature, by no means unknown on the mainland, proliferated widely on Ceylon.

Of a different style are panegyrics and war poems, the earliest of which is the *Parakumbasirita* ("History of Parakramabahu VI," king in Jayavardhanapura from 1410 to 1468). Again reminiscent of the mainland and the religious tradition are the plentiful eulogies of the Buddha. Popular, too, was didactic verse, among the most notable of which is the *Kusajātaka,* 687 stanzas of epigrams and exempla by the 17th-century poet Alagiyavanna Mohoṭṭāla.

MODERN PERIOD: 19TH AND 20TH CENTURIES

The modern period was ushered in by the arrival of the British, the influence of Western models becoming discernible in the early 19th century. Reform-minded Hindus, led by Rammohan Ray, took a positive attitude to Western literature and urged on their countrymen a Western type of education. Newly formed literary clubs spread the influence of predominantly British works, thereby opening the Indian educated elite to Western culture and literature in general. After a period of translation, authors sought to imitate Western models and eventually to be independently creative in the new styles.

Effects of Westernization

The most striking result of Westernization was the introduction of prose on a major scale. Vernacular prose, rarely looked upon previously as a medium for art, was now used as a literary vehicle, and such hitherto unknown forms as the novel, novella, and short story began to emerge. In poetry the thrall of tradition was stronger, and verse in the older forms continued to be written. With modernity, realism appeared, as well as symbolism in some quarters, and there was new psychological and social interest.

From Bengal spread a new sense of national purpose, which became the principal motivation for much English as well as vernacular literature. Three trends can be distinguished in the products of this increasing literary activity. The old traditionalism was transformed into romanticism, which looked to the past, to Indian history, for inspiration and sought to preserve what was considered valuable in the past; a tendency to mysticism went hand in hand with the romantic mood (a mood that was also widespread in 19th-century Europe). Greater social awareness in European literature was reflected in the literature of Indian progressives, in whose works a somewhat romantic Marxism prevailed. Finally, there was a humanistic trend. The teachings of Mahatma Gandhi, combining social concerns with traditional ethics, later exerted a very great influence on literature.

In the years preceding and following India's independence (1947) and control of the princely states, the fervour of writers sometimes turned to an increasingly articulate progressivism of various Marxist schools; sometimes to disappointment and bitterness; and most recently, it appears, to a mood of introspection. These developments, which occurred with a different pace in different regions, are described briefly below. A complete coverage of the most modern literature has not been attempted, but an endeavour has been made to mention persons who are considered to be representative.

Bengali. Except for the iconoclastic poet Michael Madhusudan Datta, poetic activity in the mid-19th century was giving ground to experimentation with the new prose style learned from English. During this period, Bengali literature produced a spate of novels—satiric, social, and picaresque. While Michael's work *Mēghanādavadh* (1861; a long poem on the Rāma theme in which Rāma and Lakṣmaṇa become the villains and Rāvaṇa the hero) caused a stir, the literary event of the period was the appearance on the scene of Bankim Chandra Chatterjee, whose first novel, *Durgeśanandinī* ("Daughter of the Lord of the Fort"), appeared in 1865. While not at first overtly nationalist, Bankim Chandra became more and more an apologist for the Hindu position. In *Kṛṣṇacaritra,* Christ suffers in comparison with Krishna, and in his best known work, *Ānanda-maṭh* (1892; "The Abbey of Bliss"), the motherland in the person of the goddess Durgā is extolled.

Chatterjee and Tagore

Perhaps first among novelists of the late 19th and early 20th centuries is Saratchandra Chatterjee, whose social concerns with the family and other homely issues made his work popular. But the early 20th century is certainly best known for the poet who towers head and shoulders above the rest, Rabindranath Tagore. Poet, playwright, novelist, painter, essayist, musician, social reformer, Rabindranath produced works, still not completely collected, that fill 26 substantial volumes. The winner of the Nobel Prize for Literature in 1913, primarily for his little book of songs called *Gītāñjali,* which was much praised by Ezra Pound and William Butler Yeats, Tagore is more known for these devotional poems than for the wit and clear thought with which his later work is filled. He was the last of an era, looking back as he did to the religious and political history of Bengal for his inspiration. Those who followed him were more concerned with introspection and dramatic imagery.

If Tagore was the last poet in the Bengali tradition, Jibanananda Das was the first of a new breed. Musing and melancholy, yet known for vivid and unusual imagery Jibananada is a poet who has much influence on younger writers in Bengal. There have been many other poets in the 20th century who are equally powerful but stand somewhat apart from the mainstream. One of these was Sudhindranath Datta, a poet much like Pound in careful and etymological use of language; another is the poet and prose writer Buddhadeva Bose.

Bose has been termed a progressive, and indeed he consciously turned away from the tradition orientation of Tagore and sought inspiration in schools foreign to Bengal—for example, the French Symbolists. He is the leader of an artistic faction, the Kallol school, and editor of an influential literary magazine, *Kavitā.* Unjustifiably called obscene, his writing has been experimental, probing into social and psychological realities of Bengali life.

While there have been, and still are, literary factions associated with political positions, they have been less definitive than some in other parts of India. Bengali writers in the 20th century have tended to be personal and individual rather than propagandist for political positions.

Assamese. Assamese literature began with Hemchandra Baruwa, a satirist and playwright, author of the play *Bahiri-Rang-Chang Bhitare Kowabhaturi* (1861; "All That Glitters Is Not Gold"). The most outstanding among the early modern writers was Lakshminath Bezbaruwa, who founded a literary monthly, *Jōnāki* ("Moonlight"), in 1889, and was responsible for infusing Assamese letters with 19th-century Romanticism. Later 20th-century writers have tried to remain faithful to the ideals of *Jōnāki.* The short story in particular has flourished in the language; notable practitioners are Mahichandra Bora and Holiram Deka.

The year 1940 marked a shift toward psychology, but World War II effectively put an end to literary development. When writers resumed after the war, there was a clear break with the past, in experimental verse and the growth of the novel form.

Hindi. Modern Hindi literature began with Harishchandra in poetry and drama, Mahavir Prasad Dvivedi in criticism and other prose writings, and Prem Chand in fiction. This period, the second half of the 19th century, saw mainly translations from Sanskrit, Bengali, and English. The growth of nationalism and social reform movements of the Arya Samaj led to the composition of long narrative poems, exemplified by those of Maithili Sharan Gupta; dramas, by those of Jayashankar Prasad; and historical novels, by those of Prasad, Chatureen Shastri, and Vrindavan Lal Varma. The novels drew mainly on the periods of the Maurya, Gupta, and Mughal empires.

andhi's
fluence
ı
ɔetry

This period was followed by the Non-cooperation and *satyāgraha* movements of Mahatma Gandhi, which inspired poets such as Makhan Lal Chaturvedi, Gupta, and Subhadra Chauhan and novelists such as Prem Chand and Jainendra Kumar. Eventual disillusionment with Gandhian experiments and the increasing influence of Marxism on European literature influenced writers such as Yashpal, Rangeya Raghava, and Nagarjuna.

S.N. Pant, Prasad, Nirala, and Mahadevi Varma, the most creative poets of the 1930s, drew inspiration from the Romantic tradition in English and Bengali poetry and the mystic tradition of medieval Hindi poetry. Reacting against them were the Marxist poets Ram Vilas Sharma and Nagarjuna and experimentalists such as H.S. Vatsyayan "Agyeya" and Bharat Bhuti Agarwal. Nirala, who developed from a mystic-romantic into a realist and experimentalist, was the most outstanding poet of the 1950s; and Sarweshwar Dayal Saxena, Kailash Vajpeyi, and Raghubir Sahay were the most creative poets of the 1960s.

Two trends, represented by the work of Prem Chand and Jainendra Kumar, led Hindi fiction in two different directions: while social realists like Yashpal, Upendranath Ashk, Amritlal Nagar, Mohan Rakesh, Rajendra Yadav, Kamleshwar, Nagarjuna, and Renu faithfully analyzed the changing patterns of Indian society, writers such as Ila Chandra Joshi, "Agyeya," Dharm Vir Bharati, and Shrikant Varma explored the psychology of the individual, not necessarily within the Indian context.

Among the dramatists of the 1930s and 1940s were Govind Ballabh Pant and Seth Govind Das; because of their highly Sanskritized language, their plays have had a limited audience. Plays by minor writers such as Ramesh Mehta, however, are repeatedly staged by professional theatres. In between these extremes there are some notable playwrights.

Gujarati. In Gujarāt, too, the advent of British rule deeply influenced the literary scene. The year 1886 saw the *Kusumamālā* ("Garland of Flowers"), a collection of lyrics by Narsingh Rao. Other poets include Kalapi, Kant, and especially Nanalal, who experimented in free verse and was the first poet to eulogize Gandhi. Gandhi, himself a Gujarati, admonished poets to write for the masses and thus inaugurated a period of poetic concern with changes in the social order. Many incidents in Gandhi's life inspired the songs of poets. The Gandhi period in Gujarāt as elsewhere gave way to a period of progressivism in the class-conflict poetry of R.L. Meghani and Bhogilal Gandhi. In post-independence India, poetry has tended to become subjectivist and alienated without, however, fully superseding the traditional verse of devotion to God and love of nature.

Among novelists, Govardhanram stands out; his *Sarasvatīchandra* is a classic, the first social novel. In the novel form, too, the influence of Gandhiism is clearly felt, though not in the person of Kanaiyalal Munshi, who was critical of Gandhian ideology but still, in several *Purāṇa*-inspired works, tended to preach much the same message. In the period after independence the modernists embraced existentialistic, surrealistic, and symbolistic trends and gave voice to the same kind of alienation as the poets.

Marathi. The modern period in Marathi poetry began with Kesavasut and was influenced by 19th-century British Romanticism and liberalism, European nationalism, and the greatness of the history of Mahārāshtra. Kesavasut declared a revolt against traditional Marathi poetry and started a school, lasting until 1920, that emphasized home and nature, the glorious past, and pure lyricism. After that, the period was dominated by a group of poets called the Ravikiraṇ Maṇḍal, who proclaimed that poetry was not for the erudite and sensitive but was instead a part of everyday life. Contemporary poetry, after 1945, seeks to explore man and his life in all its variety; it is subjective and personal and tries to speak colloquially.

Among modern dramatists, S.K. Kolhatkar and R.G. Gadkari are notable. Realism was first brought to the stage in the 20th century, by Mama Varerkar, who tried to interpret many social problems.

Novelists
of the
20th
century

The *Madhalī Sthiti* (1885; "Middle State"), of Hari Narayan Apte, began the novel tradition in Marathi; the work's message was one of social reform. A high place is held by V.M. Joshi, who explored the education and evolution of a woman (*Suśīlā-cha Diva,* 1930) and the relation between art and morals (*Indu Kāḷe va Saralā Bhoḷe,* 1935). Important after 1925 were N.S. Phadke, who advocated art for art's sake, and V.S. Khandehar, who countered the former with an idealistic art for life's sake. Noteworthy contemporary novelists are S.N. Pendse, V.V. Shirwadkar, G.N. Dandekar, and Ranjit Desai.

Punjabi. Modern Punjabi literature began around 1860. A number of trends in modern poetry can be discerned. To the more traditional genres of narrative poetry, mystic verse, and love poems was added nationalist poetry in a humorous or satiric mood and experimental verse. Among the more important Punjabi poets are Bhai Vir Singh, in the 19th century, and Purana Singh, Amrita Pritam, and Baba Balwanta, in the 20th century.

Modern prose is represented by Bhai Vira Singha, Charana Singha, and Nanaka Singha, all of whom wrote novels; the same writers, as well as Gurbhaksh Singh and Devendra Satyarathi, also wrote short stories. Among playwrights mention may be made of I.C. Nanda, Harcharan Singh, and Santa Singh Sekhon.

Rajasthani. It is generally agreed that modern Rajasthani literature began with the works of Suryamal Misrama. His most important works are the *Vamsa Bhaskara* and the *Vira satsaī*. The *Vamsa Bhaskara* contains accounts of the Rājput princes who ruled in what was then Rājputāna (at present the state of Rājasthān), during the lifetime of the poet (1872–1952). The *Vira satsaī* is a collection of couplets dealing with historical heroes. Two other important poets in this traditional style are Bakhtavara Ji and Kaviraja Muraridana.

The period of nationalist strife against the British inspired a number of poets to verse that was both nationalist and in the traditional heroic vein; among them are Hiralala Sastri, Manikyalala Varma, and Jayanarayana Vyasa. This period was followed by one in which progressive social ideals inspired such poets as Ganeshilala Vyasa, Murlidhara Vyasa, and Satyaprakasha Jodhi.

Primarily known for their lyrics are Kanhaya Lal Sethiya and Megharaja Mukula, among others, and known for their narrative poems are Manohara Sharma, Shrimanta Kumara, and Naraina Singha Bhati.

Modern prose is represented in the novel, short story, and play. Among the novelists are Shiva Candra Bharatiya, Shri Lal Jodhi, Vijaya Dana Detha, and Yadavendra Sharma Chandra; the short-story writers are Rani Lakshmi Kumari Chandavata, Narasingh Rajapurohita, Dinadayala Ojha, and Purushottama Lala Menariya. Vijaya Dana Detha and Rani Lakshmi Kumari Chandavata are also known for their retelling of Rajasthani folktales. Among the playwrights is Shivachandra Bharatiya.

Tamil. In the second half of the 19th century two tendencies were present in Tamil literature. One was the old traditional prose style of the *Patiṉeṇ-kīḻkkaṇakku*, or "Eighteen Ethical Works" (see above *Dravidian literature: 1st–19th century*), learned and severely scholastic; among others, V.V. Svaminatha Iyer and Arumuga Navalar wrote in this style. Another tendency, begun by Aruṇācala Kavirāyar in the 18th century, sought to bring the spoken and written languages together. This tendency developed on one side into such works as the operatic play *Nantaṉār Carittarak Kīrttaṉai* by Gopalakrishna, and on the other into ballads, often based on the lore of the Sanskrit *Purāṇas*. Despite attempts to effect a synthesis between the two languages, however, the scholastic style has continued to have a profound influence on modern Tamil literature; the normal spoken language, in fact, never became a literary medium.

First Tamil novel

The first novel in Tamil appeared in 1879, the *Piratāpamutaliyār Carittiram*, by Vetanayakam Pillai, who was inspired by English and French novels. In important respects Pillai's work is typical of all early modern Tamil fiction: his subject matter is Tamil life as he observed it, the language is scholastic, and the inspiration comes from foreign sources. Not strictly a novel, his work, which has a predominantly moral tone, is a loosely gathered string of narratives centred around an innocent hero.

Quite different is the *Kamalāmpāḷ Carittiram* ("The Fatal Rumor"), by Rajam Aiyar, whom many judge to be the most important prose writer of 19th-century Tamil literature. In this work, the author created a series of characters that appear to have become classics; the story is a romance, yet life in rural Tamil country is treated very realistically, with humour, irony, and social satire. In language Aiyar follows the classical style, which he intermixes with informal conversation, a style that has been imitated by modern authors.

The turn of the century saw the development of the *centamiḻ* style, which in many respects is a continuation of the medieval commentatorial style. The best representative is V.V. Swaminathan, who also is responsible for the rediscovery of the Tamil classical legacy, usually called "Tamil Renaissance," which tended to direct the mood of writers back to the glorious past. The pride in Tamil subsequently gave rise to a purist tradition and a second style, called *tuyattamiḻ*, or "pure Tamil." With exaggerated Tamilian self-consciousness, the language was purged of all non-Tamil loanwords, particularly Sanskrit, which removed the literary language even further from the spoken one. This style was not ineffective in verse but led easily to rhetoric.

The purist trend brought forth a reaction in *putumaṇipravāla naṭai*, "the new *maṇipravāla*" (see above *Dravidian literature: 1st–19th century*), which was Sanskritized with a vengeance and is of little literary interest.

The scholastic and formalist character of Tamil prose was predominant in the literature until the advent, in the early 20th century, of the poet and prose writer Subrahmanya Bharati. Bharati sought to synthesize the popular and the scholastic traditions of Tamil literature, and he created thereby a Tamil that was amenable to all literary expression. This synthesis, however, did not extend to the literary language itself, which in grammar continued the formal language, though for syntax, vocabulary, etc., he drew upon colloquial speech. In doing so he saved the language from the Sanskrit tradition of *Purāṇa* writing. His style is the *maṟumaḷarcci naṭai*, the "renaissance style."

In the first half of the 20th century, R. Krishnamurthy was an immensely popular writer. Under the pseudonym Kalki, he was an influential journalist who wrote voluminous historical romances.

In the 1930s there was a literary movement inspired by a journal called *Manikkoti*. Writers in this movement contributed extremely important new works, both in verse and prose, to Tamil letters. Among them was Putumaippittan, who wrote realistically, critically, and even bitterly about the failings of society.

Contemporary literature is represented by T. Janakiraman, who writes novels, short stories, and plays with themes from urban Tamil middle-class family life; Jayakanthan, a sharp and passionate writer, with a tendency to shock his readers; and L.S. Ramatirthan, probably the finest stylist at work in Tamil today, who started by writing in English.

Malayalam. In Malayalam the modern movement began in the late 19th century with Asan, who was temperamentally a pessimist—a disposition reinforced by his metaphysics—yet all his life was active in promoting his downtrodden Ezhava community. Ullor wrote in the classical tradition, on the basis of which he appealed for universal love, while Vallathol (died 1958) responded to the human significance of social progress.

Contemporary poetry records the encounter with problems of social, political, and economic life. The tendency is toward political radicalism.

Drama, native in Malayalam tradition, emerged in the modern period as farce, comedy, and satire but turned in the 1920s to a more sombre appraisal of outdated social conventions. The novel dates back to the late 1880s and was early concerned with social realism. At present the general tendency is introspective.

Emergence of modern Kannada poetry

Kannada. Modern Kannada poetry emerged about the beginning of the 20th century and showed a spirit of national purpose that pervaded other literature as well. By 1920, after major translations from Western models had been published, new literary forms such as the lyric and the short story came to the fore in the works of Panje Mangesh Rao and B.M. Srikantiah. Other prominent Kannada writers were D.V.G. Masti, Govinda Pai, and K.V. Puttappa ("Kuvempu"). In recent years a modernist movement has influenced the literature.

Urdu. The modern period in Urdu literature coincides with the mid-19th-century emergence of a middle class that saw in Western thought and science a means to needed social reform. Naẓīr Aḥmad wrote novels about the conflicts of Muslim middle class people. Shiblī, a poet and critic, wrote on the lives of great Muslims. The more famous novelists of the later period are Ratan-Nāth Sharshār, ʿAbd-ul-Ḥalīm Sharar, and Mīrzā Ruswā. The fathers of modern Urdu poetry were Ḥālī and Muḥammad

Ḥusayn Āzād, the latter particularly characterized by a fine sensitivity for the past.

The greatest modern poet is Iqbāl. Writing in the early 20th century, he was influenced by the general sense of national purpose and the freedom movement. His poetic imagery, the power of his expression, and his philosophical outlook won the admiration of his fellow Muslims. In prose the most important writer of short stories was Prem Chand, who late in his career took to writing in Hindi. The 1930s saw the influence of progressivism, which attempted to make literature an arm of social revolution. Among the representative writers of this period are Sajjad Zahir, Upendranath Ashk, and Ismat Chughtai, the last a woman who is considered among the best.

English. There has been Indian literary activity in English for the last 200 years. It began with the insistence of the reformist Rammohan Ray and other like-minded Hindus that, for India to take its rightful place among nations, a knowledge of and education in English were essential. English literary activity took on a new aspect with the independence movement, whose leaders and followers found in English the one language that united them.

Among the first poets were Henry Derozio, Kashiprasad Ghose, and Michael Madhusudan Datta, all of whom wrote narrative verse. In the following generation there was Toru Dutt, important among women poets in this genre. Carrying on her work was Sarojini Naidu, judged by many the greatest of women poets; among her writings are *The Golden Threshold* (1905), *The Bird of Time* (1912), and *The Broken Wing* (1917). Best known of the Indian poets in English was the Bengali Rabindranath Tagore (see above *Bengali*), who, however, wrote most of his verse first in Bengali, and then translated it. A very different figure from Tagore is Sri Aurobindo, who started out as an ardent nationalist and was jailed by the British. After his conversion from activism to introspection, which took place in jail, he established a hermitage in Pondicherry. He left behind a rich *oeuvre* of verse that has inspired a contemporary school of mystic poets. Other modern poets show the influence of T.S. Eliot and Ezra Pound.

The independence movement gave strong impetus to expository prose. Important contributors to this genre were Bal Gangadhar Tilak, who edited the English journal *Mahratta,* Lala Lajpat Rai, Kasturi Ranga Iyengar, and T. Prakasam. Mahatma Gandhi, too, wrote widely in English and edited *Young India* and the *Harijan.* He also wrote the autobiography *My Experiments with Truth* (originally published in Gujarati, 1927–29), now an Indian classic. In this he was followed by Jawaharlal Nehru, whose *Discovery of India* is justly popular.

ose
tion

Prose fiction in English began in 1902 with the novel *The Lake of Palms,* by Romesh Chunder Dutt. The next important novelist is Mulk Raj Anand, who fulminated against class and caste distinction in a series of novels, *The Coolie* (1936), *Untouchable* (1935), *Two Leaves and a Bud* (1937), and *The Big Heart* (1945). Less fierce, though a better craftsman, is R.K. Narayan, who has published nine novels (as well as many short stories), among them *The Guide* (1958), *The Man-Eater of Malgudi* (1961), and *The Vendor of Sweets* (1967); his work has a wider circle of readers outside India than within. Other Indian novelists in the English medium are Santha Rama Rau, Manohar Malgonkar, Kamala Markandeya, and Khushwant Singh. The most popular is Raja Rao, whose novels *Kanthapura* (1938), *The Cow of the Barricades* (1947), and *The Serpent and the Rope* have attracted a wide following.

Sinhalese. Traditional contemporary poetry continues to be Buddhist in subject matter and sentiment. A more modern literature arose under the influence of Western models; notable among the contemporary representatives of Sinhalese literature are Kumaranatunga, a critic, Matin Wickremasinghe, a novelist, and Tennakoon, a poet.

(J.A.B.v.B.)

Music

FOLK, CLASSICAL, AND POPULAR MUSIC

Rural areas. The wide field of musical phenomena in South Asia ranges from the relatively simple two- or three-tone melodies of some of the hill tribes in central India to the highly refined art music heard in concert halls in the large cities. This variety reflects the heterogeneous population of the subcontinent in terms of race, religion, language, and social status. In the villages, music is not just a form of entertainment but is an essential element in many of the activities of daily life and plays a prominent part in most of the rituals. These include life-cycle events, such as birth, initiation, marriage, and death; events of the agricultural cycle, such as planting, transplanting, harvesting, and threshing; and a variety of work songs. Much of this music could be described as functional, for it serves a utilitarian purpose; for instance, a harvest song might well give thanks to God for a bountiful harvest, but underlying this is the idea that singing this song in its traditional manner will help to ensure that the next harvest will be equally fruitful. These songs are usually sung by all of the members participating in the activity and are not sung for a human audience. They are often sung in the form of leader and chorus, and the musical accompaniment, if any, is generally provided by drone instruments (those sustaining or reiterating a given note or notes), usually of the lute family, or percussion instruments, such as drums, clappers, and pairs of cymbals. Occasionally, a fiddle or flute might also accompany the singers, who often dance while they sing.

In each area and even within a single area, different social groups have their own individual songs whose origins are lost in antiquity. The songs are passed on from one generation to another, and in most cases the composers are unknown. Apart from folk songs, one also hears outdoor instrumental music in villages. The music is provided by an ensemble of varying size, which consists basically of an oboe type of instrument (usually a *sheh'nai* in North India and *nagaswaram* in the south) and a variety of drums. Sometimes straight, curved, or S-shaped horns may be added. These groups play at weddings, funerals, and religious processions. The musicians are professional or semiprofessional and usually belong to a very low caste. Such ensembles are found in tribal as well as folk societies and in villages as well as in cities.

Itinerant
musicians

Other professional music is also found in the rural regions. Most areas are visited by religious mendicants, many of whom travel around the countryside singing devotional songs, accompanying themselves either with a one-, two-, or three-stringed lute that generally provides only a drone or with a frame (tamborine-like) drum. They carry with them a small begging bowl and maintain themselves entirely on what they receive in alms. There are also itinerant magicians, snake charmers, acrobats, and storytellers who travel in the rural areas, often providing the only entertainment available in the villages. Music is often involved in their acts, and the storyteller generally sings his tales, which may be taken from the two epics, the *Mahābhārata* and the *Rāmāyaṇa,* or from the *Purāṇas,* the legends that describe the adventures of the incarnations of God as they rid the world of evil. Sometimes the narrative songs are concerned with historical characters and describe the wars and the heroic deeds of the regional rulers. Some storytellers specialize in generally tragic stories of romance and of lovers.

During certain religious festivals, the villages might be visited by a travelling band of players who enact some of the mythological episodes connected with the festival. Such performances are accompanied by music and may also include dances. During the festivals villagers may visit neighbourhood shrines or temples, there encountering religious mendicants singing devotional songs and perhaps watching elaborate enactments of the episodes connected with the festival. Thus, the villagers become familiar with the mythological and philosophical aspects of their religion, in spite of the low level of literacy in the rural areas and the difficulties of communication, often limited to a narrow dirt road traversed by bullock carts.

In modern times, rural areas are being influenced to a greater extent by urban culture. The principal impact has come through the introduction of relatively inexpensive transistorized radios, which have found their way into fairly remote villages. In addition, travelling cinemas, set up quickly and easily in tents, have visited the rural areas for some years. As a result, the traditional rural forms of music and dance are in the process of change.

Classical music. In the cities many different forms of music can be heard. Of these the best known in the West are the classical music of North India, including Pakistan, sometimes called Hindustani music, and that of South India, or Karnatic music. Both classical systems are supported by an extensive body of literature and elaborate musical theory. Until modern times, classical music was patronized by the princely courts and to some extent also by the wealthy noblemen. Since India gained independence in 1947, and with the abolition of the princely kingdoms, the emphasis has shifted to the milieu of large concert halls. The concertgoer, radio, and the cinema are now the main patrons of the classical musicians. In recent times the growth of university music programs, particularly involving classical music, has placed greater emphasis on music history and theory and has provided a further source of income for musicologists and musicians. The traditional system of private instruction, however, still continues to this day.

Raga and *tāla*

Classical music is based on two main elements, raga and *tāla*. The word raga is derived from a Sanskrit root meaning "to colour," the underlying idea being that certain melodic shapes, involving specific intervals of the scale, produce a continuity of emotional experience and "colour" the mind. Since neither the melodic shapes nor their sequence are fixed precisely, a raga serves as a basis for composition and improvisation. Indian music has neither modulation (change of key) nor changing harmonies; instead, the music is invariably accompanied by a drone that establishes the tonic, or ground note, of the raga and usually its fifth (*i.e.,* five notes above). These are chosen to suit the convenience of the main performer, as there is no concept of fixed pitch. While a raga is primarily a musical concept, specific ragas have acquired, particularly in North Indian music, a number of extramusical elements and are associated with particular periods of the day, seasons of the year, colours, deities, and specific moods.

The second element of Indian music, *tāla,* is best described as time measure and has two main constituents; the duration of the time measure in terms of time units that vary according to the tempo chosen; and the distribution of stress within the time measure. *Tāla,* like raga, serves as a basis for composition and improvisation.

Indian classical music is generally performed by small ensembles of not more than five or six musicians. Improvisation plays a major part in a performance, and great emphasis is placed on the creativity and sensitivity of the soloist. A performance of a raga usually goes through well-defined stages, beginning with an improvised melodic prelude that is followed by a composed piece set in a particular time measure. The composition is generally quite short and serves as a frame of reference to which the soloist returns at the conclusion of his improvisation. There is no set duration for the performance of a raga. A characteristic feature of North Indian classical music is the gradual acceleration of tempo, which leads to a final climax.

Film music

Nonclassical music of the cities. Classical music interests only a small proportion of the peoples of South Asia, even in the cities. Since about the 1930s a new genre, associated with the cinema, has achieved extraordinary popularity. Most Indian films are very much like Western musicals and generally include six or more songs. Film music derives its inspiration from a number of sources, both Indian and Western; classical, folk, and devotional music are the main Indian sources, while Western influence is seen most obviously in the use of large orchestras that employ both Western and Indian instruments. The influence of Western popular music, too, is very evident. In spite of the eclectic nature of Indian film music, most of the songs maintain an Indian feeling that arises largely from the vocal technique of the singers and the ornamentation of the melody line. This music is an experimental and developing form, and there have been attempts to add harmony and counterpoint, some of which may seem rather naïve to the Western ear. But the film music differs from typical Western music in that the melody line is generally not dictated by harmonic progressions and in that the harmonies used are incidental additions.

Aside from classical and film music, there are several other forms of urban music, some of which closely resemble the music of the rural areas. In city streets one is likely to encounter an outdoor band of oboes and drums announcing a wedding or a funeral. Street musicians, religious mendicants, snake charmers, storytellers, and magicians perform at every available opportunity, and work songs are sung by construction workers and other labourers. In private homes still other forms of music are performed, ranging from religious chanting to traditional folk and devotional songs. In public places of entertainment, the listener may encounter, apart from classical and film music, theatrical music from one of the relatively modern forms of regional theatre; and in the lowbrow places of entertainment courtesans still sing and dance in traditional fashion. In the larger cities there are performances of Western chamber music and occasionally symphony concerts, as well as popular dance music, rock, and jazz in the night clubs.

ANTIQUITY

In a musical tradition in which improvisation predominates, and written notation, when used, is skeletal and more a tool of the theorist than of the practicing musician, the music of past generations is irrevocably lost. References to music in ancient texts, aesthetic formulations, and depictions and written discussions of musical instruments can offer clues. In rare instances an ancient musical style may be preserved in unbroken oral tradition. For most historical eras and styles, surviving treatises explaining musical scales and modes—the framework of melody—provide a particularly important means of recapturing at least a suggestion of the music of former times, and tracing the musical theory of the past makes clear the position of the present musical system.

Little is known of the musical culture of the Indus Valley civilization of the 3rd and 2nd millennia BC. Some musical instruments, such as the arched, or bow-shaped, harp and more than one variety of drum, have been identified from the small terra-cotta figures and among the pictographs on the seals that were probably used by merchants. Further, it has been suggested that a bronze statuette of a dancing girl represents a class of temple dancers similar to those found much later in Hindu culture. It is known that the Indus civilization had established trade connections with the Mesopotamian civilizations, so that it is possible that the bow harp found in Sumeria would also have been known in the Indus Valley.

Vedic chant. *Compilation of hymns.* It is generally thought among scholars that the Indus Valley civilization was terminated by the arrival of bands of semi-nomadic tribesmen, the Aryans, who descended into India from the northwest, probably in the first half of the 2nd millennium BC. An important aspect of Aryan religious life was the bard-priest who composed hymns in praise of gods, to be sung or chanted at sacrifices. This tradition was continued in the invaders' new home in northern India until a sizable body of oral religious poetry had been composed. By about 1000 BC this body of chanted poetry had apparently grown to unmanageable proportions, and the best of the poems were formed into an anthology called Rigveda, which was then canonized. It was not committed to writing, but text and chanting formula were carefully handed down by word of mouth from one generation to the next, up to the present period. The poems in the Rigveda are arranged according to the priestly families who used and, presumably, had composed the hymns. Shortly after this a new Veda, called the Yajurveda, basically a methodical rearrangement of the verses of the Rigveda with certain additions in prose, was created to serve as a kind of manual for the priest officiating at the sacrifices. At approximately the same time, a third Veda, the Sāmaveda, was created for liturgical purposes. The Sāmaveda was also derived from the hymns of the Rigveda, but the words were distorted by the repetition of syllables, pauses, prolongations, and phonetic changes, as well as the insertion of certain meaningless syllables believed to have magical significance. A fourth Veda, the Atharvaveda, was accepted as a Veda considerably later and is quite unrelated to the other three. It represents the more popular aspects of the Aryan religion and consists mostly of magic spells and incantations.

tual
curacy
citation

Each of these Vedas has several ancillary texts, called the *Brāhmaṇas, Āraṇyakas,* and *Upaniṣads,* which are also regarded as part of the Vedas. These ancillary texts are concerned primarily with mystical speculations, symbolism, and the cosmological significance of the sacrifice. The Vedic literature was oral and not written down until very much later, the first reference to a written Vedic text being in the 10th century AD. In order to ensure the purity of the Vedas, the slightest change was forbidden, and the priests devised systems of checks and counterchecks, so that there has been virtually no change in these texts for about 3,000 years. Underlying this was the belief that the correct recitation of the Vedas was "the pivot of the universe" and that the slightest mistake would have disastrous cosmic consequence unless expiated by sacrifice and prayer. The Vedas are still chanted by the Brahmin priests at weddings, initiations, funerals, and the like, in the daily devotions of the priests, and at the now rarely held so-called public sacrifices.

From the Vedic literature it is apparent that music played an important part in the lives of the Aryan peoples, and there are references to stringed instruments, wind instruments, and several types of drums and cymbals. Songs, instrumental music, and dance are mentioned as being an integral part of some of the sacrificial ceremonies. The bow harp (*vīṇā*), a stringed instrument (probably a board zither) with 100 strings, and the bamboo flute were the most prominent melody instruments. Little is known of the music, however, apart from the Vedic chanting, which can still be heard today.

Chant intonation. The chanting of the Rigveda and Yajurveda shows, with some exceptions, a direct correlation with the grammar of the Vedic language. As in ancient Greek, the original Vedic language was accented, with the location of the accent often having a bearing on the meaning of the word. In the development of the Vedic language to Classical Sanskrit, the original accent was replaced by an automatic stress accent, whose location was determined by the length of the word and had no bearing on its meaning. It was thus imperative that the location of the original accent be inviolate if the Vedic texts were to be preserved accurately. The original Vedic accent occurs as a three-syllable pattern: the central syllable, called *udātta,* receives the main accent; the preceding syllable, *anudātta,* is a kind of preparation for the accent; and the following syllable, *svarita,* is a kind of return from accentuation to accentlessness. There is some difference of opinion among scholars as to the nature of the original Vedic accent; some have suggested that it was based on pitch, others on stress; and one theory proposes that it referred to the relative height of the tongue.

rbal
cent and
usical
tch

In the most common style of Rigvedic and Yajurvedic chanting found today, that of the Tamil Aiyar Brahmins, it is clear that the accent is differentiated in terms of pitch. This chanting is based on three tones; the *udātta* and the nonaccented syllables (called *pracaya*) are recited at a middle tone, the preceding *anudātta* syllable at a low tone, and the following *svarita* syllable either at the high tone (when the syllable is short) or as a combination of middle tone and high tone. The intonation of these tones is not precise, but the lower interval is very often about a whole tone, while the upper interval tends to be slightly smaller than a whole tone but slightly larger than a semitone. In this style of chanting the duration of the tones is also relative to the length of the syllables, the short syllables generally being half the duration of the long.

The more musical chanting of the Sāmaveda employs five, six, or seven tones and is said to be the source of the later secular and classical music. From some of the phonetic texts that follow the Vedic literature, it is apparent that certain elements of musical theory were known in Vedic circles, and there are references to three octave registers (*sthāna*), each containing seven notes (*yama*). An auxiliary text of the Sāmaveda, the *Nāradīśikṣā,* correlates the Vedic tones with the accents described above, suggesting that the Samavedic tones possibly derived from the accents. The Samavedic hymns as chanted by the Tamil Aiyar Brahmins are based on a mode similar to the D mode (D-d on the white notes of the piano; *i.e.,* the ecclesiastical Dorian mode). But the hymns seem to use three different-sized intervals, in contrast to the two sizes found in the Western church modes. They are approximately a whole tone, a semitone, and an intermediate tone. Once again, the intervals are not consistent and vary both from one chanter to another and within the framework of a single chant. The chants are entirely unaccompanied by instruments, and this may account for some of the extreme variation of intonation.

The changes brought by the 20th century have weakened the traditional prominent position of the Vedic chant. The Atharvaveda is seldom heard in India now. Samavedic chant, associated primarily with the large public sacrifices, also appears to be dying out. Even the Rigveda and Yajurveda are virtually extinct in some places, and South India is now the main stronghold of Vedic chant.

Music and
drama

The classical period. The ritual of the Vedas involves only the three upper classes, or castes, of Aryan society: the Brahmin, or priestly class; the Kṣatriya, or prince-warriors; and the Vaiśya, or merchants. The fourth caste, the Śūdras, or labourers, were excluded from Vedic rites. The primary sources of religious education and inspiration for the Śūdra were derived from what is sometimes called the fifth Veda: the epic poems *Rāmāyaṇa* and *Mahābhārata,* as well as the collections of legends, called the *Purāṇas,* depicting the lives of the various incarnations of the Hindu deities. The *Rāmāyaṇa* and the *Mahābhārata* were originally secular in character, describing the heroic deeds of kings and noblemen, many of whom are not recorded in history. Subsequently, religious matter was added, including the very famous sermon *Bhagavadgītā* ("Song of the Lord"), which has been referred to as the most important document of Hinduism; and many of the heroes of the epics were identified as incarnations of the Hindu deities. The legends were probably sung and recited by wandering minstrels and bards even before the advent of the Christian Era, in much the same way as they still are. The stories were also enacted on the stage, particularly at the time of the religious festivals. The earliest extant account of drama is to be found in the *Nāṭya-śāstra* ("Treatise on the Dramatic Arts"), a text that has been dated variously from the 2nd century BC to the 5th century AD and even later. It is virtually a handbook for the producer of stage plays and deals with all aspects of drama, including dance and music.

Theatrical music of the period apparently included songs sung on stage by the actors, as well as background music provided by an orchestra (which included singers) located offstage, in what was very like an orchestra pit. Melodies were composed on a system of modes, or *jāti*s, each of which was thought to evoke one or more particular sentiments (*rasa*) by its emphasis on specific notes. The modes were derived in turn from the 14 *mūrchanā*s—seven pairs of ascending seven-note series beginning on each of the notes of two closely related heptatonic (seven-note) parent scales, called *ṣaḍjagrāma* and *madhyamagrāma.* The *mūrchanā*s were thus more or less analogous to the European modal scales that begin progressively on D, E, F, G, etc. A third parent scale, *gāndhāragrāma,* was mentioned in several texts of the period and some even earlier but is not included in the system laid out in the *Nāṭya-śāstra.*

Qualities of the scales. The two parent scales differed in the positioning of just one note, which was microtonally flatter in one of the scales. The microtonal difference, referred to as *pramāṇa* ("measuring") *śruti,* presumably served as a standard of measurement. In terms of this standard it was determined that the intervals of the *mūrchanā*s were of three different sizes, consisting of two, three, or four *śruti*s, and that the octave comprised 22 *śrutis*. An interval of one *śruti* was not used. Several modern scholars have suggested that the *śruti*s were of unequal size; from the evidence in the *Nāṭya-śāstra,* it would appear, however, that they were thought to be equal. There has been no attempt to determine the exact size of the *śruti*s in any of the traditional Indian musical treatises until relatively modern times (18th century). The term *śruti* was also used to define consonance and dissonance, as these terms were understood in the period. In this connection, four terms are mentioned: *vādī,* comparable to the Western term sonant, meaning "having sound"; *saṃvādī,* to the Western consonant (concordant; reposeful); *vivādī,* to dissonant (discordant; lacking repose); and *anuvādī,* to assonant (neither consonant nor dissonant). As in the ancient Greek Pythagorean system, which influenced Western music, only fourths and fifths (intervals of four or five tones in a Western scale) were considered consonant. In the Indian system of measurement, tones separated by either nine or 13 *śruti*s correspond in size to Western fourths and fifths and are described as being consonant to each other. "Dissonant" in this system referred only to the minor second, an interval of two *śruti*s, and to its inversion (complementary interval), the major seventh (20 *śruti*s). All other tones, including the major third, were thought to be assonant.

Melodically important microtonal differences

The musical difference between the two parent scales is best seen not in terms of the microtonal deviation mentioned earlier but rather in terms of a musically influential consonance found in one but lacking in the other and vice versa. In each of the parent scales there are two nonconsonances, one of which is the tritone (interval of three Western whole tones, such as F-B) of 11 *śruti*s inevitable in all diatonic scales (seven-note scales of the major scale and *mūrchanā* type) and which in Europe during the Middle Ages was described as *diabolus in musica* ("the devil in music").

The second is a microtonal nonconsonance unique to this ancient Indian system. It can be illustrated by referring in the subsequent explanation to Table 1, in which the seven Indian notes *ṣaḍja, ṛṣabha, gāndhāra, madhyama, pañcama, dhaivata,* and *niṣāda* are given in their commonly abbreviated forms, *ṣa, ṛi, ga, ma, pa, dha,* and *ni.*

The nonconsonance arises from variances of one *śruti* from the fundamental consonances of the fourth and the fifth—a variance of about a quarter tone. In the *ṣaḍjagrāma* scale the interval *ṛi-pa* (E^- to A) contains 10 *śrutis; i.e.,* one more than the nine of the consonant fourth. Comparably, in the *madhyamagrāma* scale the interval *ṣa-pa* (D to A^-) contains 12 *śruti*s, or one fewer than the consonant fifth. These variances involve the consonant relationships of two melodically prominent notes, the first and the fifth. In the *madhyamagrāma* the first note, *ṣa,* has no consonant fifth, and perhaps for this reason this scale is said to begin not on the *ṣa* (D) but on its fourth, the note *ma* (G); hence, it resembles the G mode—*i.e.,* the ecclesiastical Mixolydian mode—whereas the *ṣaḍjagrāma* resembles the D mode, the ecclesiastical Dorian.

There is a striking resemblance of the *ṣaḍjagrāma* scale to the intervals used by the Tamil Aiyar Brahmins in their chanting of the Sāmaveda. Not only are their hymns set in a mode similar to the D mode, but they seem to use three different-sized intervals, the intermediate one corresponding to the three-*śruti* interval. The *Nāṭya-śāstra* claims to have derived song (*gīta*) from the chanting of the Sāmaveda, and the resemblances between the two may not be entirely fortuitous.

The two parent scales are complementary and between them supply all the consonances found in the ancient Greek Pythagorean scale. Thus, if in a mode the consonance *ṛi-pa* (E–A) were needed, one would tune to the *madhyamagrāma* scale. But, if the consonance *ṣa-pa* (D–A) were important, it could be obtained with the *ṣaḍjagrāma* tuning.

Introduction of new notes

There was a further development in this system caused by the introduction of two additional notes, called *antara ga* (F♯) and *kākalī ni* (C♯), which could be substituted for the usual *ga* (F) and *ni* (C). The *antara ga* eliminates the 11-*śruti* tritone between *ga* and *dha* (F–B), but its use creates a further tritone between F♯ and C. The second additional note, *kākalī ni* (C♯), eliminates this tritone but once again creates a new one, this time between C♯ and G. This process of adding notes, if carried further, would eventually lead to the circle, or, rather, the spiral, of fourths or fifths found in Western music (whereby a sequence of fifths, such as C–G, G–D, D–A, etc., leads eventually back to a microtonally out-of-tune C); there is no evidence that such a circle or spiral was known in ancient India.

Mode, or jāti. From each of the two parent scales were derived seven modal sequences (the *mūrchanās* described above), based on each of the seven notes. The two *mūrchanās* of a corresponding pair differed from each other only in the tuning of the note *pa* (A), the crucial distinction in the tunings of the two parent scales. One of each pair was selected as the basis for a "pure" mode, or *śuddha-jāti;* in the groups of seven pure modes, four used the tuning of the *ṣaḍjagrāma* and three that of the *madhyamagrāma.* In addition to these seven pure modes, a further 11 "mixed" modes, or *vikṛta-jāti*s, are also mentioned in the *Nāṭya-śāstra.* These were derived by a combination of two or more pure modes, but the text does not explain just in what way these derivations were accomplished.

The *jāti*s were similar to the modern concept of raga in that they provided the melodic basis for composition and, presumably, improvisation. They were not merely scales, but were also assigned 10 melodic characteristics: *graha,* the initial note; *aṃśa,* the predominant note; *tāra,* the note that forms the upper limit; *mandra,* the note that forms the lower limit; *nyāsa,* the final note; *apanyāsa,* the secondary final note; *alpatva,* the notes to be used infrequently; *bahutva,* the notes to be used frequently; *ṣāḍavita,* the note that must be omitted in order to create the hexatonic (six-note) version of the mode; and *auḍavita,* the two notes that must be omitted to create the pentatonic (five-note) version of the mode.

No written music survives from this early period. It is not clear from the description whether or not the music was like that of the present period. There is no mention of a drone, nor do the instruments of the orchestra—consisting of the *vipañcī* and *vīṇā* (bow harps?), bamboo flute, a variety of drums, and singers—appear to include any specifically drone instrument, such as the modern tamboura. The evidence tends rather to suggest, from the emphasis on consonance and some of the playing techniques, that some form of organum (two or more parts paralleling the same melody at distinct pitch levels) and even some type of rudimentary harmony may have been characteristic.

Table 1: Intervals of Ṣaḍjagrāma and Madhyamagrāma Parent Scales*

Ṣaḍjagrāma									**10 *śrutis***								
Indian notes:	(*ni*)		*ṣa*		*ṛi*		*ga*		*ma*		*pa*		*dha*		*ni*		(*ṣa*)
Śruti intervals:		4		3		2		4		4		3		2		(4)	
Comparable Western notes†:	(C)		D		E^-		F		G		A		B^-		c		(d)
										11 *śrutis*							
Madhyamagrāma								**12 *śrutis***									
Indian notes:	(*ni*)		*ṣa*		*ṛi*		*ga*		*ma*		*pa*		*dha*		*ni*		(*ṣa*)
Śruti intervals:		4		3		2		4		3		4		2		(4)	
Comparable Western notes†:	(C)		D		E		F		G		A^-		B^-		c		(d)
									11 *śrutis*								

*Minus signs indicate slightly lower pitch. †Without reference to precise pitch.

MEDIEVAL PERIOD

Precursors of the medieval system. It is not clear just when the *jāti* system fell into disuse, for later writers refer to *jāti*s merely out of reverence for Bharata, the author of the *Nāṭya-śāstra.* Later developments are based on musical entities called *grāma-rāga*s, of which seven are mentioned in the 7th-century Kuṭimiyāmalai rock inscription in Tamil Nadu state. Although the word *grāma-rāga* does not occur in the *Nāṭya-śāstra,* the names applied to the individual *grāma-rāga*s are all mentioned. Two of them, *ṣadjagrāma-rāga* and *madhyamagrāma-rāga,* are obviously related to the parent scales of the *jāti* system. The other five seem to be variants of these two *grāma-rāga*s in which either or both the altered forms of the notes *ga* and *ni* (F♯ and C♯) are used. In the *Nāṭya-śāstra,* the reference to the various *grāma-rāga*s is far removed from the main section in which the *jāti* system is discussed, and there is no obvious connection between the two. Each of the *grāma-rāga*s is said to be used in one of the seven formal stages of Sanskrit drama. They have been reconstructed as shown in Table 2.

Further development of the grāma-rāgas. In the next significant text on Indian music, the *Bṛhaddeśī,* written by the theorist Mātaṅga about the 10th century AD, the *grāma-rāga*s are said to derive from the *jāti*s. In some respects, at least, the *grāma-rāga*s resemble not the *jāti*s but their parent scales. The author of the *Bṛhaddeśī* claims to be the first to discuss the term raga in any detail. It is clear that raga was only one of several kinds of musical entities in this period and is described as having "varied and graceful ornaments, with emphasis on clear, even, and deep tones and having a charming elegance." The ragas of this period seem to have been named after the different peoples living in the various parts of the country, suggesting that their origin might lie in folk music. Mātaṅga appears to contrast the two terms *mārga* and *deśī. Mārga* (literally "the path") apparently refers to the ancient traditional musical material, whereas *deśī* (literally "the vulgar dialect spoken in the provinces") designates the musical practice that was evolving in the provinces, which may have had a more secular basis. Although the title *Bṛhaddeśī* ("The Great Deśī") suggests that the latter music might have been the focus of the treatise and that the *grāma-rāga*s were possibly out of date by the time it was written, the surviving portion of the text does not support such a theory.

The mammoth 13th-century text *Saṅgītaratnākara* ("Ocean of Music and Dance"), composed by the theorist Śārṅgadeva, is often said to be one of the most important landmarks in Indian music history. It was composed in the Deccan (south central India) shortly before the conquest of this region by the Muslim invaders and thus gives an account of Indian music before the full impact of Muslim influence. A large part of this work is devoted to *mārga*—that is, the ancient music that includes the system of *jāti*s and *grāma-rāgas*—but Śārṅgadeva mentions a total of 264 ragas. Despite the use in both the *Bṛhaddeśī* and the *Saṅgītaratnākara* of a notation equivalent to the Western tonic sol–fa (*i.e.,* with syllables, as do–re–mi . . .) to illustrate the ragas, modern scholars have not yet been able to reconstruct them with assurance.

The basic difficulty scholars face lies in determining the intervals used in each of the ragas. In the ancient system, the *jāti*s were something like the ancient Greek and medieval church modes in that each was derived from a parent scale by altering the ground note and the tessitura (range). In modern Indian music, however, the ragas are all transposed to a common ground note. This change may well be connected with the introduction of the drone and the evolution of the long-necked-lute family on which the drone is usually played. In the old system, with the changing ground note, it would have been necessary to retune drone instruments from one raga to another, which would have been a cumbersome and impractical operation to carry out during a recital. It may have been this factor that provided the impetus for the change to the standard-ground-note system. There is no conclusive evidence to show just when this change might have taken place, and it is not clear whether the *Bṛhaddeśī* and the *Saṅgītaratnākara* are using the old ground-note system or one similar to that used in modern times.

THE ISLĀMIC PERIOD

Impact on musical genres and aesthetics. The Muslim conquest of India can be said to begin in the 12th century, although Sind (now in Pakistan) had been conquered by the Arabs as early as the 8th century. Muslim writers such as al-Jāḥiẓ and al-Masʿūdī had already commented favourably on Indian music in the 9th and 10th centuries, and the Muslims in India seem to have been very much attracted by it.

In the beginning of the 14th century the great poet Amīr Khosrow, who was considered to be extremely proficient in both Persian and Indian music, wrote that Indian music was superior to the music of any other country. Further, it is stated that, after the Muslim conquest of the Deccan under Malik Kāfūr (*c.* 1310), a large number of Hindu musicians were taken with the royal armies and settled

Table 2: Grāma-Rāgas

	scale														
Madhyama-grāma-rāga															
Indian notes	*ṣa*		*ṛi*		*ga*		*ma*		*pa*		*dha*		*ni*		*ṣa*
Śruti values		3		2		4		3		4		2		4	
Comparable Western notes	D		E$^-$		F		G		A$^-$		B$^-$		c		d
Ṣadja-grāma-rāga															
Indian notes	*ṣa*		*ṛi*		*ga*		*ma*		*pa*		*dha*		*ni*		*ṣa*
Śruti values		3		2		4		4		3		2		4	
Comparable Western notes	D		E$^-$		F		G		A		B$^-$		c		d
Ṣādava-grāma-rāga					(a)*										
Indian notes	*ṣa*		*ṛi*		*ga*		*ma*		*pa*		*dha*		*ni*		*ṣa*
Śruti values		3		4		2		3		4		2		4	
Comparable Western notes	D		E$^-$		F♯		G		A$^-$		B$^-$		c		d
Pañcama-grāma-rāga					(a)*										
Indian notes	*ṣa*		*ṛi*		*ga*		*ma*		*pa*		*dha*		*ni*		*ṣa*
Śruti values		3		4		2		4		3		2		4	
Comparable Western notes	D		E$^-$		F♯		G		A		B$^-$		c		d
Kaiśika-grāma-rāga					(a)*								(k)*		
Indian notes	*ṣa*		*ṛi*		*ga*		*ma*		*pa*		*dha*		*ni*		*ṣa*
Śruti values		3		4		2		3		4		4		2	
Comparable Western notes	D		E$^-$		F♯		G		A$^-$		B$^-$		c♯		d
Sādhārita-grāma-rāga					(a)*								(k)*		
Indian notes	*ṣa*		*ṛi*		*ga*		*ma*		*pa*		*dha*		*ni*		*ṣa*
Śruti values		3		4		2		4		3		4		2	
Comparable Western notes	D		E$^-$		F♯		G		A		B$^-$		c♯		d
Kaiśika-madhyama-grāma-rāga					(a)*								(k)*		
Indian notes	*ṣa*		*ṛi*		*ga*		*ma*				*dha*		*ni*		*ṣa*
Śruti values		3		4		2			7			4		2	
Comparable Western notes	D		E$^-$		F♯		G				B$^-$		c♯		d

*(a) and (k) refer to *antara* and *kākalī*, the variant forms of the notes *ga* and *ni*.

in the North. Although orthodox Islām considered music illegal, the acceptance of the Ṣūfī doctrines, in which music was an accepted means to the realization of God, enabled Muslim rulers and noblemen to extend their patronage to this art. At the courts of the Mughal emperors Akbar, Jahāngīr, and Shāh Jahān, music flourished on a grand scale. Apart from Indian musicians, there were also musicians from Persia, Afghanistan, and Kashmir in the employ of these rulers; nevertheless, it appears that it was Indian music that was most favoured. Famous Indian musicians, such as Svāmī Haridās and Tānsēn, are legendary performers and innovators of this period. After the example set by Amīr Khosrow, Muslim musicians took an active interest in the performance of Indian music and added to the repertoire by inventing new ragas, *tāla*s, and musical forms, as well as new instruments.

Influence of religious movements

The Muslim patronage of music was largely effective in the north of India and has had a profound influence on North Indian music. Perhaps the main result of this influence was to de-emphasize the importance of the words of the songs, which were mostly based on Hindu devotional themes. In addition, the songs had been generally composed in Sanskrit, a language that had ceased to be a medium of communication except among scholars and priests. Sanskrit songs were gradually replaced by compositions in the various dialects of Hindi, Braj Bhasa, Bhojpuri, and Dakhani, as well as in Urdu and Persian; nevertheless, the problems of communication, in terms of both language and subject matter, were not easily reconciled. A new approach to religion was, in any case, sweeping through India at about this time. This emphasized devotion (*bhakti*) as a primary means to achieving union with God, bypassing the traditional Hindu beliefs of the transmigration of the soul from body to body in the lengthy process of purification before it could achieve the Godhead. The Islāmic Ṣūfī movement was based on an approach similar to that of the *bhakti* movements and also gained many converts in India. A manifestation of these devotional cults was the growth of a new form of mystic-devotional poetry composed by wandering mendicants who had dedicated their lives to the realization of God. Many of these mendicants have been sanctified and are referred to as poet-saints or singer-saints, since their poems were invariably set to music. A number of devotional sects sprang up all over the country, some Muslim, some Hindu, and others merging elements from both. These sects emphasized the individual's personal relationship with God. In their poetry, man's love for God was often represented as a woman's love for man and, specifically, the love of the milkmaid Rādhā for Krishna, a popular incarnation of the Hindu god Vishnu. In the environment of the royal courts, there was a less idealistic interpretation of the word love, and much of the poetry, as well as the miniature painting, of the period depicts the states of experience of the lover and the beloved.

This attitude is also reflected in the musical literature of the period. From early times, both *jāti*s and ragas in their connection with dramatic performance were described as evoking specific sentiments (*rasa*) and being suitable for accompanying particular dramatic events. It was this connotational aspect, rather than the technical one, that gained precedence in this period. The most popular method of classification was in terms of ragas (masculine) and their wives, called *rāgiṇī*s, which was extended to include *putra*s, their sons, and *bhāryā*s, the wives of the sons. The ragas were personified and associated with particular scenes, some of which were taken from Hindu mythology, while others represented aspects of the relationship between two lovers. The climax of this personification is found in the *rāgamāla* paintings, usually in a series of 36, which depict the ragas and *rāgiṇī*s in their emotive settings.

Theoretical developments. From the middle of the 16th century, a new method of describing ragas is found in musical literature. It was also at about this time that the distinction between North and South Indian music became clearly evident. In the literature, ragas are described in terms of scales having a common ground note. These scales were called *mela* in the South and *mela* or *thāṭa* in the North.

Emergenc of moder theoretica concepts

It was in the South that a complete theoretical system of *mela*s was introduced, in the *Caturdaṇḍiprakāśika* ("The Illuminator of the Four Pillars of Music"), a text written in the middle of the 17th century. This system was based on the permutations of the tones and semitones, which had by this time been reduced to a basic 12 in the octave. The octave was divided into two tetrachords, or four-note sequences, C–F and G–c, and six possible tetrachord species were arranged in an order showing their relationship with each other. It will be noted in the sequence of tetrachords shown below that each lower tetrachord has an analogous upper tetrachord and that the outer notes of each are constant, whereas the inner notes change their pitch.

1.	C	D♭	E♭♭	F	and	G	A♭	B♭♭	c
2.	C	D♭	E♭	F		G	A♭	B♭	c
3.	C	D♭	E	F		G	A♭	B	c
4.	C	D	E♭	F		G	A	B♭	c
5.	C	D	E	F		G	A	B	c
6.	C	D♯	E	F		G	A♯	B	c

The list could have extended further, except that apparently no pitch distinction was made between the enharmonic pairs D–E♭♭, D♯–E♭, A–B♭♭, and A♯–B♭. (Enharmonic notes have different pitch names but sound either the same pitch or, in some tuning systems, have very slight differences in pitch.)

By utilizing all possible combinations of a lower with an upper tetrachord, 36 *mela*s, or raga scales, were derived; a further 36 were formed by using F♯ in place of the F in the lower tetrachord. The *mela*s were named in such a way that the first two syllables of the name, when applied in a code, gave the number of that *mela* in the sequence. The musician, given the number, could easily reconstruct the scale of the *mela*. The names of the *mela*s were often derived from prominent ragas in those *mela*s, with a two-syllable prefix that supplied the code numbers; for instance, the name of the *mela Dhīra-śaṅkarābharaṇa* is derived from the raga *Śaṅkarābharaṇa,* the two syllables *dhīra* giving the code number 29, which indicates a scale similar to the Western major scale, or C mode. The *Caturdaṇḍiprakāśika* acknowledges the theoretical nature of its analytical system and mentions clearly that only 19 of the possible 72 *mela*s were in use at the time that the text was written.

Although North Indian texts also describe ragas in terms of *mela*s or *thāṭa*s, there is no attempt to arrange them systematically. In the *Rāgataraṅgiṇi* ("The River of Rāga"), probably of the 16th century, 12 *mela*s are mentioned:

bhairavī	C	D	E♭	F	G	A	B♭	c
toṛī	C	D♭	E♭	F	G	A♭	B♭	c
gaurī	C	D♭	E	F	G	A♭	B	c
karṇāta	C	D	E	F	G	A	B♭	c
kedāra	C	D	E	F	G	A	B	c
imana	C	D	E	F♯	G	A	B	c
sāraṅga	C	D	E♯	F♯	G	A♯	B	c
megha	C	D	E	F	G	A♯	B	c
dhanāśrī	C	D♭	E	F♯	G	A♭	B	c
pūravā	C	D	E	F♯	G	A$^+$	B	c
mukhārī	C	D	E♭	F	G	A♭	B♭	c
dīpaka	no description							

Although it appears from the description of *sāraṅga* and *megha mela*s that enharmonic intervals were used, there is good reason to believe that the E♯ and A♯ in the two *mela*s really represent their chromatic counterparts, F and B♭, and that F and F♯ (and B and B♭) do not appear in sequence. The A$^+$ in the *mela pūravā* is said to be raised by one *śruti.* The description of the ragas in these *mela*s shows that the North Indian system was by this time also based on 12 semitones.

THE MODERN PERIOD

With the collapse of the Mughal Empire in the 18th century and the emergence of the British as a dominant power in India, the subcontinent was divided into many princely states. Music continued to be patronized by the rulers, although the courts were never again to achieve their former opulence.

Musically, there has been a continuous evolution from the Islāmic period to the present, and both North and

South Indian classical music have continued to expand. South Indian music has clearly been influenced more by theory than has the North. The 72-*mela* system continues to be the basis of classifying the ragas in South India, but it has had more than a classificatory significance. Many new ragas have been composed in the past few centuries, some of them inspired by the theoretical scales of the *mela* system. As a result, there are now ragas in all of the 72 *mela*s.

luence
theory
compo-
on

In North Indian music, theory has had little influence on performance practice. This can be ascribed to the language problem, an especially significant influence on the many Muslim musicians in North India, who were not able to cope with the Sanskrit musical literature. Thus, there had been no attempt to systematize the music, and there was a considerable gap between performance and theory until the present century. Vishnu Narayana Bhatkande, one of the leading Indian musicologists of this century, contributed a great deal toward diminishing the gap. Being both a scholar and a performer, he devoted much effort to collecting and notating representative versions of a number of ragas from musicians belonging to different family traditions, or *gharānā*s. Based on this collection, he concluded that most of the ragas of North Indian music can be grouped into the following scales, called *thāṭa*s (compare the South Indian *mela*s shown above):

kalyāṇa	C	D	E	F♯	G	A	B	c
bilāvala	C	D	E	F	G	A	B	c
khamāja	C	D	E	F	G	A	B♭	c
bhairava	C	D♭	E	F	G	A♭	B	c
pūrvī	C	D♭	E	F♯	G	A♭	B	c
mārvā	C	D♭	E	F♯	G	A	B	c
kāfī	C	D	E♭	F	G	A	B♭	c
āsāvarī	C	D	E♭	F	G	A♭	B♭	c
bhairavī	C	D♭	E♭	F	G	A♭	B♭	c
toṛī	C	D♭	E♭	F♯	G	A♭	B	c

The *thāṭa*s do not cover all the ragas used in North Indian music, but there is reason to believe that most of the ragas having scales other than the above are relatively modern innovations. New ragas are constantly being created, and some North Indian musicians are using the vast potential of the South Indian *mela* system as their source of inspiration.

le
ısi-
tions as
ıosed to
ı

Mela and *thāṭa* are theoretical devices for the classification of ragas. Ragas have scalar elements, such as specified ascending and descending movements, that might or might not employ adjacent steps. They may also employ oblique or zigzag movements. Ragas can be heptatonic, hexatonic, or pentatonic and may also have accidentals (sharpened or flattened notes) that occur only in specific melodic contexts. A further distinction between scale and raga is found in the varying emphasis placed on different notes in a raga. Ragas, furthermore, also have melodic elements, such as certain recurrent nuclear motives (brief melodic fragments) that enable the raga to be identified more easily. One scale type can be the basis for perhaps 20 or 30 ragas, in which case it is the nonscalar elements that provide the distinguishing features of each raga in the group.

Rhythmic organization. *South India.* Just as the system of classifying raga is better organized in South Indian music, so too is the system of classifying *tāla,* or time measure. The main group is composed of 35 *tāla*s, called the *sūlādi-tāla*s. Each *tāla* is composed of one, two, or three different units: short, medium, and long. The medium unit is twice the duration of the short; the long unit is, however, a variable and may be three, four, five, seven, or nine times the duration of the short. There are seven basic *tāla* patterns, and, because the long unit of these *tāla*s can be of five different durations, the total number of *tāla*s in this system is 35. The basic *tāla* patterns are:

dhruva-tāla—long, medium, long, long
maṭhya-tāla—long, medium, long
rūpaka-tāla—medium, long
jhampā-tāla—long, short, medium
tripuṭa-tāla—long, medium, medium
āṭa-tāla—long, long, medium, medium
eka-tāla—only a single long.

The total duration of each pattern is controlled by the duration of the variable long; thus, if the long unit is five times the short, a *tāla* pattern such as *dhruva-tāla* will be 5 + 2 + 5 + 5, or 17 units. Several of these *tāla*s have the same total duration but are distinguished from each other by their internal subdivisions. In the course of a performance, the vocalist, as well as the audience, may mark the time by clapping, hand waving, and finger counting.

In addition to the *sūlādi-tāla*s, there are four *cāpu-tāla*s that are used in South Indian classical music. Said to derive from folk music, they consist of two sections of unequal length, 1 + 2, 2 + 3, 3 + 4, and 4 + 5. Of these, the 3 + 4 combination is the most prominent. On rare occasions a performer may use one of the "classical" *tāla*s referred to in Sanskrit texts. These generally involve long time cycles composed of as many as 100 short units. The most frequently heard time measures, however, are *ādi-tāla,* a modified eight-beat version of *tripuṭa-tāla* (4 + 2 + 2); *miśra-cāpu-tāla* (3 + 4); and *rūpaka-tāla* (4 + 2). The difficult and long *tāla*s are used primarily as a tour de force. Each *tāla* may be performed in either slow, medium, or quick tempo; there is no gradual acceleration as in North Indian music.

North India. In North Indian music the *tāla*s are fewer and not organized in any systematic manner. As in South Indian music, the two main factors are the duration of the time cycle and the subdivisions within the cycle. Each of these subdivisions is marked by a clap or a wave, with the greatest emphasis falling on beat 1 of the cycle, which is called *sam.* North Indian *tāla*s have a further feature, the *khālī* ("empty"), a conscious negation of stress occurring at one or more points in each *tāla* where one would expect a beat. It often falls at the halfway point in the time cycle and is marked by a wave of the hand. There is nothing comparable to the *khālī* in the South Indian system. A further distinguishing feature found only in North Indian *tāla*s is the emphasis placed on the characteristic drum pattern of each *tāla,* called *ṭhekā.* Two *tāla*s might have the same duration and subdivisions but might, nevertheless, be differentiated from each other by different characteristic drum patterns. In addition, the *tāla*s are also associated with different forms of song and even particular tempi. The usual North Indian *tāla*s range from six to 16 time units in duration. The most popular are *tīn-tāla* (4 + 4 + 4 + 4), *eka-tāla* (2 + 2 + 2 + 2 + 2 + 2), *jhap-tāla* (2 + 3 + 2 + 3), *kaharavā* (4 + 4), *rūpaka-tāla* (3 + 2 + 2), and *dādrā* (3 + 3). *Tīn-tāla* should not be confused with Western $\frac{4}{4}$, or common time, for the time cycle repeats only after 16 units and is more like four bars of common time.

Characteristic North Indian rhythmic features

Musical forms and instruments. *South India.* Both raga and *tāla* provide bases for composition and improvisation in Indian classical music. A performance usually begins with an improvised section, called *ālāpa,* played in free time without accompaniment of drums. It may have various sections and might on occasion last half an hour or longer. It is followed by a composed piece in the same raga, set in a particular *tāla.* In South Indian music all composed pieces are primarily for the voice and have lyrics. In North India, however, there are also some purely instrumental compositions, called *gat* and *dhun.* The emphasis on the composition varies in the different forms of song and, to some extent, in the interpretation of the performer. In South Indian music the composed piece is generally emphasized more than in the North. Much of the South Indian repertoire of compositions stems from three composers, Tyagaraja, Muthuswami Dikshitar, and Syama Sastri, contemporaries who lived in the second half of the 18th and the beginning of the 19th centuries. The devotional songs that they composed, called *kṛti,* are a delicate blend of text, melody, and rhythm and are the most popular items of a South Indian concert. The composed elements in these songs sometimes include sections such as *niraval,* melodic variations with the same text, and *svara-kalpana,* passages using the Indian equivalent of the sol–fa syllables, which are otherwise improvised.

The longest item in the South Indian concert, called *rāgam-tānam-pallavi,* is, on the other hand, mostly improvised. It begins with a long *ālāpa,* called *rāgam* in

Interaction of composition and improvisation

this context, presumably because this elaborate, gradually developing *ālāpa* is intended to display the raga being performed in as complete a manner as possible, without the limitations imposed by a fixed time measure. This is followed by another improvised section, *tānam,* in which the singer uses meaningless words to produce more or less regular rhythms, but still without reference to time measure. This section, too, is without drum accompaniment. The final section, *pallavi,* is a composition of words and melody set in a particular *tāla,* usually a long or complex one. The *pallavi* may have been composed by the performer himself and be unfamiliar to his accompanists, usually a violinist who echoes the singer's phrases and a drummer who plays the mridanga, a double-ended drum. The statement of the composition is followed by elaborate rhythmic and melodic variations that the accompanists are expected to follow. It is customary to have a drum solo at the end of the *pallavi,* and the performance concludes with a brief restatement of the *pallavi.*

Other forms used in South Indian classical music derive largely from the musical repertoire of *bhārata-nāṭyam,* the classical South Indian dance. The *varṇam,* a completely composed piece, serves mainly as a warming up and is performed at the beginning of a concert. *Pada* and *jāvali* are two kinds of love songs using the poetic imagery characteristic of the romantic-devotional movement mentioned earlier. *Tillānā* has a text composed mostly of meaningless syllables, which may include the onomatopoeic syllables used to represent the different drum sounds. This is a very rhythmic piece and is usually sung in fast tempo.

Melodic, drone, and rhythmic instruments

The ensemble used in present-day South Indian classical music consists of a singer or a main melody instrument, a secondary melody instrument, one or more rhythmic percussion instruments, and one or more drone instruments. The most commonly heard main melody instruments are the *vīṇā,* a long-necked, fretted, plucked lute with seven strings; the *veṇu,* a side-blown bamboo flute; the *nagaswaram,* a long, oboe-like, double-reed instrument with finger holes; the violin, imported from the West about 200 years ago, played while seated on the floor with the scroll resting on the player's left foot; and the *goṭṭuvādyam,* a long-necked lute without frets, played like the Hawaiian guitar, with a sliding stop in the left hand.

The violin is by far the most common secondary melody instrument in South India. It plays in unison where the passage is composed but imitates the voice or main melody instrument in the improvised passages. Of the rhythm instruments, the mridanga, a double-conical, two-headed drum, is the most common. Others include the *kañjīrā,* a tambourine; the *ghaṭam,* an earthenware pot without skin covering; the *morsiṅg,* a metallic jew's harp; and the *tavil,* a slightly barrel-shaped, double-ended drum, which accompanies the *nagaswaram.* The most prominent drone instrument is the four-stringed tamboura, a long-necked lute without frets. It accompanies the voice and all melody instruments, except the *nagaswaram,* which is usually accompanied by the *oṭṭu,* a longer version of the *nagaswaram* but without finger holes. A hand-pumped harmonium drone, called *śruti* or *śruti* box, sometimes replaces the *oṭṭu* or the tamboura.

North India. The most common vocal form in North Indian classical music at the present time is the *khyāl,* a Muslim word meaning "imagination." The *khyāl* is contrasted with the *dhruvapada* (now known as *dhrupad*), which means "fixed words." The two forms existed side by side in the Islāmic period, and it is only in the last century or two that *khyāl* has achieved ascendancy. There are two types of *khyāl.* The first is sung in extremely slow tempo, with each syllable of the text having extensive melisma (prolongation of a syllable over many notes), so that the words are virtually unrecognizable. It is not usually preceded by a lengthy *ālāpa;* instead, *ālāpa*-like phrases are generally sung against the very slow time measure to the accompaniment of the drums. Also characteristic of the *khyāl* are the *sargam tānas,* passages using the Indian equivalent of the sol-fa syllables, and the *ā-kār tānas,* which are rapid runs sung to the syllable *aah.* The second type of *khyāl,* which may be as much as eight times faster than the slow and is generally set in a different *tāla,* follows the slow. Its composed portion is usually quite short, and the main features of the improvisation are the *ā-kār tānas.* Occasionally, a composition called *tarānā,* made up of meaningless syllables, may replace the fast-tempo *khyāl.*

The *ṭhumrī* is another North Indian vocal form and is based on the romantic-devotional literature inspired by the *bhakti* movement. The text is usually derived from the Rādha-Krishna theme and is of primary importance. The words are strictly adhered to, and the singer attempts to interpret them with his melodic improvisations. It is quite usual for a singer to deviate momentarily from the raga in which the composition is set, by using accidentals and evoking other ragas that might be suggested by the words, but he always returns to the original raga.

Some of the North Indian musical forms are very like the South Indian. The vocal forms *dhrupad* and *dhamār* resemble the *rāgam-tānam-pallavi.* They begin with an elaborate *ālāpa* followed by the more rhythmic but unmeasured *non-tom* using meaningless syllables such as *te, re, na, nom,* and *tom.* Then follow the four composed sections of the *dhrupad* or *dhamār,* the latter being named after *dhamār-tāla* of 14 units (5 + 5 + 4) in which it is composed, the former name derived from *dhruvapada.* The song, usually in slow or medium tempo, is first sung as composed; then the performer introduces variations, the words often being distorted and serving merely as a vehicle for the melodic and rhythmic improvisations. Although the *dhrupad-dhamār* form has been out of favour for over a century, it is now apparently being revived.

Instrumental forms

Instrumental music has gained considerable prominence in North India in recent times. The most common instrumental form is the *gat,* which seems to have derived its elements from both *dhrupad* and *khyāl.* It is usually preceded by *ālāpa* and *joṛ,* which resemble the *ālāpa* and *non-tom* sections of the *dhrupad.* On plucked stringed instruments these two movements are often followed by *jhālā,* a fast section in which the rhythmic plucking of the drone strings is used to achieve a climax. The performer usually pauses before the composed *gat* is introduced. Like the *khyāl,* the *gat* can be in slow or fast tempo. The composition is generally short, and the emphasis is on the improvisations of the melody instrumentalist and the drummer, who for the most part alternate in their extemporizing. The final climax may once again be achieved by a *jhālā* section, in which the tempo is accelerated quite considerably. Other forms played on instruments are the *ṭhumrī,* basically an instrumental rendering of a vocal *ṭhumrī,* and *dhun,* which is derived from a folk tune and does not usually follow a conventional raga. One may also hear a piece called *rāga-māla* (literally, "a garland of ragas"), in which the musician modulates from one raga to another, finally concluding with a return to the original raga.

Musical instruments of North India

The most prominent melody instruments used in North Indian classical music are the sitar, a long-necked fretted lute; *surbahār,* a larger version of the sitar; the sarod, a plucked lute without frets and a shorter neck than that of the sitar; the *sāraṅgī,* a short-necked bowed lute; the bansuri, a side-blown bamboo flute with six or seven finger holes; the *sheh'nai,* a double-reed wind instrument similar to the oboe, but without keys; and the violin, played in the same manner as in South India. Secondary melody instruments are used only in vocal music, the two most common being the *sāraṅgī* and the keyboard harmonium, an import from the West. The violin and the *surmaṇḍal,* a plucked board zither, are also used in this context. In recent times, instrumental duets, in which the musicians improvise alternately, have grown in popularity. In these duets the musicians may imitate each other's phrases, temporarily creating something of the effect of a secondary melody instrument.

As with South Indian music, the drone is usually provided by a tamboura (Bengali *tanpura*) or a hand-pumped reed drone similar to the harmonium but without a keyboard, called *sur-peṭī* in North India. The *sheh'nai* is usually accompanied by one or more drone *shehnais,* called *sur.*

The rhythmic accompaniment is usually provided on the tabla, a pair of small drums played with the fingers. As accompaniment to the somewhat archaic *dhrupad,* however, the pakhavāj, a double-conical drum, similar to the South Indian mridanga, is generally used. The *sheh'nai* in classical music is usually accompanied by a small pair of kettledrums, called *ḍukaṛ-ṭikaṛ.*

Interaction with Western music. It is in the sphere of musical instruments that the influence of Western music is most obvious. In addition to the violin and the harmonium, many other Western instruments are occasionally used. These include the clarinet, saxophone, trumpet, guitar, mandolin, and organ. Scholars have criticized the use of some of these instruments on the ground that their tuning, being based on the Western tempered scale (having 12 equal semitones), is not suitable for the performance of Indian music, and All-India Radio has forbidden the use of the harmonium in its programs. Most of the leading North Indian singers, however, have been using the harmonium as a secondary melody instrument for many years and have continued to do so in concerts and on recordings.

Apart from the area of musical instruments, Indian music appears to have absorbed very little of Western music. It is possible that some modern developments in classical music might have been inspired by Western music. These include the slightly increased use of chromaticism (using a succession of semitones) and some of the new drone tunings in which the major third is added (making for example, the drone on the first, third, and fifth notes of the scale, rather than on the first and fifth only). But the evidence is not conclusive, and it could equally be argued that these are natural developments within the system. Western technology has, of course, had a profound influence on Indian music. Sound-amplification devices have made concerts available to large audiences, and the intimate atmosphere in which the music was traditionally performed is now seldom encountered. The Indian musician has been obliged to adapt his music, once played before a select and musically educated group of listeners, to new circumstances involving a mass of people, many of whom are unable to appreciate the finer points of the music. The use of microphones during concerts has had a marked effect on voice production, and, since the voice no longer needs to project over distances, many modern singers now sing with a relaxed throat and produce a more mellow tone.

ıpact
ıtside
India

Since the mid-1950s, Indian classical music has been performed fairly regularly in the West. Initially, the audiences were composed mainly of South Asians, but gradually an increasing number of Westerners have been attending the concerts. Perhaps the music would not have reached beyond a very limited audience were it not for the interest shown by the American violinist Yehudi Menuhin, who sponsored a number of programs in the West, and the British popular-music group the Beatles, who attempted to incorporate the sound of the sitar and other elements of Indian culture into the world of Western popular music. At the same time, several North Indian instrumentalists, such as Ravi Shankar, Ali Akbar Khan, Vilayat Khan, Imrat Khan, and Nikhil Banerjee, were received with overwhelming enthusiasm by Western audiences. By the end of the 1960s the sitar and tabla were heard frequently in Western pop music, jazz, cinema, and television programs, as well as in radio and television advertisements.

Within three or four years, the mass Western involvement with Indian music was over. It is perhaps too early to assess the full impact of this period. It would seem that Indian music has not as yet had any significant influence on Western music. A few modern composers have attempted to incorporate elements of Indian music into their compositions, but their works remain as experiments. The fusing of Indian music and jazz would seem to have more possibilities, since improvisation is an important factor in both, but present attempts have not fulfilled this expectation. Most of these attempts seem to be premature and based on an inadequate understanding of one or the other musical system. (N.A.J.)

Dance and theatre

Theatre and dance in South Asia stem principally from Indian tradition. The principles of aesthetics and gesture language in the *Nāṭya-śāstra,* a 2,000-year-old Sanskrit treatise on dramaturgy, have been the mainstay of all the traditional dancers and actors in India. Even folk performers follow some of its conventions; *e.g.,* the *Kandyan* dancers of Ceylon (now Sri Lanka), who preserve some of the whirls and spins described in this ancient Indian text. Despite the influence of the different religious waves that swept the subcontinent through the centuries, the forms of South Asian dance and theatre were always able to preserve their ancient core.

Traditionally, dance and acting are inseparable. The classical South Asian dancer, equipped with a repertoire of gesture language, alternates between *nṛtta,* pure dance; *nṛtya,* interpretive dance; and *nāṭya,* dance with a dramatic element. (The Sanskrit word *naṭa* means a dancer-actor.) Traditional theatre throughout both South and Southeast Asia is a combination of music, dance, mime, stylized speech, and spectacle. The classical and folk actor must be a dancer, a singer, and a mime in one.

Influence
on South-
east Asia

Between the 2nd century BC and the 8th century AD, South Indian kings sent overseas trade missions, priests, court dancers, and sometimes armies to Southeast Asia. During these years of cultural expansion, Indian dance forms, mythological lore, and the language of gesture flourished in Burma, Cambodia, Java, Sumatra, and Bali. Later, when India's economic and political power shrank, its cultural empire remained intact. Even when these Southeast Asian countries embraced Buddhism or Islām, they continued performing dance dramas with Hindu gods and goddesses, adding to these their own local myths, costumes, and masks. The two Hindu epics, the *Rāmāyaṇa* and the *Mahābhārata,* storehouses of dramatic personae of traditional dramas, have been absorbed by these countries as part of their own cultural heritage. Some dance forms and gesture vocabulary that died out in their land of birth have been preserved in Bali. For a discussion of the dance and theatre of Southeast Asia see the article SOUTHEAST ASIAN ARTS.

THE PERFORMING ARTS IN INDIA

The royal courts and temples of India traditionally have been the chief centres of the performing arts. In ancient times, Sanskrit dramas were staged at seasonal festivals or to celebrate special events. Some kings were themselves playwrights; the most notable of the playwright-kings was Śūdraka, the supposed 4th-century author of *Mṛcchakaṭika* (The Little Clay Cart). Other well-known royal dramatists include Harṣa, who wrote *Ratnāvalī* in the 7th century; Mahendravikramavarman, author of the 7th-century play *Bhagavad-Ajjukiya;* and Viśākhadatta, creator of the 9th-century drama *Mudrārākṣasa.*

Royal
patronage

In the 4th century BC, Kauṭilya, the chief minister of Emperor Candragupta, referred in his book on the art of government, the *Artha-śāstra,* to the low morals of players and advised the municipal authorities not to build houses in the midst of their villages for actors, acrobats, and mummers. But, in the glorious era of the Hindu kings during the first eight centuries after Christ, actors and dancers were given special places of distinction. This tradition continued in the princely courts of India even under British rule. *Kathākali* dance-drama, for instance, was created by the Raja of Kottarakkara, ruler of one of the states of South India in the 17th century. The powerful peshwas of the Marāthā kingdom in the 18th century patronized the *tamasha* folk theatre. Nawab Wajid Ali Shah (flourished mid-19th-century) was an expert *kathak* dancer and producer of Krishnalore plays in which his palace maids danced as the *gopīs* (milkmaids who were devotees of Krishna). Maharajas of Travancore and Mysore competed with each other for the excellence of their dance troupes. In the 20th century, the Maharaja of Banaras (Vārānasi) carried on this tradition by being patron and producer of the spectacular *rāmlīlā,* a 31-day cycle play on Rāma's life that he witnessed every night while sitting on his royal elephant. On special nights the spectators numbered more than 30,000.

Dance is a part of all Hindu rituals. Farmers dance for a plentiful harvest, hunters for a rich bag, fishermen for a good catch. Seasonal festivals, religious fairs, marriages, and births are celebrated by community dancing. A warrior dances before the image of his goddess and receives her blessings before he leaves for battle. A temple girl dances to please her god. The gods dance in joy, in anger, in triumph. The world itself was created by the Cosmic Dance of Lord Śiva, who is called Naṭarāja, the king of dancers, and worshipped by actors and dancers as their patron.

Religious festivals

Religious festivals are still the most important occasions for dance and theatrical activity. The *rāmlīlā kṛṣṇalīlā* and *rāslīlā* in North India (Uttar Pradesh, Delhi, Rājasthān, Haryana, and Punjab), the *chhau* masked dance-drama in Saraikela region in Bihār, and the *bhagavatha mela* in Melatur village in Tamil Nadu are performed annually to celebrate the glory of their particular deities. During the Daśaharā festival every village in North India enacts for a fortnight the story of Rāma's life, with songs, dances and pageants. The *jātrā* in West Bengal is a year-round dramatic activity, but the number of troupes swells to many thousands in Calcutta during the Pūjā festival. The hill and tribal people dance all night to celebrate their community festivals and weddings rich in masks, pageants, and carnivals. In far-flung areas of South Asia, people may not have seen a drama, but there will be hardly a person who has not witnessed or taken part in a community dance.

Audience participation

In folk theatre, traditional dance, classical music, and poetical symposia (especially the Urdu *mushā'irah*), performances are held in the open air or in a well-lit canopied courtyard so that the players can see the spectators and be motivated by their reactions.

For the usually all-night folk dramas, people come with their children, straw mats, and snacks, making themselves at home. At these performances there is a constant inflow and outflow of spectators. Some go to sleep, asking their neighbours to awaken them for favourite scenes. Stalls selling betel leaves, peanuts, and spicy fried things, adorned with flowers and incense and lighted by oil lamps, surround the open-air arena. The clown, an essential character in every folk play, comments on the audience and contemporary events. Zealous spectators offer donations and gifts in appreciation of their favourite actor or dancer, who receives them in the middle of the performance and thanks the donor by singing or dancing a particular piece of his choice. The audience thus constantly throws sparks to the performer, who throws them back. People laugh, weep, sigh, or suddenly fall silent during a moving scene.

In both folk and classical forms of drama, the performer may lengthen or shorten his piece according to audience response. During a *kathak* dance, the drummer, in order to test the perfection of the dancer, disguises the main beat of his drum by slurs and offbeats, a secret he shares with the audience and announces by a loud thump that is synchronized with the dancer's stamping of the foot. At this point in the dance the spectators shout, swaying their heads in admiration. They show their approval and disapproval through delighted groans or sullen headshakes as the performance goes on. In the *rāslīlā,* the audience joins in singing the refrain and marks the beat by hand clapping. At a climactic point the people rock and sway, rhythmically clapping and singing. These practices bind the performers, chanters, and spectators together in a sense of aesthetic pleasure.

Instrumental music and singing

Instrumental music and singing are integral parts of Indian dance and theatre. Musicians, chanters, and drummers sit on the stage in view, a tradition observed throughout almost all of Asia. They watch the dancer and play on their instruments following his movements, whereas in the West the movements of a ballerina are timed and controlled by the already-written music. An Indian dancer is constantly reacting to the accompanying musician, and vice versa. He may signal the chanters and drummers and even instruct them during the performance without spoiling its aesthetic effect.

In some classical dance forms, such as *kuchipudi,* the dancer sings in voiceless whispers as she dances. In *bhārata-nāṭyam* the dance movements are like sculpted music in space, and the accompanying musician is invariably a dance guru (teacher). In *kathak* the rhythmic syllables beaten out by the dancer with her feet are vocalized by the singer and then chirped out by the drummer. No folk dancing is complete without the use of drum and vocal singing. Women's folk singing such as the *giddha* in the Punjab and the men's *kīrtan* in West Bengal takes the form of dance when the rhythm becomes fast.

In folk theatre this relationship is even more apparent. *Rāslīlā* dance sequences are interspersed with the singing as a decorative frill, to accentuate emotional appeal, or to mark the climax of a song. The *yakṣagāna* hero gives a brisk dance number to announce his entry. In many folk forms of opera (*bhavai, terukkūttu,* and *nautanki*), the characters sing and dance at the same time or alternate. Ballad singers from the states of Orissa and Andhra Pradesh dramatize their singing by strong facial gestures and rhythmic ankle bells and execute dance phrases between the narrative singing. On the other hand, no one can imagine a dancer who is not at the same time a musician. This double aesthetic discipline enriches both of these arts, and the Indian audience is conditioned to this tradition.

INDIAN DANCE

Dance in India can be organized into three categories: classical, folk, and modern. Classical dance forms are among the best preserved and oldest practiced in the 20th century. The royal courts, the temples, and the guru to pupil teaching tradition have kept this art alive and unchanged. Folk dancing has remained in rural areas as an expression of the daily work and rituals of village communities. Modern Indian dance, a product of the 20th century, is a creative mixture of the first two forms, with freely improvised movements and rhythms to express the new themes and impulses of contemporary India.

The popularity of dance in 20th-century India can be judged from the fact that there is hardly any Indian motion picture that does not have half a dozen dances in it. In the typical "boy meets girl" film the heroine dances everywhere and anywhere. A film company may not have a script writer (in some cases the financier writes the story himself), but it must have a dance director. To provide ample dance opportunities, motion pictures have been made on the lives of poets, courtesans, and temple dancers and on mythological themes. For these the services of expert dancers are sought.

In the 20th century, classical dance has left the temples and royal courts and is now presented regularly on the stage in large cities. Rich industrialists, international hotels, and the wealthy families of the upper class are the chief patrons. It is not uncommon to have a classical dance recital by a major performer at a business dinner or for the annual function of a club. Some universities have dance as a regular subject in their curricula. Women learn it as a social grace, and young girls learn a few classical dances for greater eligibility in marriage. Folk dancing has also become more common as a contemporary cultural event in the cities. Most colleges have their folk-dance troupes, and even the police of the Punjab have their folk-dance groups to perform the *bhangra.* Folk dance, cut off from its rural settings, has lost much of its original vigour and beauty, but that is the inevitable result of cross-fertilization of regional cultures through folk-troupe exchanges at an interstate level.

Classical dance. *The dance-drama.* India has evolved through its classical and folk traditions a type of dance drama that is a form of total theatre. The actor dances out the story through a complex gesture language, a form that, in its universal appeal, cuts across the multilanguage barrier of the subcontinent. Some of the classical dance-drama forms (*e.g., kathākali, kuchipudi, bhagavatha mela*) enact well-known stories derived from Hindu mythology. The 20th-century dancers Uday Shankar and Shanti Bardhan have created ballets that were inspired by such traditional dance-dramas. Contemporary Indian directors and writers are re-examining traditional dance forms and are using these in their current works for greater psychological appeal and deeper artistic impact. Millions in villages are still entertained by dance-dramas. In spite of the popu-

'he
'āṭya-
āstra

larity of straight prose plays in the cities, the appeal of dance-drama is unquestionably deeper and more satisfying to the rural Indian, whose aesthetics is still rooted in tradition.

The chief source of classical dance is Bharata Muni's *Nāṭya-śāstra* (1st century BC to 1st century AD), a comprehensive treatise on the origin and function of *nāṭya* (dramatic art that is also dance), on types of plays, gesture language, acting, miming, theatre architecture, production, makeup, costumes, masks, and various *bhāvas* ("emotions") and *rasas* ("sentiments"). No other book of ancient times contains such an exhaustive study of dramaturgy.

Techniques and types of classical dance. According to the *Nāṭya-śāstra,* the dancer-actor communicates the meaning of a play through four kinds of *abhinaya* (histrionic representations): *angika,* transmitting emotion through the stylized movements of parts of the body; *vācika,* speech, song, pitch of vowels, and intonation; *āhārya,* costumes and makeup; and *sāttvika,* the entire psychological resources of the dancer-actor.

The actor is equipped with a complicated repertoire of stylized gestures. Conventionalized movements are prescribed for every part of the body, the eyes and hands being the most important. There are 13 movements of the head, seven of the eyebrows, six for the nose, six for the cheek, seven for the chin, nine for the neck, five for the breasts, and 36 for the eyes. There are 32 movements of feet, 16 on the ground and 16 in the air. Various positions of the feet (strutting, mincing, tromping, splaying, beating, etc.) are carefully worked out. There are 24 single-hand gestures (*asaṃyuta-hasta*) and 13 for combined hands (*saṃyuta-hasta*). One gesture (*hasta*) may mean over 30 different things quite unrelated to each other. The *patāka* gesture of the hand, for example, in which all the fingers are extended and held close together with the thumb bent, can represent heat, rain, a crowd of men, the night, a forest, a horse, or a flight of birds. The *patāka* hand with the third finger bent (*tripatāka*) can mean a crown, a tree, marriage, fire, a door, or a king. In *karkaṭa* ("crab"), one of the combined hand gestures, the fingers of the hands are interlocked, and this may indicate a honeycomb, yawning after sleep, or a conch shell. Of course, for each of these different meanings a *hasta* is given a different body posture or action.

The male or female classical dancer portraying a story in a solo performance simultaneously plays two or three principal characters by alternating facial expressions, gestures, and moods. Krishna, his jealous wife Satyabhāmā, and his gentle wife Rukmiṇī, for example, may be played by one person.

āva
d
sa

The aesthetic pleasure of Hindu dance and theatre is determined by how successful the artist is in expressing a particular emotion (*bhāva*) and evoking the *rasa.* Literally *rasa* means "taste" or "flavour" and is that exalted sentiment or mood that the spectator experiences after witnessing a performance. The critics do not generally concern themselves so much about plot construction or technical perfection of a poem or play as about the *rasa* of a particular work. There are nine *rasa:* erotic, comic, pathetic, furious, heroic, terrible, odious, marvellous, and spiritually peaceful. There are nine corresponding *bhāvas:* love, laughter, pathos, anger, energy, fear, disgust, wonder, and quietude.

Four distinct schools of classical Indian dance—*bhārata-nāṭya, kathākali, kathak,* and *manipuri*—exist in the 20th century, along with two types of temperament—*tāṇḍava,* representing the fearful male energy of Śiva, and *lāsya,* representing the lyrical grace of Śiva's wife Pārvatī. *Bhārata-nāṭya,* which takes its name from Bharata's *Nāṭya-śāstra,* has the *lāsya* character, and its home is Tamil Nadu, in South India. *Kathākali,* a pantomimic dance-drama in the *tāṇḍava* mood with towering headgear and elaborate facial makeup, originated in Kerala. *Kathak* is a mixture of *lāsya* and *tāṇḍava* characterized by intricate footwork and mathematical precision of rhythmic patterns; it flourishes in the north. *Manipuri,* with its swaying and gliding movements, is *lāsya,* and it has been preserved in Manipur state in the Assam Hills. In 1958 the Sangeet Natak Akademi (National Academy of Music, Dance and Drama) in New Delhi bestowed classical status on two other schools of dance—*kuchipudi,* from Andhra Pradesh, and *Odissi,* from Orissa. These two styles overlap the *bhārata-nāṭya* school and therefore are not as distinctly different in temperament and style as other forms.

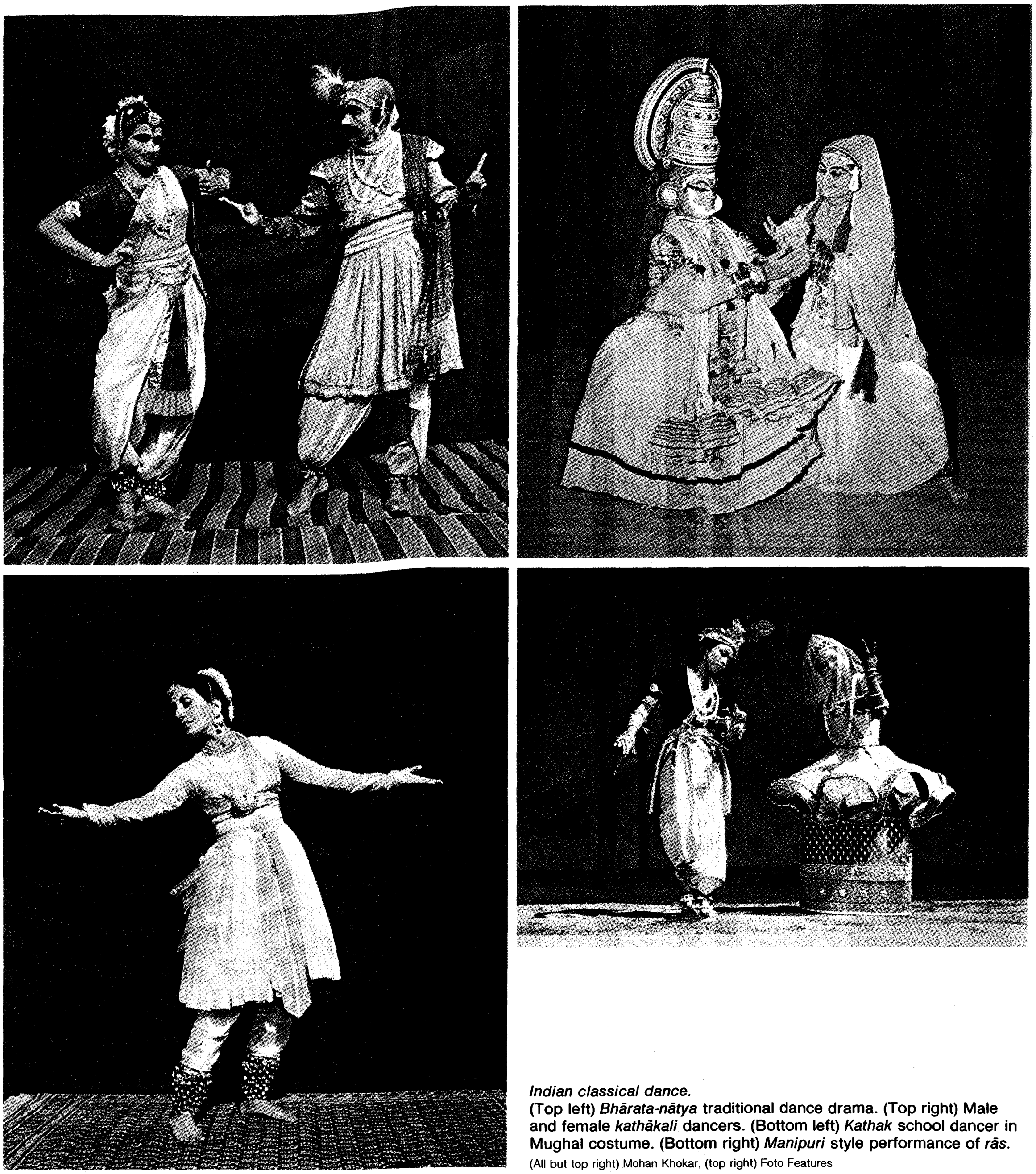
The bhārata-nāṭya school. *Bhārata-nāṭya* (also called *dasi attam*) has survived to the present through the *devadāsīs,* temple dancing girls who devoted their lives to their gods through this medium. Muslim invasions from the north destroyed the powerful Hindu kingdoms in the south but could not disrupt their arts, which took shelter in the temples. After the 16th century the Muslims overpowered the south completely until the British came, thus giving a setback to Hindu dance. Slowly the institution of *devadāsī* fell into disrepute, and temple dancing girls became synonymous with prostitutes. In the latter half of the 19th century in Tanjore, Chinniyah, Punniah, Vadivelu, and Shivanandam, four talented dancers who were brothers, revived the original purity of *dasi attam* by studying and following the ancient texts and temple friezes, with missing links supplied by the socially spurned *devadāsīs.* Their popularized form of *dasi attam* was called *bhārata-nāṭya.*

A performance of *bhārata-nāṭya* lasts for about two hours and consists of six parts, beginning with *allarippu* (Telegu language, "to decorate with flowers"), a devotional prologue that shows off the elegance and grace of the dancer. The second part is *jātisvaram,* a brilliant blaze of *jāti*s ("dance phrases") with *svara*s ("musical sounds"). This is followed by *shabdam,* the singing words that prepare the dancer to interpret through *abhinaya* (gesture language) interspersed with pure dance. The fourth part is *varṇam,* a combination of expressive and pure dance. Then follow the *padam*s, songs in Telegu, Tamil, or Kanarese that the dancer dramatizes by facial expressions and hand gestures. The accompanying singer chants the line again and again, and the dancer enacts the clashing and contrasting meanings. Her virtuosity consists of exhausting all possible shades of suggestion. The performance ends with *tillānā,* a pure dance accompanied by meaningless musical syllables chanted to punctuate the rhythm. The dancer explodes into leaps and jumps forward and backward, from right and left, in a state of ecstasy. *Tillānā* ends with three clangs of the cymbals while the dancer executes a triple blaze of *jāti*s, thumping her feet with a jingling flourish of ankle bells.

Bhārata-nāṭya has attained world recognition as one of the most exquisite forms of classical dance. Its aspirants go to Tamil Nadu to learn from gurus who still live in villages. Because of its *lāsya* character, performing artists have always been women. But their teachers have invariably been old men who chant the lines to tiny cymbals, controlling the complex rhythm without dancing themselves.

The major 20th-century performers associated with the *bhārata-nāṭya* school of dance are T. Balasaraswathi, especially known for her *abhinaya* (expressive interpretation) of *padams*; Rukmini Devi, who popularized *bhārata-nāṭya* among the upper classes in the 1930s; Yamini Krishnamurthi; and Shanta Rao. Two of the most important gurus were Minakshisundaram Pillai, who injected vigour into *bhārata-nāṭya* by his choreography, and his son-in-law, Chokkalingam Pillai.

The kathākali school. *Kathākali* (*kathā,* "story"; *kali,* "performance") originated in the 17th century in Kerala, the lush tropical coastal strip of South India washed by the Arabian Sea. It was devised by the Raja of Kottarakkara, who, angry over the refusal of a neighbouring prince to allow his dancers to perform a Sanskrit dance-drama in his court, decided to create his own dance troupe using Malayalam, the spoken language of the people. This school of dance has its own *hasta*s, based on a regional text influenced by the *Nāṭya-śāstra* and later treatises. It also has marked elements of energetic ritualistic dances. The makeup has its roots in the grotesque Dravidian demon masks of the pre-Hindu period. Themes are taken mainly from the *Rāmāyaṇa,* the *Śiva-Purāṇa,* the *Bhāgavata-Purāṇa,* the *Mahābhārata,* and other religious texts. The superhuman characters represent primal forces of good and evil at war. Because of its terrifying vigour, men play all the roles.



Indian classical dance.
(Top left) *Bhārata-nātya* traditional dance drama. (Top right) Male and female *kathākali* dancers. (Bottom left) *Kathak* school dancer in Mughal costume. (Bottom right) *Manipuri* style performance of *rās*.
(All but top right) Mohan Khokar, (top right) Foto Features

Characters and costumes

Most *kathākali* characters (except those of women, Brahmins, and sages) wear towering headgear and billowing skirts and have their fingers fitted with long silver nails to accentuate hand gestures. The principal characters are classified into seven types. (1) *Pacca* ("green") is the noble hero whose face is painted bright green and framed in a white bow-shaped sweep from ears to chin. Heroes such as Rāma, Lakṣmaṇa, Krishna, Arjuna, and Yudhiṣṭhira fall into this category. (2) *Katti* ("knife"), haughty and arrogant but learned and of exalted character, has a fiery upcurled moustache with silver piping and a white mushroom knob at the tip of his nose. Two walrus tusks protrude from the corners of his mouth, his headgear is opulent, and his skirt is full. Duryodhana, Rāvaṇa, and Kichaka belong to this type. (3) *Chokannatadi* ("red beard"), power-drunk and vicious, is painted jet black from the nostrils upward. On both cheeks semicircular strips of white paper run from the upper lip to the eyes. He has black lips, white warts on nose and forehead, two long curved teeth, spiky silver claws, and a blood-red beard. (4) *Velupputadi* ("white beard") represents Hanuman, son of the wind god. The upper half of his face is black and the lower red, marked by a tracery of curling white lines. The lips are black, the nose is green, black squares frame the eyes, and two red spots decorate the forehead. A feathery gray beard, a large furry coat, and bell-shaped headgear give the illusion of a monkey. (5) *Karupputadi* ("black beard") is a hunter or forest dweller. His face is coal black with crisscross lines drawn around the eyes. A white flower sits on his nose, and peacock feathers closely woven into a cylinder rise above his head. He carries a bow, quiver, and sword. (6) *Kari* ("black") is intended to be disgusting and gruesome. Witches and ogresses, who fall into this category, have black faces marked with queer patterns in white and huge, bulging breasts. (7) *Minnukku* ("softly shaded") represents sages, Brahmins, and women. The men wear white or orange dhotis. Women have their faces painted light yellow and sprinkled with mica, and their heads are covered by saris.

Under a flower-decked canopy on a square, ground-level stage a tall brass worship lamp brimming with coconut oil burns brightly. The musicians and dancers bow before it before they start performing. Drummers standing in one corner pound the *cenda,* a barrel-shaped drum with a piercing, clattering sound suited for battle scenes, and continue throughout the performance, almost without respite.

Two men hold a 12-foot by six-foot (four-metre by two-metre) embroidered hand curtain from behind which the principal characters make their entrances. They dance, grab the trembling curtain, and give vivid facial expressions with fearful glances and grunts. This "peering over the curtain," called *tiranokku,* is a close-up that offers an actor full scope to display his art. At a climactic moment the curtain is whisked away and the character enters in full splendour. The performance lasts all night, the singers singing the text that the dancers act out in an elaborate gesture language.

Well-known performers of *kathākali* include Guru Chandu Panikkar, Guru Kunju Kurup, Ramunni Nair, and Kalamandalam Krishna Nair. The dancers Guru Gopi Nath and Krishnan Kutty have both emphasized simplification of the use of towering headgear and thick-crusted, elaborate makeup, so that the art may be more commonly understood.

The kathak school. Kathak, born of the marriage of Hindu and Muslim cultures, flourished in North India under Mughal influence. *Kathak* dancers retain their 17th-century costumes but are steeped in Rādhā and Krishna love lore. Krishna, playing his flute in the Vṛndāvana woods on the bank of the Yamuna River, is surrounded by the *gopīs* ("milkmaids"). Their play is the eternal game of the god and his devotees, the hide-and-seek of man and woman. This spiritual relationship is deeply passionate, with erotic love-play. Slowly the dance degenerated and found shelter in bawdy houses, where nautch girls practiced the art to make themselves more tantalizing. In the beginning of the 20th century it was reclaimed and revived, however, mainly through the efforts of Kalkaprasad Maharaj, whose three sons Achchan, Lachchu, and Shambhu, perfected the art.

Because of its mixed *lāsya* and *tāṇḍava* temperament, *kathak* is popular with both females and males. In *bhāratanāṭya,* footwork is synchronized with hand gestures and eye movements, but *kathak* has no such rigid technique. It takes its movements from life, stylizes them, and adds complex rhythmic patterns. The mathematical precision in doubling and quadrupling the beat with quick transfers and shifts makes the onlookers dizzy.

Kathak performers

A female *kathak* dancer generally wears a brocade blouse, a long, wide, shimmering silk skirt, a transparent tissue scarf of gold threads, and a heavy cluster of ankle bells. A musician, generally the guru, sits beside the drummer on the floor and vocalizes the complicated syllables of the drum that the dancer beats out with her feet. *Kathak's* basic dance posture and some of the steps can be traced to the *rāslīlā* of Braj Bhoomi. The musical refrain, which is called *lehra,* provides the base on which the drummer and the dancer execute a rich tapestry of rhythmic patterns. Beats are called *mātrās* and the footwork *tatkar.* Important elements of the dance are *chakkars, torahs,* and *tihais. Chakkar* denotes whirling with great speed and stopping for a fraction of time after each whirl within the prescribed beat while at the same time maintaining the beauty of the form. *Torah* is a composition consisting of rhythmic syllables. *Tihai* is the repetition of a phrase of rhythmic syllables used to adorn the concluding part of a *torah.* There are two styles of *kathak:* Jaipur *gharana* and Lucknow *gharana.* While the Lucknow *gharana* excels in *bhāva,* the Jaipur *gharana* specializes in brilliance of footwork.

In the 20th century the major performers of *kathak* include Shambhu Maharaj, who specialized in *bhavapradarśan* ("display of emotion"), and Sunder Prasad, who concentrated on the *tala* and *layakari* aspects of the dance. Birju Maharaj, Gopi Krishan, Sitara Devi, and Damayanti Joshi all have important reputations in India as well as abroad.

The manipuri school. Manipuri has survived in the sheltered valley of Manipur in the Assam Hills. It remained aloof not only from foreign influences but also from the main Indian trends. Its isolation was broken only in the 1920s, when Rabindranath Tagore visited the valley and invited a leading guru of the area, Atomba Singh, to teach at his school in Santiniketan. The supple movements of *manipuri* dance were suitable for Tagore's lyrical dramas, and he therefore employed them in his plays and introduced the dance as a part of the curriculum at his institution.

The *manipuri* dancer wears a large, stiff skirt that is glittering with round mirror pieces and a shimmering gauze veil. Her hair is done up in a high rolled crown that is adorned with chains of white blossoms, and her luminous cheeks and forehead are decorated with dots of sandalwood paste.

Style of the *manipuri* dance

Known for its femininity, *manipuri* is marked by a slow, swooning rhythm. The dancer, with her hips thrust back and head tilted on one side, turns and sways and glides as if in a dream. The immobility of her face, like that of a mask, is in sharp contrast with the other three schools of dance, in which the face and eyes are a major source of expression.

The *manipuri* drummer, his naked torso in a white dhoti with a red border tucked up above his knees, dances while he plays on the drum. He slaps and thumps; the drum rumbles and howls and chuckles. Drunk with its rhythm, the drummer dances in wild, frenzied leaps. His energetic and electric movements are a masculine counterpart to the slow, undulating patterns woven by the female dancer.

Chief 20th-century exponents of *manipuri* include Atomba Singh, who preserved the tradition of *rās* dancing, and Amubi Singh.

The kuchipudi school. Kuchipudi dance-dramas owe their origin to the small village of Kuchipudi (Kuchelapuram) in Andhra Pradesh. Their form was originated in the 17th century by Sidhyendra Yogi, creator of the superb dance-drama *Bhama Kalapam,* which is the story of charming Satyabhāma, jealous wife of Lord Krishna. Sidhyendra Yogi taught the art to Brahmin boys of Kuchipudi and gave a performance with them in 1675 for the Nawab of Golconda, who was so pleased that he granted Kuchipudi to the Brahmin Bhavathas for the preservation of this art. Even in the 20th century every Brahmin of Kuchipudi is expected to perform at least once in his life the role of Satyabhāma as an offering to Lord Krishna.

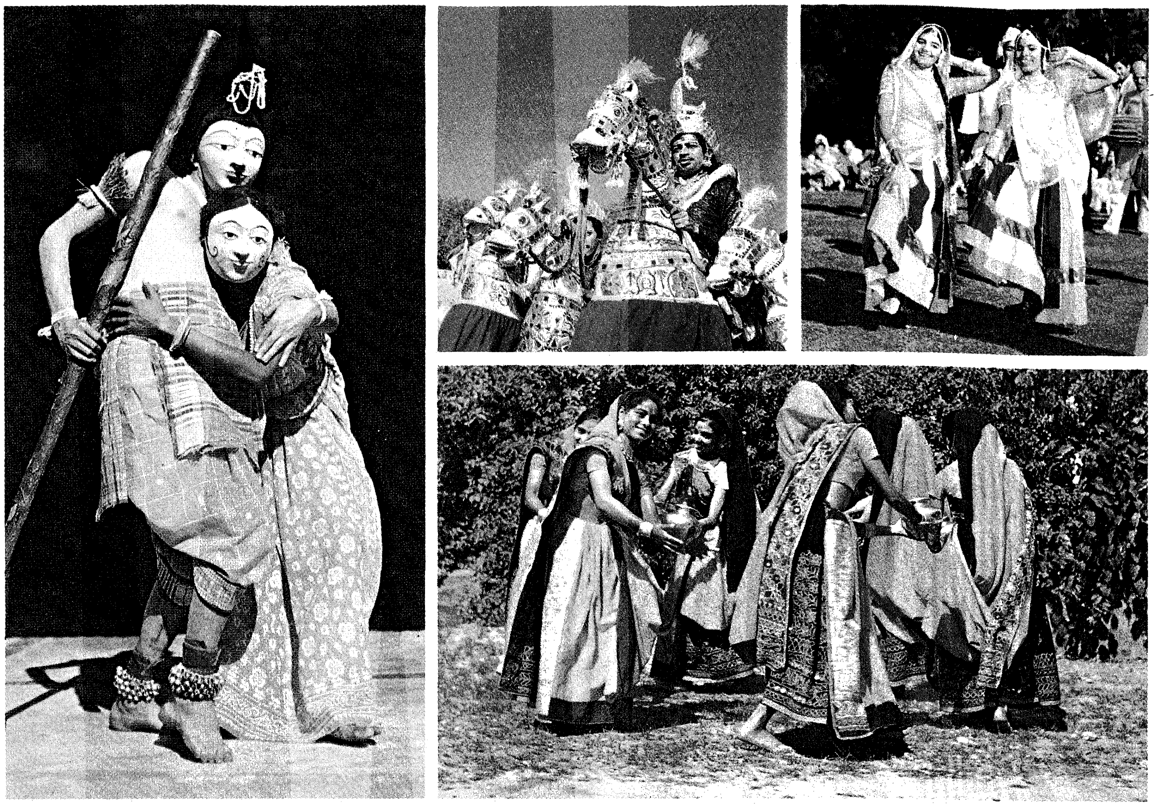
The *kuchipudi* dance begins with worship rituals. A male dancer moves about sprinkling holy water, and then incense is burned. *Indra-dhvaja* (the flagstaff of the god Indra) is planted on the stage to guard the performance against outside interference. Women sing and dance with worship lamps, followed by the worship of Gaṇeśa, the elephant god, who is traditionally petitioned for success before all enterprises. The *bhagavatha* (stage manager-singer) sings invocations to the goddesses Sarasvatī (Learning), Lakṣmī (Wealth), and Paraśakti (Parent Energy), in between chanting drum syllables.

Two men hold up the traditional coloured curtain. A long gold-embroidered braid is hung on the curtain as a challenge to anyone among the spectators who dares to act and dance. If anyone should take up this braid, the hero playing the female character Satyabhāma will cut off "her" hair. The principal characters are introduced from behind the curtain after each one has done a brisk dance, and at that time the *bhagavatha* sings out the background and function of each. All roles are traditionally played by men (but in recent times by women also), and all the four elements of *abhinaya* are used—dance, song, costume, and psychological resources. Thus, *kuchipudi* differs from other classical dances in which the performers do not sing.

Among the major *kuchipudi* dancers of the 20th century are Guru Chinta Krishnamurthi, Vedantam Satyanarayana, and Yamini Krishnamurthi.

Odissi. Odissi, practiced in Orissa, claims to be over 2,000 years old and the true inheritor of the *Nāṭya-śāstra* tradition. It originated and was initially developed in the temples and later flourished in the courts as well. Many of the 108 basic dance units (*karaṇas*) mentioned in the *Nāṭya-śāstra* can be found only in *odissi,* and many of its dance poses are sculpted on the exterior of the temples of Bhubaneswar, Konārak, and Purī. Kelu Charan Mahapatra and Indrani Rehman are the principal 20th-century figures associated with *odissi.*

Other classical dance forms. Among other classical or



Indian folk dance.
(Left) *Chhau* dance of Bihār showing boatman and wife. (Centre) *Kacchi ghori* dancers of Rājasthān. (Top right) *Ghoomar* dancers of Rājasthān. (Bottom) *Garabā* dancers of Gujarāt.
(Left, bottom) Mohan Khokar, (centre, top right) Foto Features

Bhagavatha mela, mohini attam, and kuṟavañci

semiclassical dance forms are *bhagavatha mela, mohini attam,* and *kuṟavañci.* Performed at the annual Narasimha Jayanti festival in Melatur village in Tamil Nadu, the *bhagavatha mela* uses classical gesture language with densely textured Karnatak music. Its repertoire was enriched by the musician-poet Venkatarama Sastri (1759–1847), who composed important dance-dramas in the Telugu language. *Mohini attam* is based on the legend of the Hindu mythological seductress Mohinī, who tempted Śiva. It is patterned on *bhārata-nāṭya* with elements of *kathākali.* It uses Malayalam songs with Karnatak music. *Kuṟavañci* is a dance-drama of lyrical beauty prevalent in Tamil Nadu. It is performed by four to eight women, with a gypsy fortune-teller as initiator of the story of a lady pining for her lover. Formally, it is a mixture of the folk and classical types of Indian dance.

Folk dance. Indian folk dances have an inexhaustible variety of forms and rhythms. They differ according to region, occupation, and caste. The half-naked Adivasis (aboriginal tribes) of central and eastern India (Murias, Bhīls, Gonds, Juangs, and Santāls) are the most uninhibited in their dancing. There is hardly a national fair or festival where these dances are not performed. The most impressive occasion occurs every January 26 on Republic Day, when dancers from all parts of India come to New Delhi to dance in the vast arena of the National Stadium and along a five-mile parade route.

Types of folk dances

It is difficult to categorize Indian folk dances, but generally they fall into four groups: social (concerned with such labours as tilling, sowing, fishing, and hunting); religious; ritualistic (to propitiate an angry goddess or demon with magical rites); masked (a type that appears in all the above categories).

The *kolyacha* is among the better known examples of social folk dance. A fisherman's dance indigenous to the Konkan coast of western central India, the *kolyacha* is an enactment of the rowing of a boat. Women wave handkerchiefs to their male partners, who move with sliding steps. For wedding parties young Kolis dance in the streets carrying household utensils for the newlywed couple, who join the dance at its climax.

The national social folk dance of Rājasthān is the *ghoomar,* danced by women in long full skirts and colourful *chuneri*s (squares of cloth draping head and shoulders and tucked in front at the waist). Especially spectacular are the *kacchi ghori* dancers of this region. Equipped with shields and long swords, the upper part of their bodies each arrayed in the traditional attire of a bridegroom and the lower part concealed by a brilliant-coloured papier-mâché horse built up on a bamboo frame, they enact jousting contests at marriages and festivals. Bawaris, by tradition a criminal tribe, generally are expert in this form of folk dance.

In the Punjab, the most electrifying social folk dance is the male harvest dance, *bhangra,* which is also popular in the Punjab province of Pakistan. This dance is always punctuated by a song. At the end of every line the drum thunders. The last line is taken up by all the dancers in a chorus. In ecstasy they spring, bellow, shout, and gallop in a circle, madly wiggling their shoulders and hips. Any man of any age can join.

The Lambadi Gypsy women of Andhra Pradesh wear mirror-speckled headdresses and skirts and cover their arms with broad, white bone bracelets. They dance in slow, swaying movements, with men acting as singers and drummers. Their social dance is imbued with impassioned grace and lyricism and is less wild than that of Gypsies in other parts of the world.

The bison-horn dance of the Muria tribe in Madhya Pradesh is performed by both men and women, who traditionally have lived on equal terms. The men wear a horned headdress with a tall tuft of feathers and a fringe of cowry shells dangling over their faces. A drum shaped like a log is slung around their necks. The women, their heads surmounted by broad, solid-brass chaplets and their breasts covered with heavy metal necklaces, carry sticks in their right hands like drum majorettes. Fifty to 100 men and women dance at a time. The male "bisons" attack and fight each other, spearing up leaves with their horns and chasing the female dancers in a dynamic interpretation of nature's mating season.

The Juang tribe in Orissa performs bird and animal dances with vivid miming and powerful muscular agility.

Religious folk dance

Some major examples of religious folk dances are the *dindi* and *kala* dances of Mahārāshtra, which are expressions of religious ecstasy. The dancers revolve in a circle, beating short sticks (*dindi*s) to keep time with the chorus leader and a drummer in the middle. As the rhythm accelerates, the dancers form into two rows, stamp their right feet, bow, and advance with their left feet, making geometric formations. The *kala* dance features a pot symbolizing fecundity. A group of dancers forms a double-tiered circle with other dancers on their shoulders. On top of this tier a man breaks the pot and splashes curds over the naked torsos of the dancers. After this ceremonial opening, the dancers twirl sticks and swords in a feverish battle dance.

Garabā, meaning a votive pot, is the best known religious dance of Gujarāt. It is danced by a group of 50 to 100 women every year for nine nights in honour of the goddess Ambā Mātā, known in other parts of India as Durgā or Kālī. The women move in a circle bending, turning, clapping their hands, and sometimes snapping their fingers. Songs in praise of the goddess accompany this dance.

Of the endless variety of ritualistic folk dances, many have magical significance and are connected with ancient cults. The *karakam* dance of Tamil Nadu state, mainly performed on the annual festival in front of the image of Māriyammai (goddess of pestilence), is to deter her from unleashing an epidemic. Tumbling and leaping, the dancer retains on his head without touching it a pot of uncooked rice surmounted by a tall bamboo frame. People ascribe this feat to the spirit of the deity, which, it is believed, enters his body. The Therayattam festival in Kerala is held to propitiate the gods and demons recognized by the pantheon of the Malayalis. The dancers, arrayed in awe-inspiring costumes and hideous masks, enact weird rituals before the village shrine. A devotee makes an offering of a cock. The dancer grabs it, chops off its head in one stroke, gives a blessing, and hands the bloody gift back to the devotee. This ceremony is punctuated by a prolonged and ponderous dance.

Masked folk dances

The greatest number of masked folk dances are found in Arunachal Pradesh (formerly North East Frontier Agency) union territory of India, where the influence of Tibetan dance may be seen. The yak dance is performed in the Ladākh section of Kashmir and in the southern fringes of the Himalayas near Assam. The dancer impersonating a yak dances with a man mounted on his back. In *sada topo tsen* men wear gorgeous silks, brocades, and long tunics with wide flapping sleeves. Skulls arranged as a diadem are a prominent feature of their grotesquely grinning wooden masks representing spirits of the other world. The dancers rely on powerful, rather slow, twirling movements with hops. The *chhau,* a unique form of masked dance, is preserved by the royal family of the former state of Saraikela in Bihār. The dancer impersonates a god, animal, bird, hunter, rainbow, night, or flower. He acts out a short theme and performs a series of vignettes at the annual Chaitra Parva festival in April. *Chhau* masks have predominantly human features slightly modified to suggest what they are portraying. With serene expressions painted in simple, flat colours, they differ radically from the elaborate facial makeup of *kathākali* or the exaggerated ghoulishness of the Nō and Kandyan masks. His face being expressionless, the *chhau* dancer's body communicates the total emotional and psychological tensions of a character. His feet have a gesture language; his toes are agile, functional, and expressive, like those of an animal. The dancer is mute; no song is sung. Only instrumental music accompanies him. In another form of *chhau,* practiced in the Mayūrbhanj district of Orissa, the actors do not wear masks, but through deliberately stiff and immobile faces they give the illusion of a mask. The style of their dance is vigorous and acrobatic.

Modern Indian dance. While in the West the theatrical elements of spoken words, music, and dance developed independently and evolved in the forms of drama, opera, and ballet, Indian theatrical tradition continued to combine the three in its dramas. Indian films still follow this rule (the heroine suddenly bursts into a song or dances for the hero), which offends Western sensibility, but in fact they are following their own classical and folk tradition. Recently, dance in the form of ballet with complex choreography in the Western sense has emerged as a distinct form.

Modern ballet

Modern Indian ballet started with Uday Shankar, who went to England to study the plastic arts and was chosen by the Russian ballerina Anna Pavlova to be her partner in the ballet *Radha and Krishna.* Young Shankar returned to India fired with enthusiasm. After studying the essentials of the four major styles of classical dance, he created new ballets with complex choreography and music, mixing the sounds from wooden clappers and metal cymbals with those of traditional instruments. He used classical and folk rhythms. Employing Western stage techniques, he presented his ballets with a skill and polish previously unknown to Indian audiences. These ballets included *Shiva-Parvati* and *Lanka Dahan* ("The Burning of Lanka"), in which he used wooden masks from Ceylon. In *Rhythm of Life* (1938) and in *Labour and Machinery* (1939), he employed contemporary political and social themes. He established a culture centre at Almora in 1939 and during its four years' existence created a whole generation of modern dancers.

Shanti Bardhan, a junior colleague of Uday Shankar, produced some of the most imaginative dance-dramas of the modern period. After founding the Little Ballet Troupe in Andheri, Bombay, in 1952 he produced *Ramayana,* in which the actors moved and danced like puppets. His posthumous production *Panchatantra (The Winning of Friends)* is based on an ancient fable of four friends (Mouse, Turtle, Deer, and Crow), in which he used masks and the mimed movements of animals and birds.

Mohan Khokar

"Ramayana," puppet-style modern dance-drama originally produced and choreographed by Shanti Bardhan, *c.* 1952.

Narendra Sharma and Sachin Shankar, both pupils of Uday Shankar, have continued his tradition. Other important figures who have shaped modern Indian dance include Menaka, Ram Gopal, and Mrinalini Sarabhai, who has experimented with conveying modern themes through the *bhārata-nāṭya* and *kathākali* styles.

Dance-training centres. Dance training in small academies and local *kala kendras* ("art centres") is available all over contemporary India. Most universities have introduced dance as a subject in their curricula. The gurus still impart specialized training to pupils who go to live with them in villages and learn the art over a number of years. But there are many state-run or public-financed training centres organized in the 20th century that attract students from all over the world. Among the most important of these are Kerala Kalamandalam (Kerala Institute of Arts), near Shoranūr; Kalakshetra at Adyar, Tamil Nadu; Kathak Kendra, a dance branch of the Bharatiya Kala Kendra in New Delhi; Triveni Kala Sangam (Centre of Music, Dance, and Painting), at New Delhi; Darpana Academy in Ahmadābād, Gujarāt; Visva-Bharati (founded by Rabindranath Tagore), at Santiniketan, West Bengal;

and the Jawaharlal Nehru Manipuri Dance Academy, at Imphāl.

INDIAN THEATRE

Classical theatre. Classical Sanskrit theatre flourished during the first nine centuries of the Christian era. Aphorisms on acting appear in the writings of Pāṇini, the Sanskrit grammarian of the 5th century BC, and references to actors, dancers, mummers, theatrical companies, and academies are found in Kauṭilya's book on statesmanship, the *Artha-śāstra* (4th century BC). But classical structure, form, and style of acting and production with aesthetic rules were consolidated in Bharata Muni's treatise on dramaturgy, *Nāṭya-śāstra.* Bharata defines drama as a

Bharata Muni's rules

> mimicry of the actions and conduct of people, rich in various emotions, depicting different situations. This relates to actions of men good, bad and indifferent and gives courage, amusement, happiness, and advice to all of them.

Bharata classified drama into ten types. The two most important are *nāṭaka* ("heroic"), which deals with the exalted themes of gods and kings and draws from history or mythology (Kālidāsa's *Śakuntalā* and Bhavabhūti's *Uttararāmacarita* fall into this caterory), and *prakaraṇa* ("social"), in which the dramatist invents a plot dealing with ordinary human beings, such as a courtesan or a woman of low morals (Śūdraka's *Mṛcchakaṭika,* "The Little Clay Cart," belongs to this type). Plays range from one to ten acts. There are many types of one-act plays, including *bhaṇa* ("monologue"), in which a single character carries on a dialogue with an invisible one, and *prahasana* ("farce"), which is classified into two categories: superior and inferior, both dealing with courtesans and crooks. King Mahendravikramavarman's 7th-century-AD *Bhagavad-Ajjukiya* ("The Harlot and the Monk") and *Mattavilāsa* ("Drunken Revelry") are examples of *prahasana.*

Types of classical theatre

There are three structural types of classical theatre: oblong, square, and triangular, each further divided into large, medium, and small sizes. According to the *Nāṭya-śāstra,* the playhouse was "like a mountain cave" with two floors at different levels, small windows so that outside noise and wind would not interfere with the acoustics, and a backstage for actors to do makeup, costumes, and offstage noise effects. Bharata disapproved of a large playhouse and recommended the medium-size structure meant for court productions.

The ancient Hindus insisted on a small playhouse, because dramas were acted in a highly stylized gesture language with subtle movements of eyes and hands. Hindu theatre differed from its Greek counterpart in temperament and method of production. The three unities rigidly followed by the Greeks were totally unknown to Sanskrit dramatists. Less time was consumed by a Greek program of three tragedies and a farce than by a single Sanskrit drama, with its subsidiary plots and wide variety of characters and moods. The Greeks laid emphasis on plot and speech, the Hindus on the four types of acting and visual demonstration. People were audiences to the Greeks and spectators to the Hindus. The aesthetic rules also differed. Aristotle's theory of catharsis bears no resemblance to Bharata's theory of *rasa.* The Greek conception of tragedy is totally absent in Sanskrit dramas, as is the aesthetic principle that prohibits any death or defeat of the hero on stage.

Kinds of Hindu production

There were two types of Hindu productions: the *lokadharmī,* or realistic theatre, with natural presentation of human behaviour and properties catering to the popular taste, and the *nāṭyadharmī,* or stylized drama, which, using gesture language and symbols, was considered more artistic. In *Śakuntalā* the king enters riding an imaginary chariot, and Śakuntalā plucks flowers that are not there; in "The Little Clay Cart" the thief breaks through a nonexistent wall, and Maitreya passes through Vasantasena's seven courtyards by miming.

A classical play traditionally opened with the *nāndī,* a benediction of eight to 12 lines of verse in praise of the gods, after which the *sūtra-dhāra* (stage manager) entered with his wife and described the place and occasion of the action. The last sentence of his prologue served as a bridge leading to the action of the play. In *Śakuntalā* he refers to the bewitching song of his wife, which has made him forget his surroundings as the pursuit of a deer has made the king forget his state affairs. At this point the king enters, riding his hunting chariot, and the spectators are plunged into action of the play.

The *vidūṣaka* (clown) is a noble, good-hearted, blundering fool, the trusted friend of the hero. A bald-headed glutton, comic in speech and manners, he is the darling of the spectators. With the decline of Sanskrit drama the folk theatre in various regional languages inherited the conventions of the opening prayer song, the *sūtra-dhāra,* and the *vidūṣaka.*

The only surviving Sanskrit drama is *kudiyattam,* still performed by the Cakkayars of Kerala. Some principles of the *Nāṭya-śāstra* are evident in their presentations.

The earliest available classical dramas are 13 plays edited in 1912 by Pandit Ganapati Sastri, who dug out their manuscripts in Trivandrum, the capital of Kerala state. These, ascribed to Bhāsa (1st century BC–1st century AD), include the one-act *Ūrubhaṅga* ("The Broken Thigh"), a tragedy that is a departure from Sanskrit convention, and the six-act *Svapnavāsavadatta* ("The Dream of Vāsavadattā").

The most acclaimed dramatist is Kālidāsa. Other important playwrights succeeding him include Harṣa, Mahendravikramavarman, Bhavabhūtī, and Viśakhādatta. An exception is King Śūdraka, whose work is perhaps the most theatrical in the entire Sanskrit range.

"The Litt Clay Cart

The title of "The Little Clay Cart" represents a departure from Sanskrit tradition, in which a *prakaraṇa* was generally named after its hero and heroine. *Mālavikāgnimitra,* for example, is the love story of Princess Mālavikā and King Agnimitra, *Vikramorvaśī* is the tale of King Purūravas and the heavenly nymph Urvaśī, and *Mālatī-Mādhava* is the love drama of Mālatī and Mādhava. Śūdraka, as if to mock tradition, chose an insignificant homely incident—the hero's son playing with a toy cart—and elevated this to the title.

"The Little Clay Cart" has a wide range of characters. The plot does not progress in a straight line but zigzags along a winding path. During its 10 acts the hero does not appear in four of them, the heroine is absent from three, and the lustful villain disappears after the first act until the eighth. Each act is an almost independent play. The device used to link the acts is that of ending them with subtitles that sum up their particular themes or plots.

"The Little Clay Cart" has been successful in the West, whereas Indian audiences, still fed on poetic-flavoured characters and romances of an ethereal type, have favoured *Śakuntalā.* Western audiences find "The Little Clay Cart" more in their own tradition of realism and individualized characterization. Its "lisping villain," gamblers, and rogues have something in common with Shakespeare's comic characters and Molière's crooks. "The Little Clay Cart" is better theatre, whereas *Śakuntalā* is better poetry.

Folk theatre. After the decline of Sanskrit drama, folk theatre developed in various regional languages from the 14th through the 19th centuries. Some conventions and stock characters of classical drama (stage preliminaries, the opening prayer song, the *sūtra-dhāra,* and the *vidūṣaka*) were adopted into folk theatre, which lavishly employs music, dance, drumming, exaggerated makeup, masks, and a singing chorus. Thematically it deals with mythological heroes, medieval romances, and social and political events, and it is a rich store of customs, beliefs, legends, and rituals. It is a "total theatre," invading all the senses of the spectators.

The most crystalized forms are the *jātrā* of Bengal, the *nautanki, rāmlīlā,* and *rāslīlā* of North India, the *bhavai* of Gujarāt, the *tamāshā* of Mahārāshtra, the *terukkūttu* of Tamil Nadu, and the *yakṣagāna* of Kanara.

Places and kinds of folk theatre

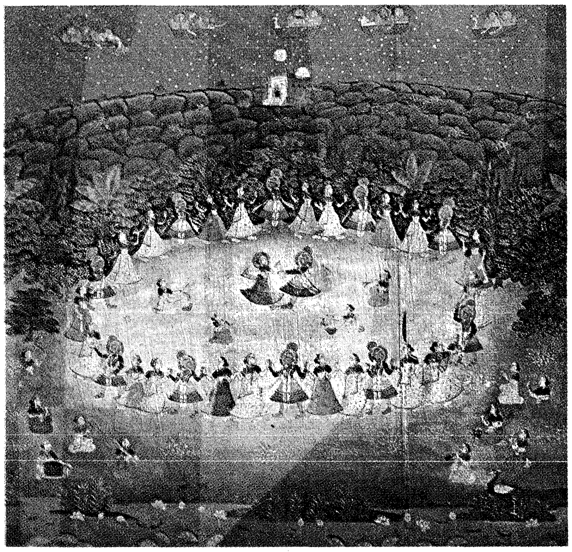
Folk theatre is performed in the open on a variety of arena stages; round, square, rectangular, multiple-set. The *bhavai,* enacted on a ground-level circle, and the *jātrā,* on a 16-foot (five-metre) square platform, have gangways that run through the surrounding audience and connect the stage to the dressing room. Actors enter and exit through these gangways, which serve a function similar to the *hanamichi* of the Japanese Kabuki theatre. In the *rāmlīlā,* the action sometimes occurs simultaneously at various levels on a multiple set. Actors in *nautanki* and *bhavai* sit on the stage in full view instead of exiting and sing or play an instrument as a part of the chorus. In the *rāmlīlā,* the actor playing Rāvaṇa removes his ten-headed mask when he is not acting and continues sitting on his throne, but for the spectators he is theatrically absent. Asides, soliloquies, and monologues abound. Scenes melt into one another, and the action continues in spite of change of locale.

In most folk forms the art of the actor is hereditary. He learns by watching his elders throughout childhood. He starts with drumming, then dancing, plays female roles, and then major roles.

All roles are played by men except that of the *tamāshā* woman, who is always a dancer-singer-actress. Recently, women have started playing female roles in the *jātrā* but have failed to achieve the artistic stature of their professional male counterparts.

In the *rāmlīlā* and *rāslīlā,* the principal characters—Rāma and Krishna—are always played by boys under 14, because tradition decreed they must be pure and innocent. They are considered representatives of the gods and are worshipped on these occasions. In the *rāmlīlā* the *vyas* ("director"), present on the stage throughout the performance, prompts and directs the characters loudly enough for the audience to hear. This is not regarded as disturbing because it is an accepted part of the tradition. Adult roles such as Rāvaṇa and Hanuman are sometimes played by the same individual throughout his life.

Mohan Khokar



Rāsīlā folk drama of northern India, watercolour, late 19th century. In the collection of Mohan Khokar, New Delhi. 1.3 × 1.3 m.

on-
ligious
lk
ınce

Of the nonreligious forms, the *jātrā* and the *tamāshā* are most important. The *jātrā,* also popular in Orissa and eastern Bihār, originated in Bengal in the 15th century as a result of the *bhakti* movement, in which devotees of Krishna went singing and dancing in processions and in their frenzied singing sometimes went into acting trances. This singing with dramatic elements gradually came to be known as *jātrā,* which means "to go in a procession." In the 19th century the *jātrā* became secularized when the repertoire swelled with love stories and social and political themes. Until the beginning of the 20th century, the dialogue was primarily sung. The length has been cut from all night to four hours. The *jātrā* performance consists of action-packed dialogue with only about six songs. The singing chorus is represented by a single character, the *vivek* ("conscience"), who can appear at any moment in the play. He comments on the action, philosophizes, warns of impending dangers, and plays the double of everybody. Through his songs he externalizes the inner feelings of the characters and reveals the inner meaning of their outer actions.

The *tamāshā* (a Persian word meaning "fun," "play," or "spectacle") originated at the beginning of the 18th century in Mahārāshtra as an entertainment for the camping Mughal armies. This theatrical form was created by singing girls and dancers imported from North India and the local acrobats and tumblers of the lower-caste Dombari and Kolhati communities with their traditional manner of singing. It flourished in the courts of Marāthā rulers of the 18th and 19th centuries and attained its artistic apogee during the reign of Bājī Rāo II (1796–1818). Its uninhibited *lavani*-style singing and powerful drumming and dancing give it an erotic flavor. The most famous *tamāshā* poet and performer was Ram Joshi (1762–1812) of Sholāpur, an upper class Brahmin who married the courtesan Bayabai. Another famous singer-poet was Patthe Bapu Rao (1868–1941), a Brahmin who married a beautiful low-caste dancer, Pawala. They were the biggest *tamāshā* stars during the first quarter of the 20th century. The *tamāshā* actress, commonly called the *nautchi* (meaning "nautch girl," or "prostitute") is the life and soul of the performance. Because of their bawdy elements, women never see *tamāshā* plays, nor do respectable men.

In the 20th century, *jātrā* and *tamāshā* both have become highly organized and are commercially run. Troupes are in heavy demand and work for nine months. Over 700 *tamāshā* troupes with 2,000 dancer-actresses tour the rural areas, providing a living for about 40,000 people. The *jātrā* is the most successful commercially. Its star actors draw more than any other professional actor in the theatrical centre of Calcutta.

Puppet
theatre
of North
India

Popular in North India are the *putliwalas* ("puppeteers") of Rājasthān, who operate marionettes made of wood and bright-coloured cloth. The puppet plays deal with kings, lovers, bandits, and princesses of the Mughal period. Generally, the puppeteer and his nephew or son operate the strings from behind, while the puppeteer's wife sits on her haunches in front of the miniature stage playing the drums and commenting on the action. The puppeteer chirps, whimpers, and squeals in animal–bird voices and creates battles and tragic moments, expresses pathos, anger, and laughter. In Andhra Pradesh the puppets, called *tholu bommalata* ("the dance of leather dolls"), are fashioned of translucent, coloured leather. These are projected on a small screen, like colour photographic transparencies. Animals, birds, gods, and demons dominate the screen. The puppeteer manipulates them from behind with two sticks. Strong lamps are arranged so that the size, position, and angle of the puppets change with the distance of the light. They are similar to the *wayang kulit* puppets of Indonesia but are much smaller and quicker moving.

In the absence of a powerful Indian city theatre (with the exception of a few in Calcutta, Bombay, and Tamil Nadu), folk theatre has kept the rural audiences entertained for centuries and has played an important part in the growth of modern theatres in different languages. The 19th-century dramatist Bharatendu Harishchandra, who was responsible for the birth of Hindi drama, used folk conventions—the opening prayer song, tableaux, comic interludes, duets, stylized speech—and combined these with Western theatrical forms in vogue at that time. Parsi companies adapted the popular folk techniques for their extravaganzas and were a major influence until the 1930s. Rabindranath Tagore, rejecting the heavy sets and realistic decor of the commercial companies, created a lyrical theatre of the imagination. Much influenced by the *baul* singers and folk actors of Bengal, he introduced the Singing Bairagi and the Wandering Poet (similar to the *vivek* of the *jātrā*) in his dramas. In the late 20th century, folk theatre has been viewed as a form that can add colour and vitality to contemporary theatre.

Modern theatre. Modern Indian theatre first developed in Bengal at the end of the 18th century as a result of Western influence. The other regional theatres more or less followed Bengal's pattern, and within the next 100 years they took the same meandering path, though they never achieved the same robust growth.

British
influence
on
Bengali
theatre

The British conquered Bengal in 1757 and influenced local arts by their educational and political systems. Their clubs performed Shakespeare, Molière, and Restoration comedies, introducing Western dramatic structure and the proscenium stage to the Indian intelligentsia. With the help of Golak Nath Dass, a local linguist, Gerasim Lebedev, a Russian bandmaster in a British military unit, produced the first Bengali play, *Chhadmabes* ("The Disguise"), in 1795 on a Western-style stage with Bengali players of both sexes. Subsequently, Bengali playwrights began synthesizing Western styles with their own folk and Sanskrit heritage. With growing national consciousness, theatre became a platform for social reform and propaganda against British rule. Among the most important playwrights were Michael Madhu Sudan (1824–73), Dina Bandhu Mitra (1843–87), Girish Chandra Ghosh (1844–1912), and D.L. Roy (1863–1913).

The success of Dina Bandhu Mitra's *Nildarpan* ("Mirror of the Indigo"), dealing with the tyranny of the British indigo planters over the rural Bengali farm labourers, paved the way for professional theatre. The actor-director-writer Girish Chandra Ghosh founded in 1872 the National Theatre, the first Bengali professional company, and took *Nildarpan* on tour, giving performances in the North Indian cities of Delhi and Lucknow. The instigatory speeches and lurid scenes of British brutality resulted in the banning of this production. To overcome censorship difficulties, playwrights turned to historical and mythological themes with veiled symbolism that was clearly understood by Indian audiences. The heroes and villains of these plays came to represent the Indian freedom fighter against the British oppressor. Girish's historical tragedies *Mir Qasim* (1906), *Chhatrapati* (1907), and *Sirajuddaulah* (1909) bring out the tragic grandeur of heroes who fail because of some inner weakness or betrayal of their colleagues. D.L. Roy emphasized the same aspect of nationalism in his historical dramas *Mebarapatan* (The Fall of Mebar), *Shahjahan* (1910), and *Chandragupta* (1911).

Girish introduced professional efficiency and showmanship. His style of acting was flamboyant, with fiery grace. Actors such as Amar Datta and Dani Babu carried his style into the early 1920s. The acting and production methods of the Star, the Minerva, and the Manmohan Theatres (all professional) were modelled on Girish's pioneer work.

The beginning of realism

The first elements of realism were introduced in the 1920s by Sisir Kumar Bhaduri, Naresh Mitra, Ahindra Chowdhuri, and Durga Das Banerji, together with the actresses Probha Devi and Kanka Vati. In his Srirangam Theatre (closed in 1954), Sisir performed two most memorable roles: the again Mughal emperor Aurangzeb and the shrewd Hindu philosopher-politician Cānakya. Sisir's style has been refined by actor-director Sombhu Mitra and his actress wife Tripti, who worked in the Left-wing People's Theatre movement in the 1940s. With other actors they founded the Bahurupee group in 1949 and produced many Tagore plays including *Rakta Karabi* ("Red Oleanders") and *Bisarjan* ("Sacrifice"), so far unattempted by any professional company.

Rabindranath Tagore (1861–1941), steeped in Hindu classics and indigenous folk forms but responsive to European techniques of production, evolved a dramatic form quite different from those of his contemporaries. He directed and acted in his plays along with his cousins, nephews, and students. These productions were staged mostly at his school, Santiniketan, in Bengal as a nonprofessional and experimental theatre. The Calcutta elite and foreign visitors were attracted to these performances.

A painter, musician, actor, and poet, Tagore combined these talents in his productions. He used music and dance as essential elements in his latter years and created the novel opera-dance form in which a chorus sat on the stage and sang while the players acted out their roles in dance and stylized movements. Sometimes Tagore himself sat on a stool acting as the *sūtra-dhāra* and chanted to the accompaniment of music and drum as the dancing players became visual moving pictures.

In northern and western India, theatre developed in the latter half of the 19th century. The Bombay Parsi companies, using Hindi and Urdu, toured all over India. Their spectacular showmanship, based on a dramatic structure of five acts with songs, dances, comic scenes, and declamatory acting, was copied by regional theatres. The Maharashtrian theatre, founded in 1843 by Visnudas Bhave, a singer-composer-wood-carver in the court of the Raja of Sāngli, was developed by powerful dramatists such as Khadilkar and Gadkari, who emphasized Marāthā nationalism. The acting style in Maharashtrian theatre remained melodramatic, passionately arousing audiences to laughter or tears.

In the south, the popularity of dance-dramas has not allowed theatrical realism to flourish. Tamil commercial companies with their song and dance extravaganzas have dominated Andhra Pradesh, Kerala, and Mysore. The most outstanding Tamil company since the independence of India in 1947 has been the T.K.S. Brothers of Madras, famous for trick scenes and gorgeous settings. Also famous is the actor-producer-proprietor Rajamanickam, who specializes in mythological plays with an all-male cast, using horses, chariots, processions, replicas of temples, and even elephants.

Urdu and Hindi drama

Urdu and Hindi drama began with the production of *Indrasabha* by Nawab Wajid Ali Shah in 1855 and was developed by the Parsi theatrical companies until the 1930s.

Parsi theatre was an amalgam of European techniques and local classical forms, folk dramas, farces, and pageants. Mythical titans thundered on the stage. Devils soared in the air, daggers flew, thrones moved, and heroes jumped from high palace walls. Vampire pits, the painted back cloth of a generalized scene, and mechanical devices to operate flying figures were direct copies of the 19th-century Lyceum melodramas and Drury Lane spectacles in London.

The star film actor Prithvi Raj Kapoor founded Prithvi Theatres in Bombay in 1944 and brought robust realism to Hindi drama, then closed down in 1960 with a sense of completion after many tours throughout India. Prithvi's sons, nephews, and old associates worked in his large company, which became a training centre for many actors who later joined the films. Among these was the outstanding stage actress Zohra Sehgal, a former dance partner of Uday Shankar in the 1930s who had tremendous emotional depth and range, rare in actresses on the Hindi stage. Out of Prithvi's eight productions, in which he always played the lead, the best was *Pathan* (1946), which ran for 558 nights. It deals with the friendship between a tribal Muslim khān and a Hindu dewan and is set in the rugged frontier from which Prithvi came. This tragedy of two archtypes in which the khān sacrifices his son to save the life of his friend's son had intensity of action, smoldering passion, and unity of mood and achieved the highest quality of realism on the Hindi stage to this day.

Among the actors who molded regional-language theatres are Shri Narayan Rao Rajhans (popularly known as the Bala Gandharva of the Mahārāshtra stage), Jayashankar Bhojak Sundari of Gujarāt, and Sthanam Narasimhrao of Andhra. All three specialized in female roles and were star attractions during the first quarter of the 20th century.

In the last half of the 20th century, two outstanding actor-directors are Ebrahim Alkazi, director of the National School of Drama in New Delhi, and Utpal Dutt, who founded the Calcutta Little Theatre Group in 1947, which originally performed plays in English and in 1954 changed to productions in Bengali. Dutt is an actor fully committed to the revolutionary ideology of the Chinese Communist leader Mao Tse-tung. He acts on open-air stages in rural areas of Bengal, where he exerts a strong artistic and political influence.

Since Lebedev in 1795 there has been a continuous stream of Western-trained actors and producers who have been revitalizing regional-language theatrical groups. Nawab Wajid Ali Shah had visiting French opera composers in his mid-19th-century court. Tagore did his first opera, *Valmiki Pratibha* ("The Genius of Vālmīkī"), in 1881, after returning from England, where he became familiar with Western harmonies. Prithvi Raj Kapoor, E. Alkazi, and Utpal Dutt all had their earlier training in English productions. Norah Richards, an Irish-born actress who came to the Punjab in 1911, produced in 1914 the first Punjabi play, *Dulhan* ("The Bride"), written by her pupil I.C. Nanda. For 50 years she promoted rural drama and inspired actors and producers, including Prithvi Raj Kapoor.

India's genius still lies in its dance-dramas, which have a unique form based on centuries of unbroken tradition. There are very few professional theatre companies in the whole of India, but thousands of amateur productions are staged every year by organized groups. Out of this intense experimental activity, the Indians hope a contemporary national theatre will emerge, influenced by Western techniques but distinctly Indian in flavour.

ntres
ining

Many centres for theatrical training have been established in the 20th century. Among the most important are the National School of Drama and the Asian Theatre Institute in New Delhi, Sangeet Natak Akademi (National Academy of Music, Dance, and Drama) in New Delhi, and the National Institute for the Performing Arts in Bombay. Bharatiya Natya Sangh, the union of all Indian theatre groups, was founded in 1949 and is centered in New Delhi. Affiliated with UNESCO's branch of the International Theatre Institute, it organizes drama festivals and seminars, as well as serving as a centre for information.

SRI LANKA (CEYLON)

The ritualistic dances of Sri Lanka have attained world fame for their weird mystical beauty. Literary drama has not flourished, because the monks of predominantly Buddhist Sri Lanka shunned theatre. Dramatic activity found expression in exorcism ceremonies and masked dramas that employed mime, song, dance, acrobatics, and bits of prose dialogue. Heavily influenced by India, Sri Lanka's Kandyan dance and *kōlam* plays have South Indian origins. But over the centuries these have been transformed and now have a distinctly Sri Lankan character.

It is difficult to divide Sri Lankan performing arts into dance and drama, because a *kōlam* play uses dance and song, and the devil dance has bits of improvised prose dialogue.

Dance. *Devil dance.* The Buddhists of Sri Lanka still believe in supernatural beings and the healing power of magical rites. Their devil dancing is the expression, however, of pre-Buddhist beliefs.

The devil dance is performed to cure a person gripped by disease, insanity, or bad luck believed to be caused by some malignant spirit and to propitiate demons and deities to bring good fortune. The dancers belong to a lower-caste community and are professional. During their performance the patient lies to one side. Several palm-leaf shrines are constructed outside his house, each dedicated to a particular demon to lure him into the arena. The role of the Vesamuni, king of all demons, is played by the chief exorcist.

Three types of supernatural beings have to be appeased: demons, deities, and others that are half demon, half deity. The most terrible is Riri Yakka (Demon of Blood), who inhabits cremation grounds and graveyards and rides a pig. His belly is smeared with blood, and he has a monkey's face and four clawed hands that hold a parrot, a sword, a rooster, and a human head.

ostumes
f the
evil
ancers

The dancers, all men, each wear around their heads a red cloth fringed with long ribbons of palm leaves hanging down like female hair, a strip of cloth around their chests, 22 yards (20 metres) of thin white cloth wound so skillfully around their hips that it never comes loose during an entire night of violent activity, and clusters of bells fitted around their calves to make a deafening jingle. Their appearance is half female, half male.

The dance is punctuated by little pieces of mimes and magical actions, with naked-chested drummers pounding to the accompaniment of a chorus of singers. The climax is reached when the dancers, holding flaming torches in both hands, whirl and spin, forming circles of fire around themselves. The flames lick their bodies, but they remain unsinged. The dancers leap and dive through the air in seeming defiance of gravity. In this surcharged atmosphere they pause to put on masks representing various demons. These have terrifying expressions. They romp and stomp in circles, describing their identity and purpose of visiting. The particular demon associated with the malady enters the patient's body. The chief exorcist questions, threatens, torutures, beseeches, and offers bribes to appease the demon until it finally leaves and its victim is healed.

The *sanni yakku* dance, exorcising the disease demon, has a series of humorous impersonations. One is of the demon as a beautiful woman, then as a pregnant woman, and finally as a mother. The exorcists ask questions about her pregnancy, and she lists all the respectable men of the village.

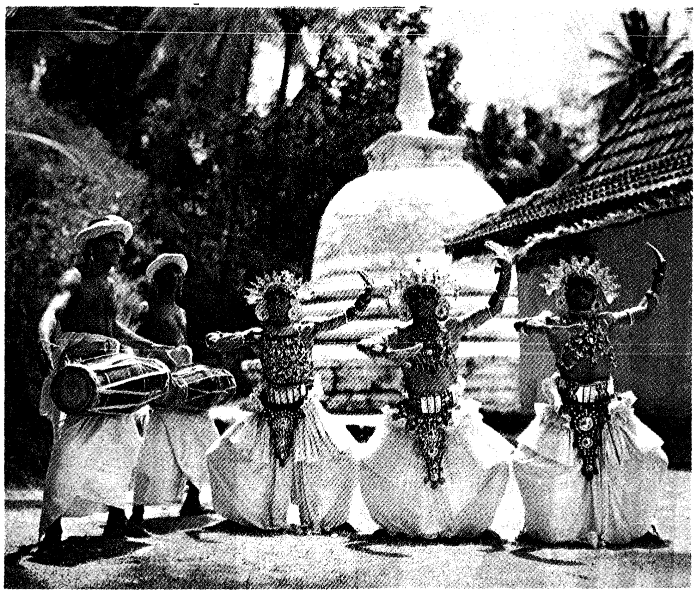
Out of many devil-dance ceremonies, most picturesque and important are the *kohomba kankariya* (or "ritual of the god Kohomba"), performed to ensure prosperity and to get rid of pestilence, and the *bali,* danced to propitiate the heavenly beings.

Kandyan dance. The *kandyan* dance, shorn of occult ceremonies, is highly sophisticated and refined. It flourished under the Kandyan kings from the 16th through the 19th centuries, and today it is considered the national dance of Sri Lanka. It has four distinct varieties: The *pantheru, naiyadi, udekki,* and *ves* (the most artistic and renowned). Its energetic movements and postures are reminiscent of India's *kathākali.* Besides the above four styles, there are 18 *vannamas* (dance enactments) including the *gajaga vannama,* depicting the elephant; the *hanuma vannama,* the monkey; and the *mayura vannama,* the peacock. These beautiful animal movements and abstract impersonations have been distilled and perfected over several hundred years.

The
vannamas

Hindu mythological themes were originally the subjects of *kandyan* dances, the most popular being Rāma's crossing over to Laṅkā with the help of his monkey general and his reunion with Sītā. Gradually, stories of kings and legendary heroes and mimes of birds and forest beasts were introduced. The Kandyan kings elevated the dance to such beauty and skill that the Buddhists began admitting it into their temple courtyards as a tribute to the glory of their religion. It became a part of the annual August Perahera festival, in which a procession of gilded elephants, palanquins, saffron-robed monks, drummers, and chanters move majestically to the Temple of the Tooth in Kandy, where Buddha's tooth is enshrined. The *kandyan* dancers are a glittering attraction as they perform en route to the temple.

Ewing Krainin—Stockpile



Kandyan dancers and drummers from Sri Lanka.

A *kandyan* dancer wears a pagoda-like conical silver headpiece with glistening forehead fringe and huge earpieces, many-stranded bead necklaces of silver and ivory across his naked torso, beaten-silver epaulets on his biceps, and hollow silver anklets filled with silver beads to make them rattle. He spins with sudden leaps and reaches violent climaxes of geometric patterns. The sudden right and left turns of his head make the onlookers dizzy. When telling a story, he sings descriptive passages and enacts them with spurts of dancing.

Masked drama. Out of the four folk-drama forms—*kōlam, sokari, nadagam,* and *pasu*—the most highly de-

The *kōlam*

veloped and significant is the *kōlam,* in which actors wear brightly painted and intricately carved wooden masks. The word *kōlam* is of Tamil origin and means "costume," "impersonation," or "guise." The performance consists of the masked representation of many isolated characters, such as kings, demons, deities, hunters, animals and birds, the washerman, the police constable, a pregnant woman—a British Museum manuscript concerning the *kōlam* lists 53 such characters. The most terrifying masks are of the demons, with twisted faces, protruding tusks, and cavelike nostrils for snorting fury. The *nāga* demon has a long, flaming, red tongue and dozens of cobras writhing around his head. Some old masks have only one large bulging eye, with a cobra hissing out from one nostril. The design of these masks lies between the strongly stylized tribal masks of Africa and the highly polished, sophisticated masks of the Japanese Nō Theatre. Five basic colours are used—red, blue, yellow, green, and black, the last two for lower rank characters. Exaggerated comicality, distortions, bulges, nightmarish whimsy, bright colours, and artful carving of the masks have been significant factors in keeping this form of drama alive.

The *kōlam* is performed once a year for seven to ten nights, starting at night after dinner and lasting through the early hours of the morning. The performance is generally held in the open courtyard of a house, to the accompaniment of two drummers, an instrumentalist, and a singing chorus with leader. After songs in praise of Lord Buddha and others (including the patron of the show), the *sabhapati* (master of ceremonies) describes the origin of *kōlam*—how an Indian king's pregnant wife expressed a desire to see a masked dance-drama and how a troupe was invited from a distant court. The *sabhapati* then introduces the masked characters as they enter and describes their various vocations and backgrounds.

Out of many, two plays are especially famous: the *Sandakinduru Katava* and the *Gothayimbala Katava.* The former deals with the legendary idyllic love between a half-human, half-bird couple singing and dancing in a forest. The King of Banaras comes hunting and, attracted by the beautiful Kinduri, kills her husband and makes advances to her. Rejected, he is ready to kill her when Lord Buddha appears and brings her husband back to life. In the *Gothayimbala Katava* the beautiful wife of the warrior Gothayimbala bathing in a pond attracts the attention of a demon, who falls in love with her. The enraged husband comes and chops off the demon's head, which, because of its magical power, reunites itself with the body every time it is cut off. Finally, the forest deity comes and rescues the warrior.

The recorded history of *kōlam* is not very old. There is only one known early eye witness account of *kōlam,* that of John Callaway, who in 1829 published 185 verses of a play with a description of the performance and some sketches of the masks and a brief introduction concerning the masquerade. According to Callaway, the dancers did not sing. The chanters described the characters in the third person and sometimes exclaimed to draw the attention of the audience to a particular action. The earliest *kōlam* text is preserved in the Colombo National Museum on palm leaves; another is in the British Museum inscribed on paper. The oldest printed text, edited in 1895 by A.G. Perera, is in the Colombo National Museum Library.

Masks are made of the light woods kaduru (Strychnos nux vomica) and ruk-attana (*Alstonia scholaris*) and after 50 years start decaying; consequently, the earlier masks are no longer in existence.

There has been an important revival of interest in drama in Sri Lanka since independence. E.R. Sarachchandara, a scholar of traditional Ceylonese theatre, has been responsible for a major breakthrough in revitalizing and adapting for the modern stage traditional dramatic forms such as the *kōlam.* Several new playwrights have become prominent in the mid-20th century. Foremost among them is Henry Jayasena. A producer-writer-actor, Jayasena has written and staged plays in Sinhalese and translations of foreign plays. But modern theatre is still very weak.

DANCE AND THEATRE IN KASHMIR

The Vale of Kashmir, predominantly populated by Muslims, has remained aloof from the main cultural currents of India. The ancient caves and temples of Kashmir, however, reveal a strong link with Indian culture at the beginning of the Christian Era. At one time the classical dances of the south are believed to have been practiced. When Islām was introduced, in the 14th century, dancing and theatrical arts were suppressed, being contrary to a strict interpretation of the Qur'ān. These arts survived only in folk forms and were performed principally at marriage ceremonies. The popular *hafiza* dance performed by Kashmiri women at weddings and festivals to the accompaniment of *sufiana kalam* (devotional music of the Muslim mystics known as Ṣūfīs) was banned in the 1920s by the ruling maharaja, who felt this dance was becoming too sensual. It was replaced by the *bacha nagma,* performed by young boys dressed like women. A popular entertainment at parties and festivals, it is also customarily included in modern stage plays.

In contrast to its natural scenic richness, Kashmir is theatrically a pauper. Theatrical productions are generally amateurish, since there is no regular performing company or any tradition of civic theatre.

Bhand jashna

There is only the *bhand jashna* ("festival of clowns"), a 300- to 400-year-old genre of Kashmiri folk theatre. Performed in village squares, it satirizes social situations through dance, music and clowning.

The Kashmiri-language theatre was founded in 1947, when a new national consciousness, the aftermath of the independence of the Indian subcontinent from Britain, inspired playwrights and folk actors to dramatize topical events and create a "visual newspaper" for the people. Left-wing propaganda plays such as *Zamin Sanz* ("Who Owns the Land?") and *Jangbaaz* ("The Warmonger"), though mediocre, had topical interest. Notable among those who tried their hand at writing for the stage is the poet Nadim, author of two operas, *Bambur-yambarzal (The Bumblebee)* and *Himal Nagraj (The Beautiful Woman and the Snake Prince).*

Since the 1960s, the Jammu and Kashmir Academy of Art, Culture, and Languages has been struggling to promote theatre in the Kashmiri and Dogri languages, but with little success. Its emphasis is on literary dramas and folk-dance festivals of regional appeal.

PAKISTAN

Muslim culture has frowned upon the performing arts, with the result that there is no Arabic or Persian classical theatre. The only possible sources of drama were the Persian passion plays dealing with the martyrdom of Ḥuysayn (grandson of Muḥammad) in the desert of Karbalā' in 680 AD, which have inspired some Urdu playwrights. Pakistan, a Muslim country, therefore either had to find a theatrical heritage in Urdu and Bengali theatre, which had been flourishing in India long before the partition, or look to the West. It did both. The Urdu-language theatre of Pakistan had started in the Lucknow court of Nawab Wajid Ali Shah in 1855 and was nurtured by both Muslim and Hindu artists. In Pakistan the *kathak* style is preferred because of its strong Muslim flavour and Mughal court associations. Cut off from Hinduism and its lore, Pakistani performers use these Indian classical dance styles to interpret the aspirations of a young nation, while their folk dances express the character of Pakistan's rural culture.

Khaṭak

Folk dance. Pakistan's dances are virile and explosive. *Bhangra* and Khaṭak are the most popular. Khaṭak is a dance of the tribal Pathans, known for their hospitality and feuds in the rugged hills of the northwest. It originated in zealous preparations for raids and celebrations of victories. In the 20th century, any joyous event is the occasion for this community dance. The Pathans, dressed in baggy *salwars,* embroidered waistcoats, and skullcapped turbans, perform it holding a rifle in both hands. In a frenzy they spin and somersault, float and whirl, with sudden bursts of swordplay to the accompaniment of drums and pipes. Because of its popularity, Khaṭak is presented to visiting dignitaries and for this purpose has been refined into choreographed productions.

Bhangra, folk dance of the Punjab region of Pakistan and India.

Lustig—FPG

male
nces

Important dances by women are the *sammi, kikli, giddha,* and *luddi.* Except for the *sammi,* which has a slow rhythm accompanied by a sad song because of its association with the tragic love legend of Princess Sammi and Prince Dhola, all the other forms are charged with energy and fast rhythms. The *kikli* is performed by teen-age girls in groups of two. The partners cross their arms, interlock their fingers, and, touching the toes of their feet, stretch backward and whirl. The *giddha* is danced in a circle, the participants keeping the rhythm by clapping their hands. Two women impulsively leave the circle, jump into the centre, and perform a hilarious mimetic dance enacting a *boli* (two-line song) and again join the circle to dance in a ring and allow another couple to take the centre. In the *luddi,* women click their fingers and clap their hands, moving in a circle by jumps and half-turns and accelerating their rhythm by stamping their feet.

Performing arts in the Punjab. The genius of Punjabis finds expression in love stories, lusty dancing, and humour. The *mirasi*s (professional wits), *naqalia*s (mummers), and *domani*s (female singer-actresses) are professional performers belonging to the lower classes. They exploit all the tricks of exaggeration, absurdity, malapropism, comic gags, and lewd references. In the performance of a *naqal* (comic sketch), two people constitute a troupe. The leader holds a leather folder and slaps his foolish partner, who leads his master to a hilarious situation through absurd replies. Expert in mime and clowning, these character types are distantly related to the Western court fool and the commedia dell'arte.

Theatre in Pakistan. Urdu theatre grew out of a spectacular production of *Indrasabha* ("The Heavenly Court of Indra"), an operatic drama written by the poet Agha Hasan Amanat and produced in 1855 in the palace courtyard of the last Nawab of Oudh, Wajid Ali Shah. The story deals with the love of a fairy and Prince Gulfam. The fairy takes her lover to heaven where the angry and jealous Indra hurls him down to earth. Finally, the fairy, through her songs and dances, wins the heart of Indra, and the two lovers are united. Wajid Ali Shah, an expert *kathak* dancer and author of many valuable treatises on stage techniques, composed some of the melodies and dances for his production and used folk conventions, gorgeous costumes, elaborate settings, and gold-inlaid masks. *Indrasabha* was a fantastic success; it was translated into almost all the regional languages, with many local variants. Its characters—Sabaz Pari (Green Fairy), Kala Deo (Black Devil), Lal Deo (Red Devil)—became a part of the theatrical vocabulary of the subcontinent.

Parsi theatre. During the second half of the 19th century, Urdu was the main spoken and written language of the northern half of the subcontinent and understood in almost all the principal cities. The Parsis (originally Zoroastrians from Iran who settled on the coast of Bombay), comprising a wealthy community with sharp business acumen, were the pioneers in establishing a commercial theatre, that lasted from 1873 to 1935 and influenced all the other regional theatres. Though located mainly in Bombay and Calcutta, the Parsi companies toured the subcontinent with huge staffs, sets, and an army of players.

The best known playwright of this period is Agha Hashr (1876–1935), a poet-dramatist of flamboyant imagination and superb craftsmanship. Among his famous plays are *Sita Banbas,* based on an incident from the *Rāmāyaṇa; Bilwa Mangal,* a social play on the life of a poet, whose blind passion for a prostitute results in remorse; and *Aankh ka Nasha* ("The Witchery of the Eyes"), about the treachery of a prostitute's love, with realistic dialogue of a brothel. Many of Hashr's plays were adapted from Shakespeare: *Sufayd Khūn* ("White Blood") was modelled on *King Lear,* and *Khūn-e Nāḥaq* ("The Innocent Murder") on *Hamlet.* His last play, *Rustam-o-Sohrab,* the tragic story of two legendary Persian heroes, Rustam and his son Sohrab, is a drama of passion and fatal irony.

Parsi
productions

Productions by Parsi theatrical companies were large-budgeted affairs. Plays opened with the actors in full makeup and costume, their hands folded and eyes closed, singing a prayer song in praise of some deity, and generally ended in a tableau. Sometimes at curtain call the director rearranged the tableau in a split second and offered a variant. Actors were required to know singing, dancing, music, acrobatics, and fencing and to possess strong voices and good physical bearing. In improvised auditoriums with bad acoustics and packed with more than 2,000 people, actors' voices reached the farthest spectator. Plays began at 10 o'clock and lasted until dawn, moving from comedy to tragedy, from pathos to farce, from songs to the rattle of swords, all interspersed with moral lessons and rhyming epigrams. The droll humour and realism of the comic interludes remain unsurpassed in contemporary Urdu drama. Important playwrights of this period were Narain Prasad Betab, Mian Zarif, and Munshi Mohammed Dil of Lucknow. All took inspiration from Hindu mythology and Persian legends, transforming these tales into powerful dramas.

Imtiaz Ali Taj (1900–70) was a bridge between Agha Hashr and contemporary Pakistani playwrights. His *Anarkali* (1922), the tragic love story of a harem girl, Anarkali, and Crown Prince Salim (son of Akbar the Great), unfolds the love-hate relationship of a domineering emperor and his rebellious son. Brilliant in treatment and character analysis, this play has been staged hundreds of times by amateur groups and has entered the list of Urdu classics.

In the absence of a professional company, Urdu theatre has found it difficult to strike roots. After 1947 many Muslim actors and writers were absorbed by the Indian film industry in Bombay, and they found it difficult to adjust their great talent to amateur theatrical clubs. All the same, plays have been staged in Karāchi, Lahore, and Rāwalpindi. The best productions have been those dealing with topical themes—refugee problems, new adjustments, the corrupt bureaucracy, the Kashmir issue, and other sociopolitical issues. Agha Babar in Rāwalpindi produced *Burra Sahib* (1961: "The Big Boss"), an adaptation of Gogol's *Government Inspector,* setting it in Pakistan. *Tere Kuce se Jub Hum Nikle* ("Thrown Out of Your Lane"), by Naseer Shamshi, describes the pathetic condition of an aristocratic family in Delhi that is forced to leave home because of communal riots. In *Lal Qile se Lalukhet Tak* ("From the Red Fort to Lalukhet"), by Khwajah Moinuddin, the comedy arises out of the pitiable condition of the refugees who leave their well-settled existence in Delhi dreaming of prosperity, take a tedious journey, and arrive homeless in Karāchi to find shelter in thatched hovels. Ali Ahmed, an avant-garde actor-director in Karāchi, presents his plays with polished stagecraft and esoteric appeal.

Lahore remains the centre of amateur theatre based on the tradition of the late directors A.S. Bokhari and G.D.

Sondhi, both former principals of the Government College in Lahore. In 1942 G.D. Sondhi built the Open-Air Theatre, situated on a small artificial hillock in the Lawrence Gardens and perhaps the best in all of South Asia. It has remained the centre of dramatic contests and festivals and is a favourite of visiting dancers and actors.

The actor-playwright Rafi Peer, with his knowledge of Western theatre as a result of his training in Berlin in the 1930s, has helped to develop Pakistani theatre. Professional in approach, he has produced radio and stage plays and has been a critical colleague of A.S. Bokhari and Imtiaz in the revival of amateur theatre.

Radio and television plays. Plays are being written for radio and television that are readily adaptable for the stage, and vice versa. Saadat Hussan Mantoo (1912–55), the greatest writer of short stories and author of over 100 radio plays and features, is still the model for contemporary writers for plot construction, bitter realism, and whimsical dialogue. His collection of plays (1942–45), including *Mantoo ke Dramay* ("Mantoo's Plays"), *Ao* ("Come"), and *Teen Aurten* ("Three Women"), have flashes of the then-unborn Theatre of the Absurd.

More dramas are written in Urdu today than are staged. The turnover is large because of their generally amateurish character and short runs. There is no major professional centre for the training of actors, nor a school for stagecraft and production. Notwithstanding, the young directors and playwrights have been enthusiastic about establishing a permanent Urdu stage.

BANGLADESH

East Bengal continued the folk *jātrā* and used this form for themes concerning current political problems and historical events. A successful example of the latter is *Nawab Sirajuddaulah,* which deals with the fall of the last Muslim ruler of Bengal in 1757 through betrayal by his ambitious brother-in-law Mīr Jaʿfar, who joined the British. This *jātrā* is popular with both rural and urban audiences. Tales of Muslim kings and lovers from Persian legends also have been rendered into *jātrās*.

Contemporary theatre inherits the tradition of the prepartition Bengali stage. The poet-playwright Nazrul Islam followed the tradition of Tagore's verse plays and dance operas. Inspired by left-wing ideology, he wrote for the People's Theatre in East Bengal, championing the cause of the poor farmer. He has dealt with psychological problems and inner tensions in his *Shilpi* ("The Artist"), in which the artist is torn between love for his wife and for his art. Especially popular are historical themes of political significance, inspiring Muslims who for centuries were subjugated by the Hindus of East Bengal. Ebrahim Khan wrote *Kamal Pasha* (1926), a play about the Turkish liberator, a symbol of hope and reawakening, and *Anwar Pasha,* about the downfall of Anwar (Enver), who could not cope with the new historical forces.

Bangladesh has a solid acting tradition and a rich repertoire of Bengali plays. Its amateur stage has professional actors, and it retains the impassioned lyricism and power of the mainstream of Bengali tradition. (B.Ga.)

Visual arts

VISUAL ARTS OF THE INDIAN SUBCONTINENT

Indian art is the term commonly used to designate the art of the Indian subcontinent, which includes the present political divisions of India, Kashmir, Pakistan, and Bangladesh. Although a relationship between political history and the history of Indian art before the advent of Islām is at best problematical, a brief review will provide a broad context. The earliest urban culture of the subcontinent is represented by the Indus Valley civilization (*c.* 2500–1800 BC), which possessed several flourishing cities not only in the Indus Valley but also in Gujarāt and Rājasthān. The circumstances in which this culture came to an end are obscure. Although there is no clear proof of historical continuity, scholars have noticed several striking similarities between this early culture and features of later Indian civilization. The period immediately following the urban Indus Valley civilization is marked by a variety of essentially rural cultures. A second urbanization began to occur only around the 6th century BC, when flourishing cities started to reappear, particularly in the Gangetic Basin. The Buddha lived and preached in this period, which culminated in the great Maurya Empire, whose relatively few works are the earliest surviving remnants of monumental art. The Maurya dynast Aśoka (died 238 BC) is considered the greatest of Buddhist kings; and the majority of the monuments of the next 500 years appear to be dedicated to the Buddhist faith, though iconographical and other details suggest that the art also drew heavily on popular religion.

The Maurya Empire spread over almost all of what is modern India and Pakistan. Territories as extensive were never possessed by any other dynasty. With its fall, the empire broke up into a number of states ruled by many dynasties, some of which acquired considerable power and fame for varying periods of time. Among these, the Śuṅgas (*c.* 2nd–1st century BC) in the north and the longer-lived Sātavāhanas in the Deccan and the south are particularly noteworthy. Though these kings were Hindu by religion, Buddhist monuments form the great majority of surviving works.

Toward the end of the 1st century BC, northern India was subjected to a series of invasions by Scythian tribes, resulting finally in the establishment of the vast Kushān (Kuṣāṇa) empire, of which Mathurā was an important centre. The new rulers seemed to have followed Indian faiths, the great emperor Kaniṣka (*c.* AD 78) being a devout Buddhist. The schools of Gandhāra and Mathurā flourished during their rule, and, though much of the work is dedicated to the Buddhist religion, the foundations of later Hindu iconography were also laid in this period. While the Kushān dynasty was sovereign in the north, the Sātavāhanas continued to rule in the south. The bulk of the work at Amarāvatī was produced during their hegemony.

The Golden Age of India

Around the mid-4th century, the Gupta dynasty, of indigenous origin, rapidly expanded its power, uprooting the last remnants of foreign rule and succeeding in bringing almost all of northern India under its sway. In the Deccan there arose at the same time the equally powerful Vākāṭakas, with whom the Guptas appear to have had friendly relations. The period extending from the 4th through the 5th centuries is marked by the most flourishing artistic activities. In addition to the Buddhist monuments, there are the first strong indications of specifically Hindu patronage. Works of remarkable beauty and elegance were produced in this period, which is commonly called the Golden Age of India.

The disintegration of these two empires toward the close of the 5th and the 6th centuries ushered in what has been called the medieval period (*c.* 8th–12th centuries), marked by the appearance of a large number of states and dynasties, often at war with each other. Their rise to power and their decline was part of a constantly recurring process, for none of them was able to hold onto a position of even relative paramountcy for any extended period of time. In the north, the great dynasties were the Gurjara-Pratīhāras, whose empire at its greatest equalled that of the Guptas; the Pālas, who ruled chiefly over northeastern India; and various other dynasties, such as the Kalacuris, the Candelas, and the Paramāras of north central India, the Cāhamānas of Rājasthān, the Cālukyas of Gujarāt. In the Deccan, also, several dynasties rose and fell, the most powerful of which were the Cālukyas of Bādāmi, the Rāṣṭrakūṭas, and the Cālukyas of Kalyāṇī. They were often at war not only with their powerful neighbours to the north but also with the great Pallava and Cōḷa kingdoms of southern India. Most of the dynasties of medieval India were Hindu, though some Jaina and a very few Buddhist kings are also known. The various faiths, however, existed in comparative harmony; and Buddhist and Jaina monuments continued to be built, though most of the surviving works are Hindu.

Although the effects of constant struggle were not as devastating as one might expect, largely as a result of the institutionalization of war and its confinement to appropriate castes, the Hindu kingdoms fell easy prey to the Islāmic invasions, which began as early as the 8th century

AD but gathered strength only in the 11th century. By the end of the 12th century, almost all of northern India had been conquered. Islāmic advances in the south were checked for a while by the Vijayanagara dynasty, but with its collapse almost all of India fell under various degrees of Islāmic hegemony. Large Hindu kingdoms enjoying differing degrees of independence continued to exist chiefly in Rājasthān and portions of southern India, but overall political supremacy was vested with the Islāmic states. The Muslim powers were also divided into many kingdoms, despite attempts made by the sultanate of Delhi, and later by the Mughals, to achieve paramountcy over large portions of India. These attempts were successful only for short periods of time. Although the initial impact of Islām on Indian art was generally destructive, Islāmic influences entering India were gradually transformed in the new environment and eventually resulted in the flowering of an extremely rich and important aspect of the Indian genius.

cenden- of the ıropean wers

The ascendency of the European powers in the 18th century, culminating in the establishment of the British Empire, laid the foundation of modern India's contacts with the West. As a whole, the European advent was marked by a relative insensitivity to native art traditions, but rising nationalism attempted a conscious revival of Indian art toward the end of the 19th century. In modern times, the absorption of European influence is a more natural, freer process that affects artistic development in a vital and profound way.

General characteristics. *The unity of Indian art.* Indian art is spread over a subcontinent and has a long, very productive history; but it nevertheless shows a remarkable unity and consistency. Works produced in the several geographical and cultural regions possess decidedly individual characteristics but at the same time have sufficient elements in common to justify their being considered manifestations of a general style. The existence of this style is evidence of the essential cultural unity of the subcontinent and to the uninterrupted contact between the various geographical units, at least from the historical period onward. Developments in one area have been quickly reflected in the others. The regional idioms have contributed to the richness of Indian art, and the mutual influences exercised by them have been responsible for the multi-faceted development of that art throughout the course of its long life.

The style of Indian art is largely determined not by a dynasty but by conditions of time and space. It has, essentially, a geographical rather than a dynastic basis, which is to say that the evolution of regional schools appears to have been largely independent of any particular dynasty that happened to rule over a specific region. The style does not change because of the conquest of one area by another dynasty; rather the influences exercised by one area on another are usually through the agency of factors other than conquest. Instances in which dynastic patronage changed the nature of a style are very few and confined mostly to the Islāmic period. The political history of India is itself quite vague, and the areas in possession of a dynasty at various points in its history are even less susceptible to precise definition. For all these reasons, the classification of Indian art adopted here is not based on dynasties, for such a division has little meaning. Nevertheless, names of certain dynasties are used, for these have passed into common usage. When this is done, however, the name must be understood as little more than a convenient way of labelling a particular period.

The materials of Indian art. Indian art employs various materials, such as wood, brick, clay, stone, and metal. Most wooden monuments of the early period have perished but have been imitated in stone. Clay and brick were also abundantly used; but, particularly in later times, the favoured material seems to have been stone, in the dressing (facing and smoothing) and carving of which the Indian artist attained great excellence. The material may have influenced the form somewhat, but essentially Indian art tends to impose the form on the material. Thus, materials are generally regarded as interchangeable: wooden and clay forms are imitated in stone, and stone is imitated in bronze, and in turn stone sculpture assumes qualities appropriate to metal. It is as though the nature of the material presented a challenge that had to be met and overcome. At the same time, Indian art stresses the plasticity of forms; sculpture is generally characterized by emphatic mass and volume; architecture is often sculpture on a colossal scale; and the elements of painting, particularly of the early period, are modelled by line and colour.

Emphasis on the plasticity of forms

Indian and foreign art. Thanks to its geographical situation, the Indian subcontinent has been constantly fed by artistic traditions emanating from West and Central Asia. The Indian artist has shown a remarkable capacity for accepting these foreign influences naturally and assimilating and transforming them to accord with the nature of his own style. The process occurred frequently: in the Maurya period; in the two centuries after Christ, when the Kushān dynasty attained imperial supremacy in the north; and at a much later period, in the 16th century, when the Mughals patronized a new school of architecture and painting.

Just as India received influences, so it transmitted its own art abroad, particularly to Ceylon and the countries of Southeast Asia. Developments of great importance were therby precipitated in Ceylon, Burma, Thailand, Indonesia, and Indochina, where the reinterpretation of Indian influences resulted in the creation of works of great originality.

Indian art and religion. Indian art is religious inasmuch as it is largely dedicated to the service of one of several great religions. It may be didactic or edificatory as is the relief sculpture of the two centuries before and after Christ; or, by representing the divinity in symbolic form (whether architectural or figural), its purpose may be to induce contemplation and thereby put the worshipper in communication with the divine. Not all Indian art, however, is purely religious, and some of it is only nominally so. There were periods when humanistic currents flowed strongly under the guise of edificatory or contemplative imagery, the art inspired by and delighting in the life of this world.

Although Indian art is religious, there is no such thing as a sectarian Hindu or Buddhist art, for style is a function of time and place and not of religion. Thus it is not strictly correct to speak of Hindu or Buddhist art, but, rather, of Indian art that happens to render Hindu or Buddhist themes. For example, an image of Vishnu and an image of Buddha of the same period are stylistically the same, religion having little to do with the mode of artistic expression. Nor should this be surprising in view of the fact that the artists belonged to nondenominational guilds, ready to lend their services to any patron, whether Hindu, Buddhist, or Jaina.

Religion, symbolism, and abstraction

The religious nature of Indian art accounts to some extent for its essentially symbolic and abstract nature. It scrupulously avoids illusionistic effects, evoked by imitation of the physical and ephemeral world of the senses; instead, objects are made in imitation of ideal, divine prototypes, whose source is the inner world of the mind. This attitude may account for the relative absence of portraiture and for the fact that, even when it is attempted, the emphasis is on the ideal person behind the human lineaments rather than on the physical likeness.

The artist and patron. Works of art in India were produced by artists at the behest of a patron, who might commission an object to worship for spiritual or material ends, in fulfillment of a vow, for the discharge of virtues enjoined by scripture, or even for personal glory. Once the artist received his commission, he fashioned the work of art according to his skill, gained by apprenticeship, and the written canons of his art, which possessed a holy character. There were prescribed rules for proportionate measurement, iconography, and the like, often with a symbolic significance. This is not to say that the individual artist was invariably aware of the symbolic meaning of the prescribed standards, based as these were on complex metaphysical and theological considerations; but the symbolism nevertheless formed part of the fabric of his work, ready to add an extra dimension of meaning to the initiated and knowledgeable spectator.

In these conditions it is not surprising that the artist as a person is for the most part anonymous, very few names of artists having survived. It was the skill with which the work of art was made to conform to established ideals, rather than the artist who possessed the skill, that held the place of first importance.

The appreciation of Indian art. According to Indian aesthetic theory, a work of art possesses distinct "flavours" (*rasa*), the "tasting" of which constitutes the aesthetic experience. Because the work of art operates at various levels, granting to the spectator what he is capable of receiving by virtue of his intellectual and emotional preparation, the appreciation of the beauty of form and line is considered an appropriate activity of the educated and cultured man. The supreme aesthetic experience, however, is believed to be much deeper and cognate to the experience of the Godhead. From this point of view, the work of art is in a sense irrelevant and unnecessary for a person at a high level of spiritual progress; and for the devout layman its excellence is measured by its efficacy in promoting spiritual development.

Architecture. The favoured material of early Indian architecture appears to have been wood, but little has survived the rigours of the climate. Wooden forms, however, affected work in other mediums and were sometimes quite literally copied, as, for example, in early cave temples of western India. The principles of wooden construction also played an important part in determining the shape of Indian architecture and its various elements and components.

Effect of wooden forms on other mediums

Baked or sun-dried brick has a history as ancient as that of wood; among the earliest remains are buildings excavated at sites of the Indus Valley civilization. The use of brick is once again evident from about the 6th century BC, and its popularity was undiminished in subsequent centuries. Many brick monuments have been discovered, particularly in areas in which good clay was easily available, such as the Gangetic Basin. Although more durable than wood, few brick buildings from before the 5th century AD have survived in a good state of preservation.

Traditions of stone architecture appear to be more recent than wood or brick, the earliest examples of the use of dressed stone for building purposes not predating the 6th century BC. The Indian architect, however, soon gained great proficiency in its use, and, by the 7th century AD, the use of stone for monumental buildings of considerable size had become quite popular. The preference for stone can also be seen in Islāmic monuments of India, which contrast markedly with the brick and tile structures popular in neighbouring West Asia.

Most surviving examples of Indian architecture before the Islāmic period are of a religious nature, consisting mainly of Buddhist shrines, or *stūpa*s, and temples. Monastic residences give some idea of civil architecture, but, surprisingly, very few examples of palaces and secular dwellings have been found.

Indus Valley civilization (c. 2500–1800 BC). From excavated remains, it is clear that the Indus Valley civilization possessed a flourishing urban architecture. The major cities associated with the civilization, notably Mohenjo-daro, Harappā, and Kalibangan, were laid out on a grid pattern and had provisions for an advanced drainage system. The residential buildings, which were serviceable enough, were mainly brick and consisted of an open patio flanked by rooms. For monumental architecture, the evidence is slight, the most important being a "sacred" tank (thought to be for ritual ablution) and associated structures. Corbel vaulting (arches supported by brackets projecting from the wall) was known, and, to a limited extent, timber was used together with brick; whatever architectural ornamentation existed must have been of brick or plaster.

The Maurya period (c. 321–185 BC). The state of Indian architecture in the period between the Indus Valley civilization and the rise of the Maurya Empire is largely unknown since most work was done in such perishable material as wood or brick. Excavations at Rājgīr, Kauśāmbī, and other sites, however, testify to the existence of fortified cities with *stūpa*s, monasteries, and temples of the type found at the later Maurya sites of Nagarī and Vidiśā; and there is evidence of the use of dressed stone in a palace excavated at Kauśāmbī. Considering the power of the Maurya Empire and the extensive territory it controlled, the architectural remains are remarkably few. The most important are *stūpa*s (later enlarged) such as a famous example of Sānchi; the ruins of a hall excavated at the site of Kumrāhar in Patna (ancient Pāṭaliputra), the capital city; and a series of rock-cut caves in the Barābar and Nāgārjunī hills near Gayā, which are interesting because they preserve in the more permanent rock some types of wooden buildings popular at that time.

The *stūpa*

The *stūpa,* the most typical monument of the Buddhist faith, consists essentially of a domical mound in which sacred relics are enshrined. Its origins are traced to mounds, or tumuli, raised over the buried remains of the dead that were found in India even before the rise of Buddhism: *Stūpa*s appear to have had a regular architectural form in the Maurya period: the mound was sometimes provided with a parasol surrounded by a miniature railing on the top, raised on a terrace, and the whole surrounded by a large railing consisting of posts, crossbars, and a coping (the capping on the top course), all secured by tenons and mortices in a technique appropriate to craftsmanship in wood. The essential feature of the *stūpa,* however, always remained the domical mound, the other elements being optional.

Along with *stūpa*s were erected roofless, or hypaethral, shrines enclosing a sacred object such as a tree or an altar. Temples of brick and timber with vaulted or domical roofs were also constructed, on plans that were generally elliptical, circular, quadrilateral, or apsidal (*i.e.,* having an apse, or semicircular plan, at the sanctum end). These structures have not survived, but some idea of their shape has been obtained from the excavated foundations and the few examples imitating wooden originals that were cut into the rock, notably the Sudāmā and the Lomas Ṛṣi caves in the Nāgārjunī and Barābar hills near Gayā. The latter has an intersesting entrance showing an edged barrel-vault roof (an arch shaped like a half cylinder) in profile supported on raked pillars, the ogee arch (an arch with curving sides, concave above and convex toward the top) so formed filled with a trellis to let in light and air. The interiors of most caves are highly polished and consist of two chambers: a shrine, elliptical or circular in plan with a domed roof (Sudāmā cave); and an adjacent antechamber, roughly rectangular and provided with a barrel vault. Remains of structural buildings have been excavated at Bairāt and Vidiśā, where wood and brick shrines with timber domes and vaults once existed. A temple (No. 40) at Sānchi was apsidal in plan and perhaps had a barrel-vault roof of timber.

A hall excavated at Kumrāhar in Patna had a high wooden platform of most excellent workmanship, on which stood eight rows of 10 columns each, which once supported a second story. Only one stone pillar has been recovered, and it is circular in shape and made of sandstone that has been polished to a high lustre. The capitals that topped them must have been similar to others found in neighbouring Lohanipur and almost certainly consisted of one or two pairs of addorsed (set back to back) animals, recalling Persepolitan examples. Indeed, there is much about Maurya architecture and sculpture to suggest Iranian influence, however substantially transformed in the Indian environment.

Early Indian architecture (2nd century BC–3rd century AD). Except for *stūpa*s, architectural remains from the 2nd century BC (downfall of the Maurya dynasty) to the 4th century AD (rise of the Gupta dynasty) continue to be rare, indicating that most of the work was done in brick and timber. Once again, examples cut into the rock and closely imitating wooden forms give a fairly accurate idea of at least some types of buildings in this period.

The *stūpa*s become progressively larger and more elaborate. The railings continue to imitate wooden construction and are often profusely carved, as at Bhārhut, Sānchi II, and Amarāvatī. These were also provided with elaborate gateways, consisting of posts supporting from one to three architraves, again imitating wooden forms and covered with sculpture (Bhārahut, Sānchi I, III). In the course of time an attempt was made to give height to the *stūpa*s by multiplying the terraces that supported the dome and by increasing the number of parasols on top. In Gandhāra and southeastern India, particularly, sculptured decoration was extended to the *stūpa* proper, so that terraces, drums, and domes—as well as railing—were decorated with figural and ornamental sculpture in bas-relief. *Stūpa*s in Gandhāra were not provided with railings but, instead, had rows of small temples arranged on a rectangular plan.

Vest ndian :ave emples

Cave temples of western India, cut into the scarp of the Western Ghāts and stretching from Gujarāt to southern Mahārāshtra, constitute the most extensive architectural remains of the period. Two main types of buildings can be distinguished, the temple proper (*caitya*) and the monastery (*vihāra, saṅghārāma*). The former is generally an apsidal hall with a central nave flanked by aisles. The apse is covered by a half dome; and two rows of pillars, which demarcate the nave, support a barrel-vault roof that covers the rest of the building. In the apsidal end is placed the object to be worshipped, generally a *stūpa,* the hall being meant for the gathered congregation. In front of the hall is a porch, separated from it by a screen wall provided with a door of considerable size, together with an arched opening on top clearly derived from wooden buildings of the Lomas Ṛṣi type and permitting air and dim light to filter into the interior. Other influences of wooden construction are equally striking, particularly in the vaulting ribs that cover the entire ceiling and that are sometimes actually of wood, as at Bhājā, where the pillars are also raked in imitation of the exigencies of wooden construction. The pillars are generally octagonal with a pot-shaped base and a capital of addorsed animals placed on a bell-shaped, or campaniform, lotus in the Maurya tradition. The most significent example is at Kārli, dating approximately to the closing years of the 1st century BC. The Bhājā *caitya* is certainly the earliest, and important examples are to be found at Beḍsā, Kondane, Pītalkhorā, Ajantā, and Nāsik. Toward the end of the period, a quadrilateral plan appears more and more frequently, as, for example, at Kuda and Sailarwāḍī.

In addition to the *caitya,* or temple proper, numerous monasteries (*vihāra*s) are also cut into the rock. These are generally provided with a pillared porch and a screen wall pierced with doorways leading into the interior, which consists of a "courtyard" or congregation hall in the three walls of which are the monks' cells. The surviving rock-cut examples are all of one story, though the facade of the great monastery at Pitalkhorā simulates a building of several stories.

Monasteries carved into the rock are also known from Orissa (Udayagiri-Khandagiri), in eastern India. These are much humbler than their counterparts in western India, and consist of a row of cells that open out into a porch, the hall being absent. At Uparkot in Junāgadh, Gujarāt, is a remarkable series of rock-cut structures dating from the 3rd–4th century AD, which appear to be secular in character and in all probability served as royal pleasure houses.

The large number of representations of buildings found on relief sculpture from sites such as Bhārhut, Sānchi, Mathurā, and Amarāvatī are a rich source of information about early Indian architecture. They depict walled and moated cities with massive gates, elaborate multi-storied residences, pavilions with a variety of domes, together with the simple, thatched-roofed huts that remained the basis of most Indian architectural forms. A striking feature of this early Indian architecture is the consistent and profuse use of arched windows and doors, which are extremely important elements of the architectural decor.

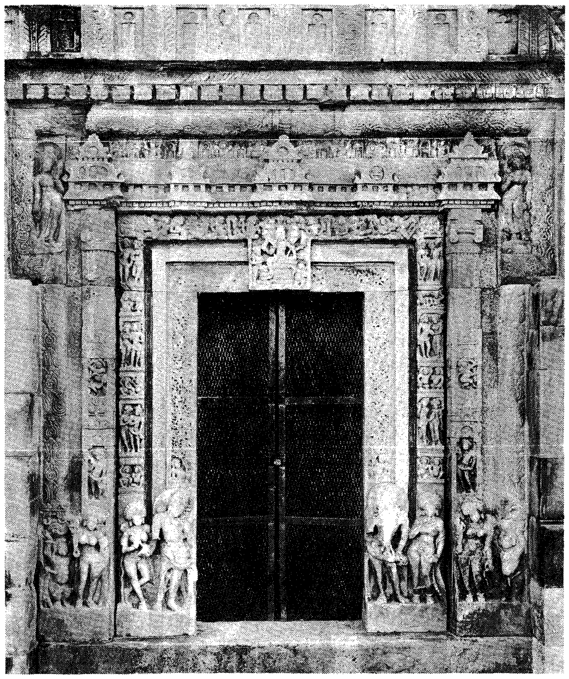
The Gupta period (4th–6th centuries AD). Dating toward the close of the 4th and the beginning of the 5th century AD is a series of temples that marks the opening phase of an architecture that is no longer content with merely imitating wooden building but initiates a new movement, ultimately leading to the great and elaborate temples of the 8th century onward.

Two main temple types have been distinguished in the Gupta period. The first consists of a square, dark sanctum with a small, pillared porch in front, both covered with flat roofs. This type of temple answers the simplest needs of worship, a chamber to house the deity and a roof to shelter the devotee. Temple No. 17 at Sānchi is a classic example of this flat-roofed type. The plain walls are of ashlar masonry (made up of squared stone blocks), composed of sizable blocks, which are spanned by large slabs that constitute the ceiling. The pillars of the porch have a campaniform lotus capital, one of the last times this form appears in Indian architecture. Another temple of this type is the Kaṅkālī Devī shrine at Tigowā, which has more elaborate pillars, provided with the overflowing vase, or the vase-and-foliage (*ghaṭa-pallava*), capital that became the basic north Indian order.

Two main temple types of the Gupta period

It is the second type of temple that points the way to future developments. It also has a square sanctum, or cella, but instead of a flat roof there is a pyramidal superstructure (*śikhara*). Among the most interesting examples are a brick temple at Bhītargaon and the Vishnu temple at Deogarh, built entirely of stone. The pyramidal superstructure of each consists essentially of piled-up cornice moldings of diminishing size, which are decorated primarily with *candraśālā* (ogee arch) ornament derived from the arched windows and doors so frequently found in the centuries immediately before and after Christ. The sanctums of both temples are square in plan, with three sides provided with central offsets (vertical buttress-like projections) that extend from the base of the walls right up to the top of the *śikhara* (spire); the section of the central offset that extends across the wall is conceived in the form of a niche, in which is placed an image. The Deogarh temple is also noteworthy for the large terrace with four corner shrines (now ruined) on which it is placed, prefiguring the quincunx, or *pañcāyatana,* grouping (one structure in each corner and one in the middle) popular in the later period. The doorway surround, too, is very elaborate, carved with several bands carrying floral and figural motifs. At the base of the surround are rows of worshippers, and in the crossette (projection at the corner) on top are images of graceful river goddesses.

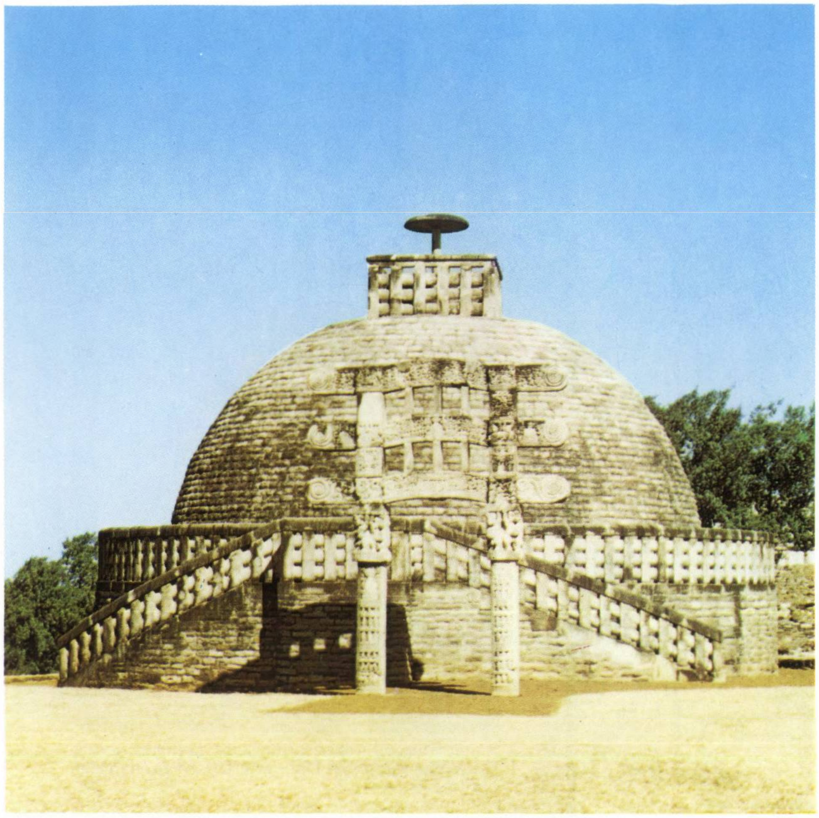
P. Chandra



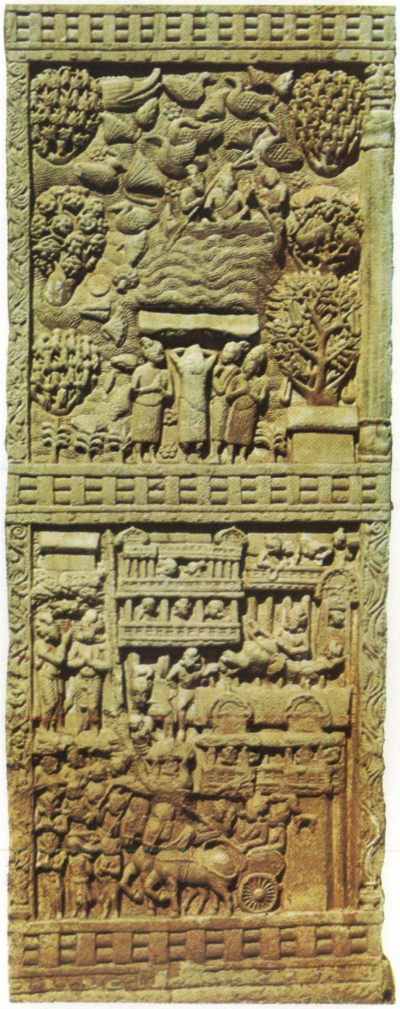
Carved sandstone doorframe of the Vishnu temple at Deogarh, Uttar Pradesh, India, 5th century AD.

The Pārvatī Devī temple at Nācnā Kuṭṭhārā, also of this period, is interesting for the covered circumambulatory provided around the sanctum and the large hall in front. When first discovered, the temple had an entire chamber above the sanctum (which subsequently collapsed). Though provided with a door, there seems to have been no access to it; thus, for all practical purposes it constituted a false story and, aside from a symbolic meaning,

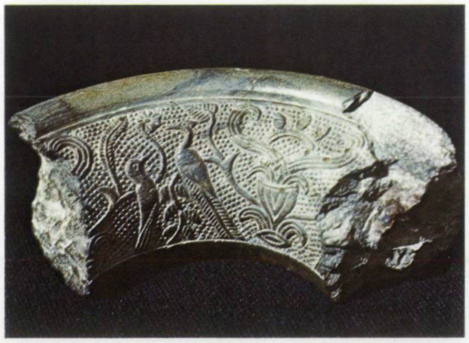
Early Indian architecture and sculpture



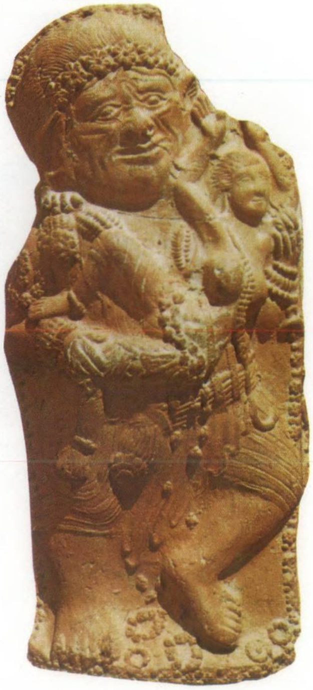
Stūpa III at Sānchi, Madhya Pradesh, sandstone, 1st century BC.



Detail of a sandstone relief sculpture from the *torana* of Stūpa I at Sānchi, Madhya Pradesh, 1st century BC.

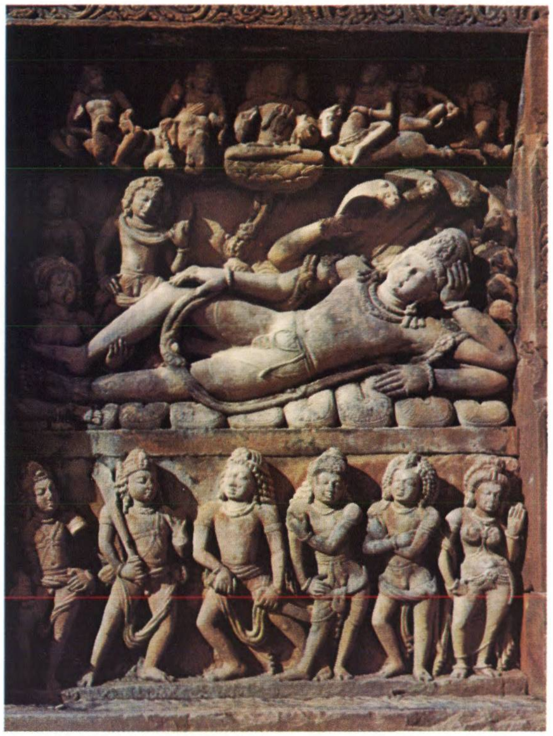


Fragment of a soapstone disc from Kausāmbī, Uttar Pradesh, 3rd century BC. In the Municipal Museum, Allāhābād, Uttar Pradesh. Height 5.7 cm.

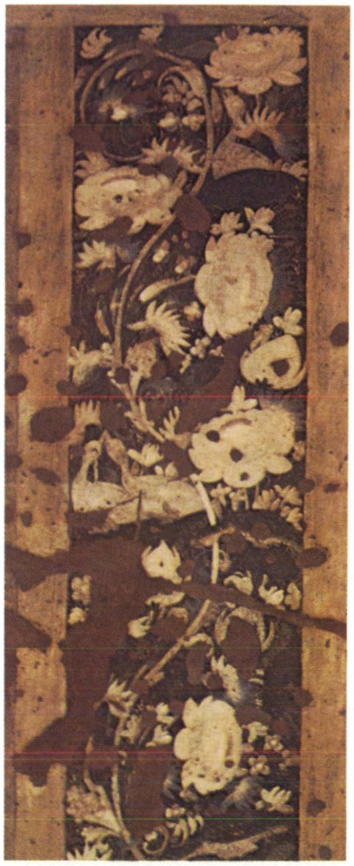


Abduction scene, terra-cotta relief from Kausāmbī, Uttar Pradesh, *c.* 2nd century BC. In the Municipal Museum, Allāhābād, Uttar Pradesh.

Plates 1–8: P. Chandra

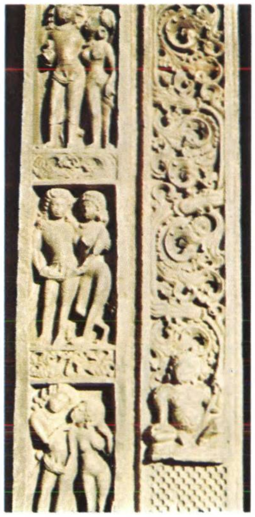


Viṣṇu lying on the serpent Sesa; sandstone relief panel on the Viṣṇu temple at Deogarh, Uttar Pradesh, 5th century AD.

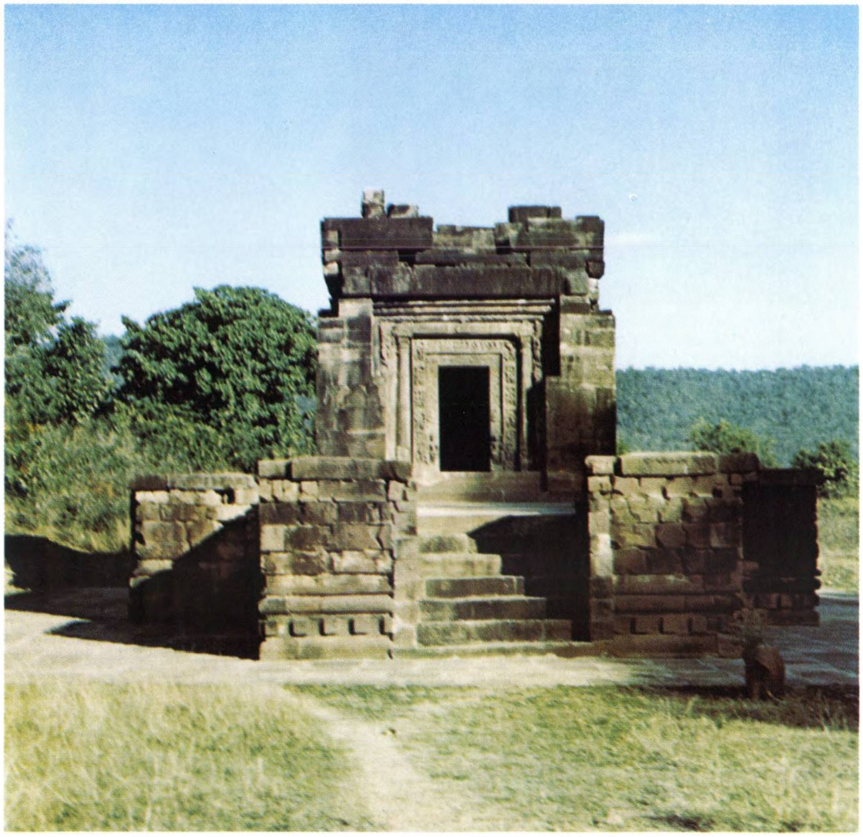


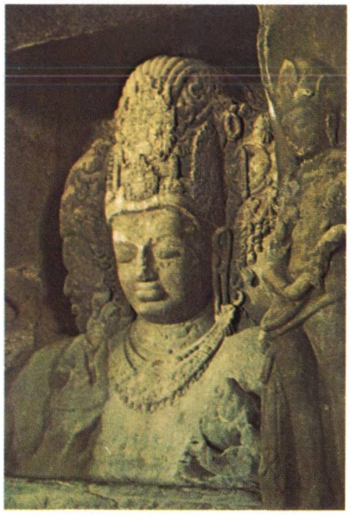
Lotus scroll painted on the ceiling of Cave 2 at Ajanta, Mahārāshtra, 5th century AD.

Indian art of the Golden Age: 5th to 6th centuries AD



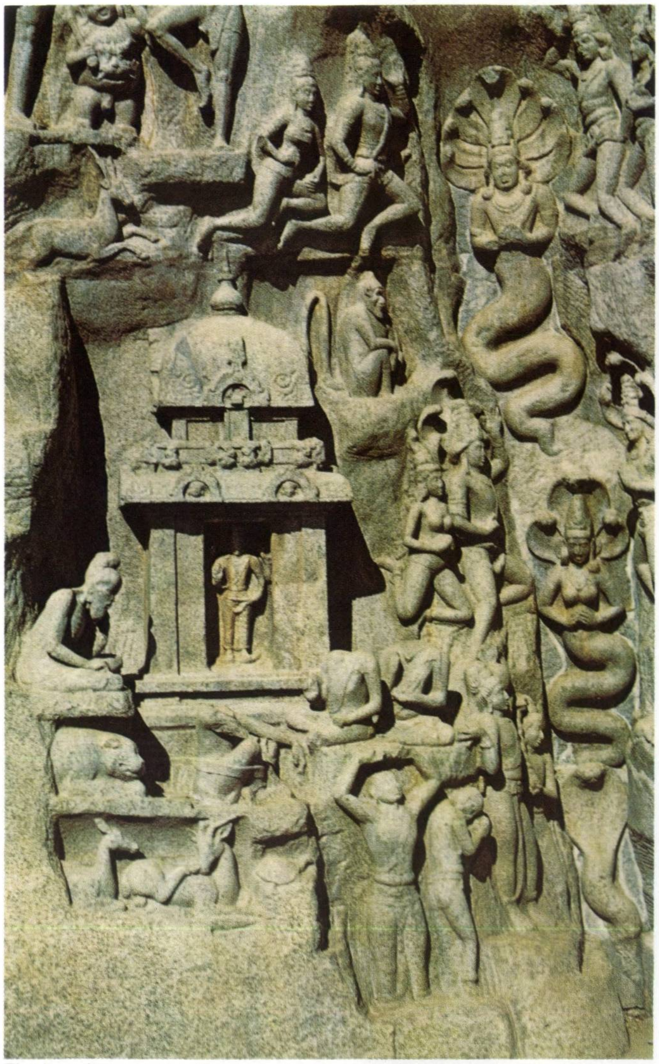
(Below) The Pārvatī temple at Nācnā Kuṭhārā, Madhya Pradesh, sandstone, 5th century AD. (Left) Amorous couples and lotus scroll, detail of the doorframe of the Pārvatī temple.



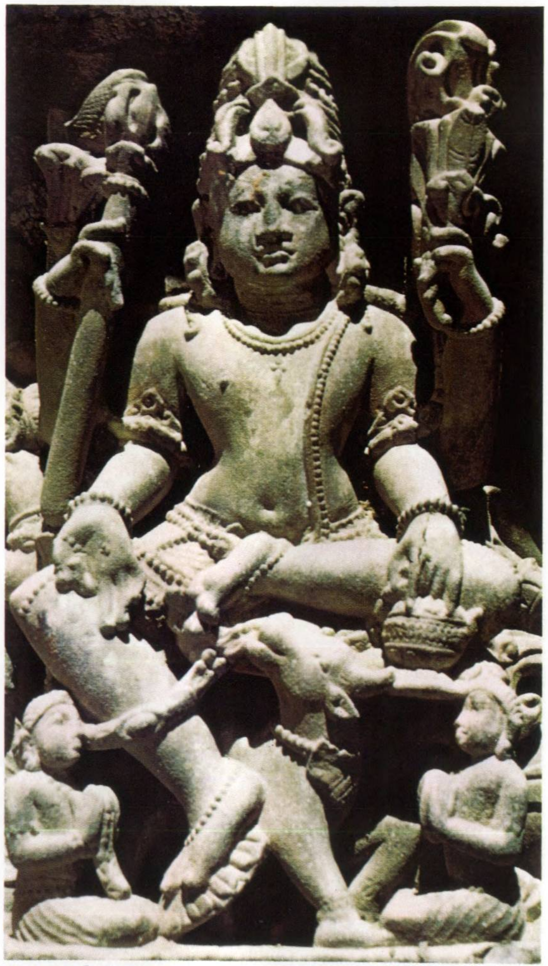


Śiva-Mahesamūrti, the main image in the cave Temple at Elephanta, Mahārāshtra, 6th century AD.

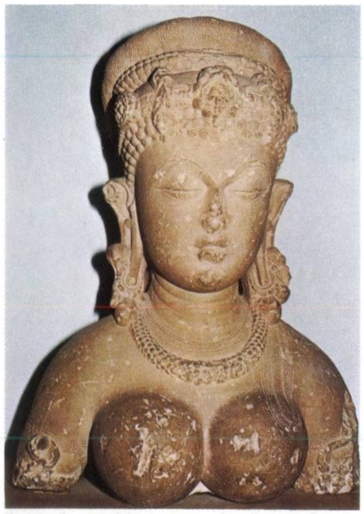
Medieval sculpture of North and South India



The hermitage by the Ganges, detail of a granite relief depicting the penance of Arjuna, from Mahābalipuram, Tamil Nadu, early 7th century AD.

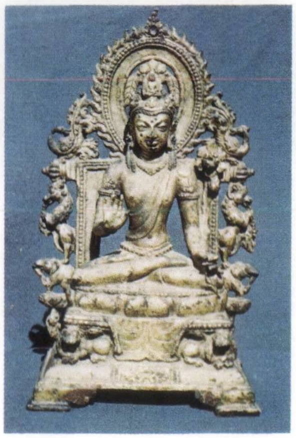


The god Śiva, detail of a doorframe from Singhpur, Madhya Pradesh, sandstone, 10th century AD.

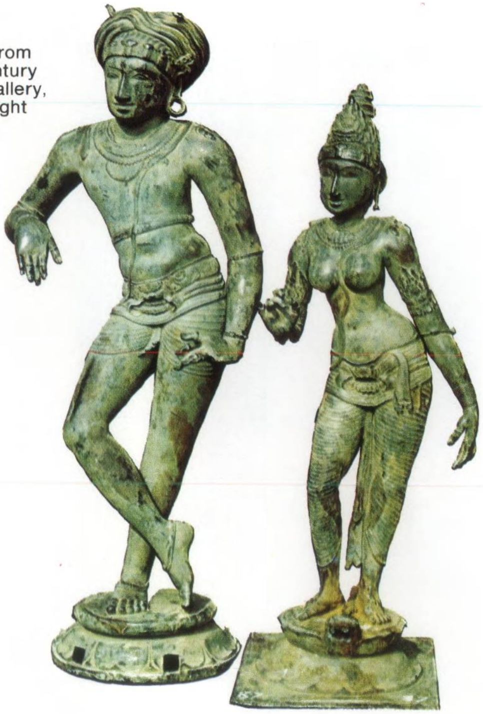


Bust of a goddess from the fort at Gwalior, Madhya Pradesh, *c.* 9th century AD. In the Gwalior Museum, Madhya Pradesh. Height 54 cm.

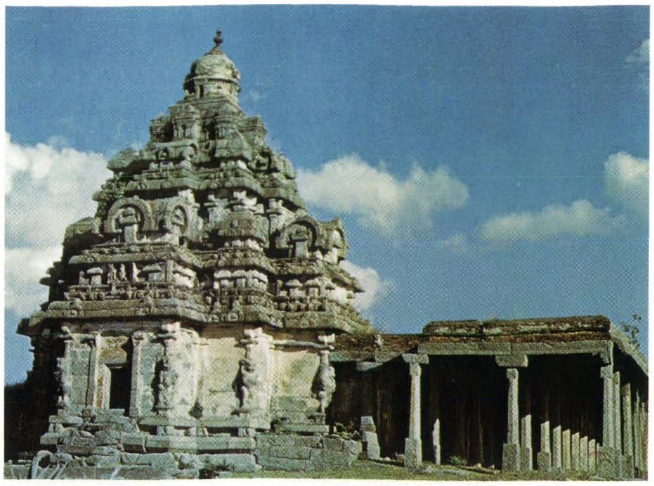
Seated Avalokiteśvara, gilt bronze sculpture from Nālandā, Bihār, 8th century AD. In the Nālandā Museum, Bihār. Height 28 cm.



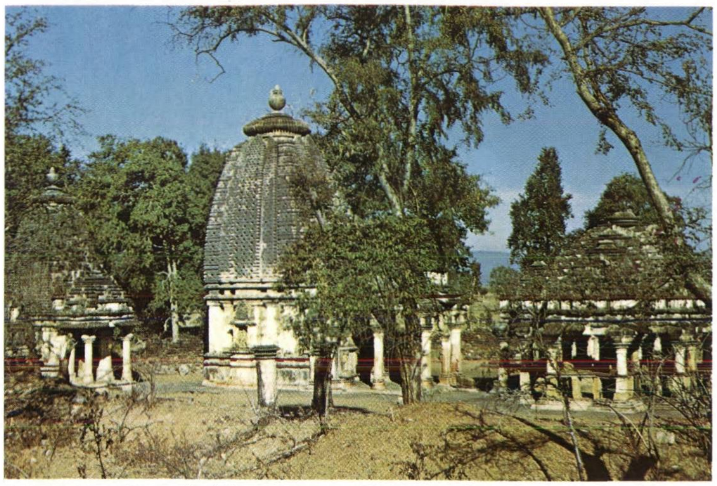
Śiva and his consort, bronze sculpture from Tiruvenkadu, Tamil Nadu, early 11th century AD. In the Thanjāvūr Museum and Art Gallery, Tamil Nadu. Height of male 106 cm., height of female 94 cm.



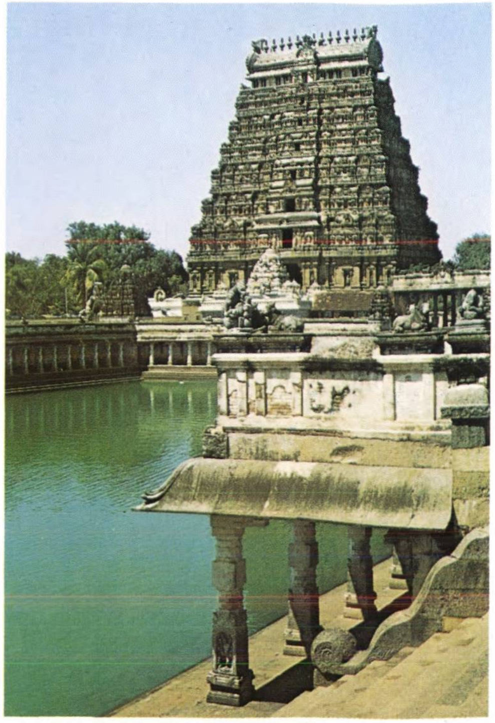
Medieval temple architecture: styles of North and South India



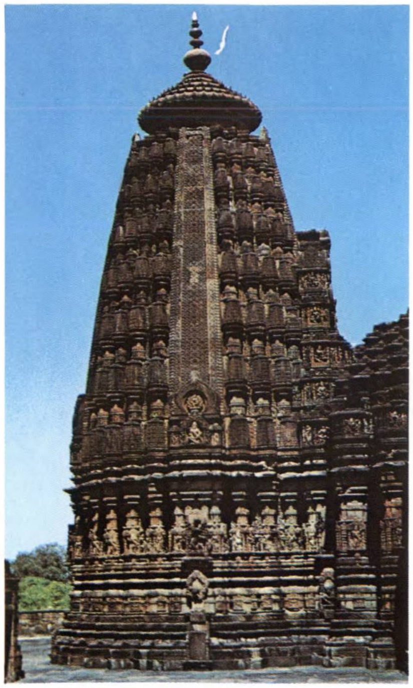
Temple with *kutina* superstructure, the Tālapurīśvara at Panamalai, Tamil Nadu, early 8th century AD.



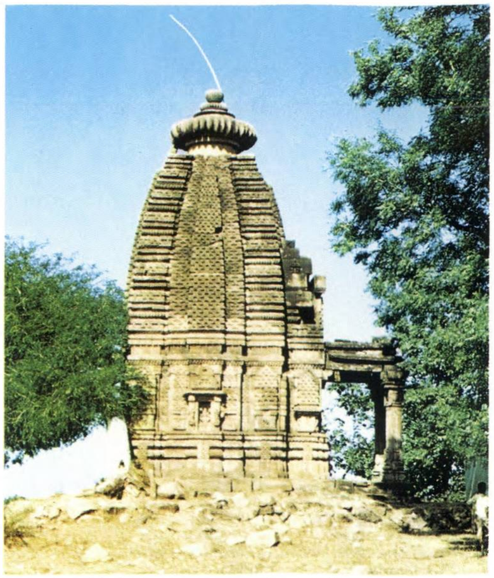
Group of temples at Badoli, Rājasthān, 10th century AD.



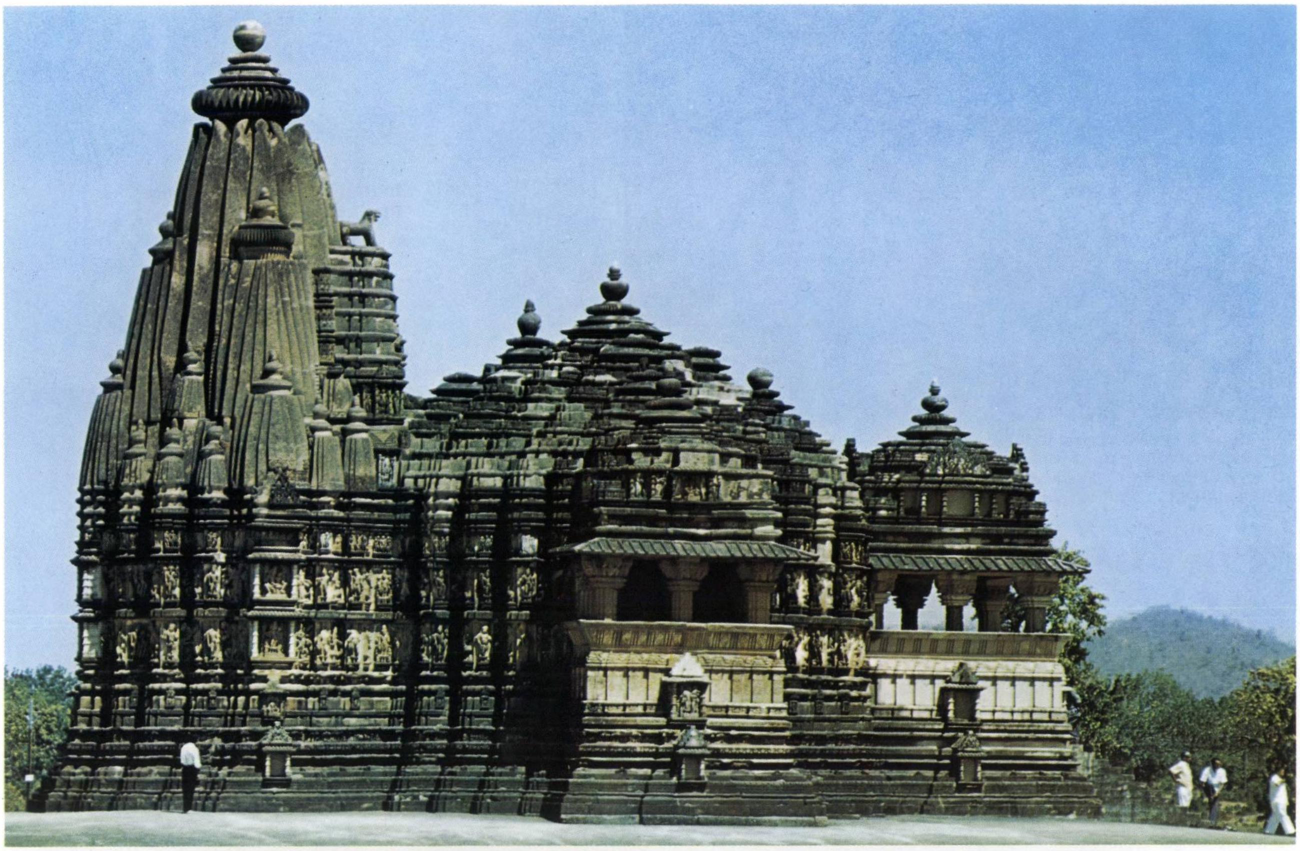
Temples, tank, and *gopura* of the Śiva temple at Cidambaram, Tamil Nadu, 12th–13th century AD.

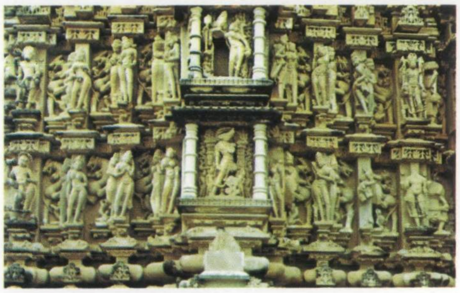


Temple with *bhūmija* superstructure, the Udayaeśvara (or Nīlakaṇṭhhaeśvara) at Udayapur, Madhya Pradesh, c. AD 1059–82.

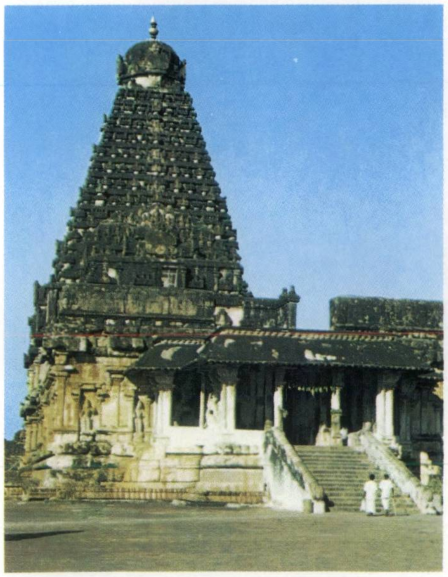


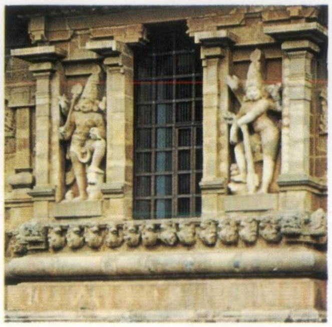
Temple with *latina* superstructure at Umrī, Madhya Pradesh, 9th century AD.



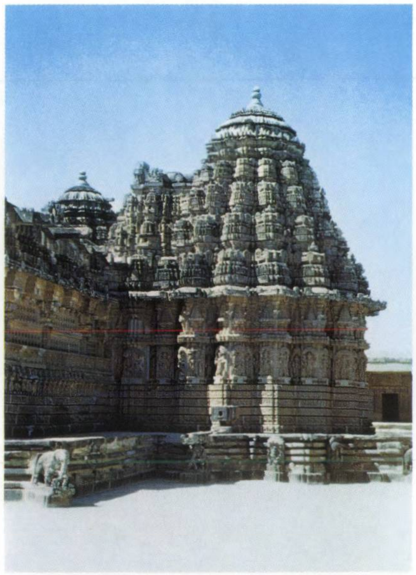


(Above) The Citragupta temple at Khajurāho, Madhya Pradesh, sandstone, 11th century AD. (Left) Detail of the temple wall.



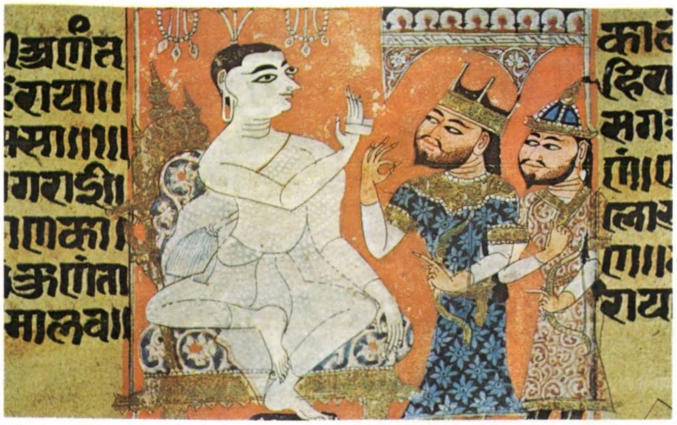


(Left) The Brhadīśvara temple at Thanjāvūr, Tamil Nadu, early 11th century AD. (Right) Detail of the temple wall.

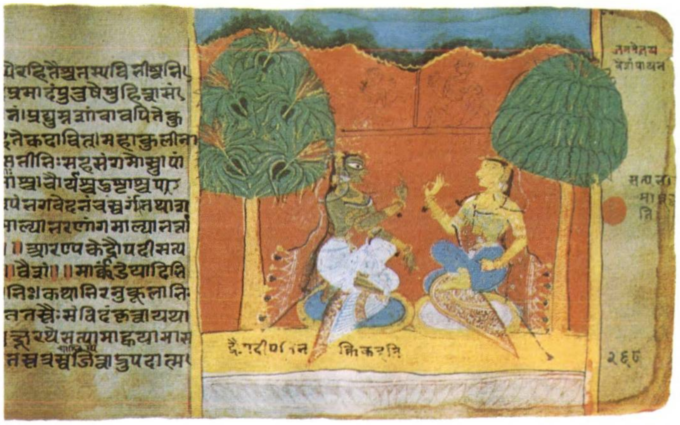


The Keśava temple at Somnāthpur, Karnataka, c. AD 1268.

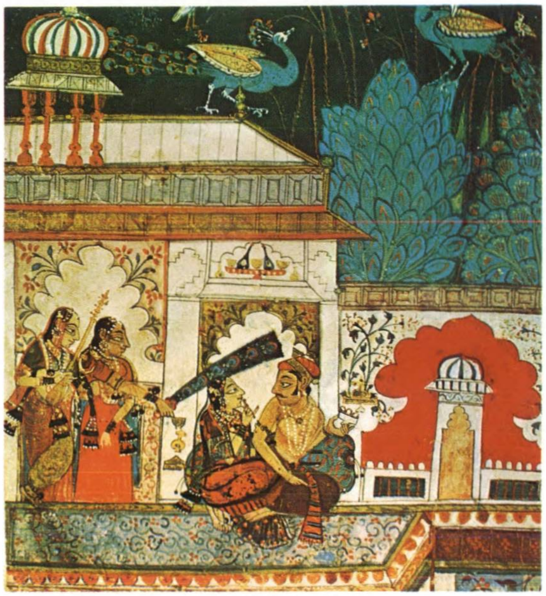
Indian miniature painting: the indigenous tradition



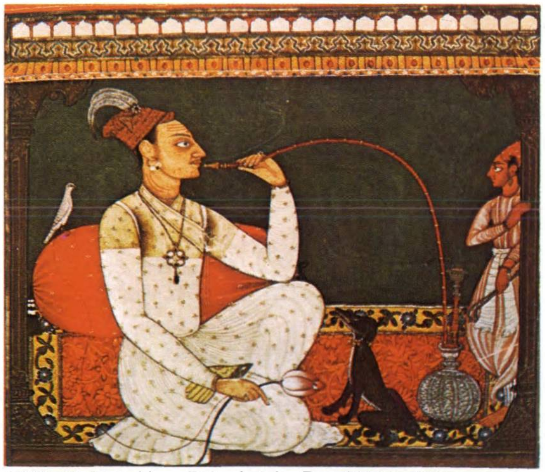
The monk Kālaka addressing the Sāhi king, detail from a folio from a *Kālakācāryakathā* manuscript, Western Indian style, late 13th century AD. In the Prince of Wales Museum of Western India, Bombay. Dimensions of miniature 7.6 × 7.6 cm.



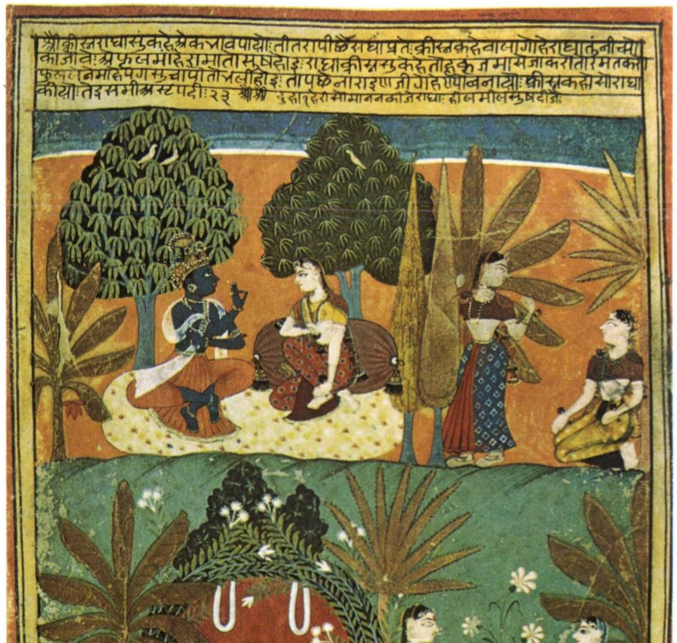
Ladies in conversation, detail from a folio from a *Mahābhārata* manuscript, AD 1516. In the collection of the Asiatic Society of Bombay. Dimensions of miniature 10.2 × 10.2 cm.



A prince and his lady, Rajasthani style, Mālwa, mid-17th century AD. In a private collection.

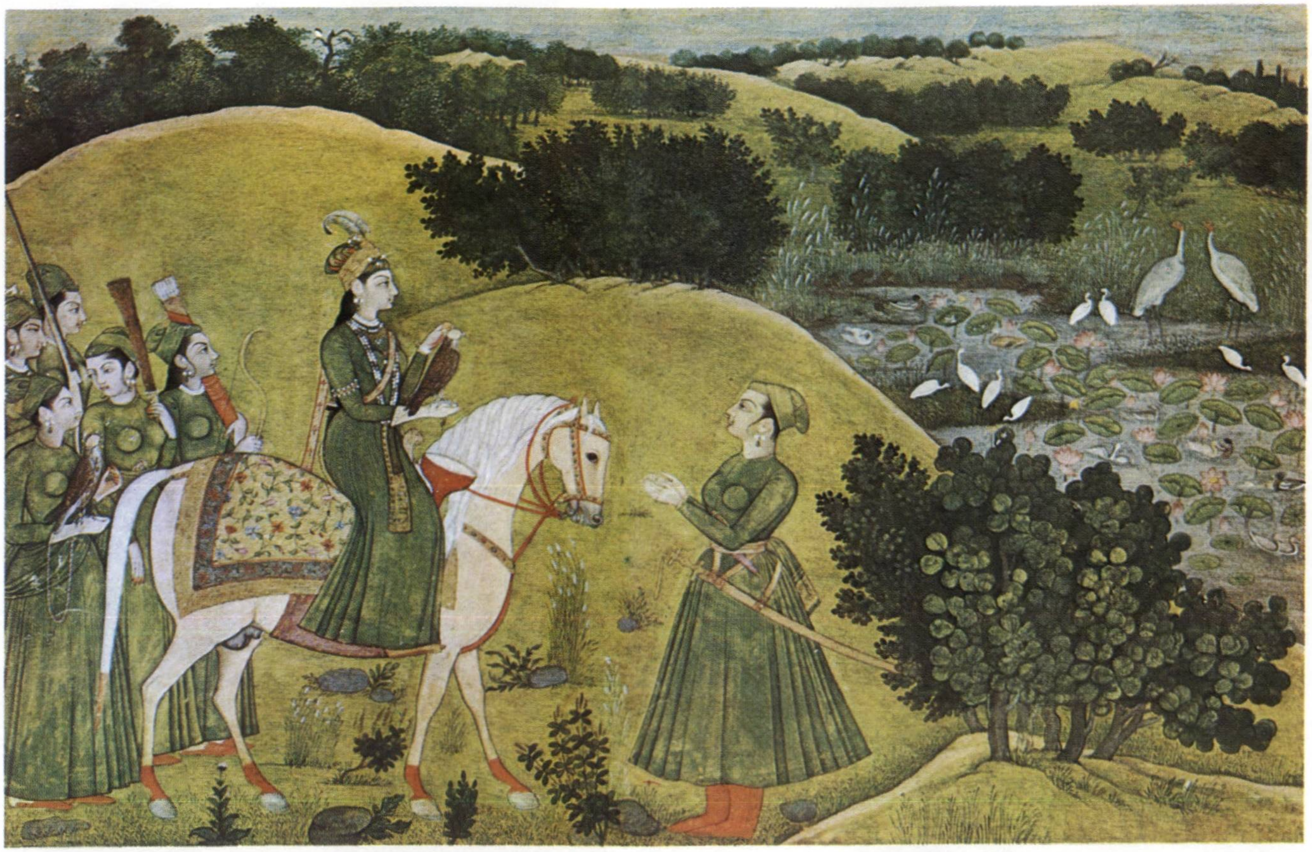


A hill chief smoking, Pahari style, Basohlī, late 17th century AD. In the National Museum of India, New Delhi.

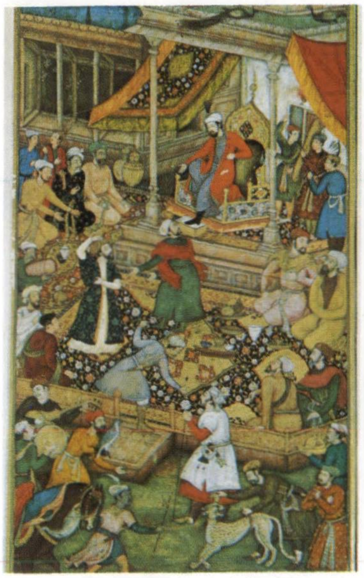


Kṛṣṇa and Rādhā, miniature from a series illustrating the *Gītagovinda,* Rajasthani style, Mewār, mid-17th century AD. In a private collection.

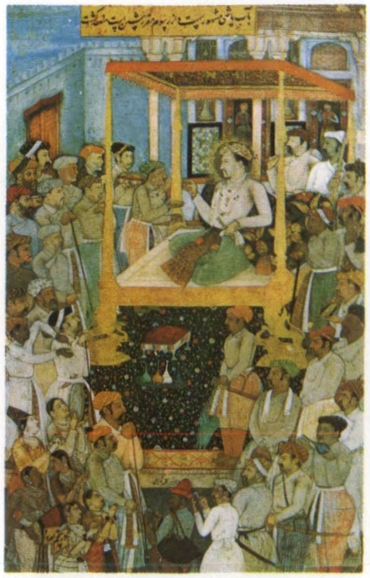
The Mughal style and its influences



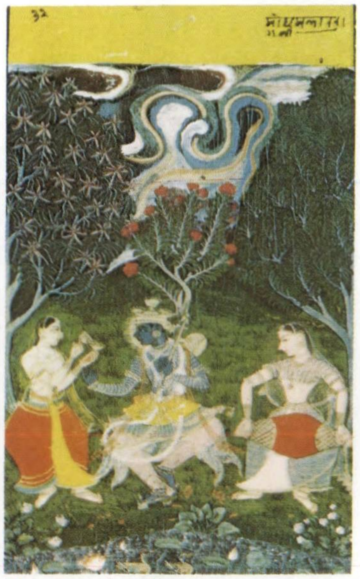
A queen on a hunting expedition, Pahari style, Kāngra, *c.* AD 1760. In the National Museum of India, New Delhi.



Court scene by Basāvan, folio from an illustrated manuscript of the *Anwār-e Suhayli,* Mughal style, AD 1596–97. In the Bharat Kala Bhavan, Vārānasi, Uttar Pradesh. 24.8 × 13.9 cm.

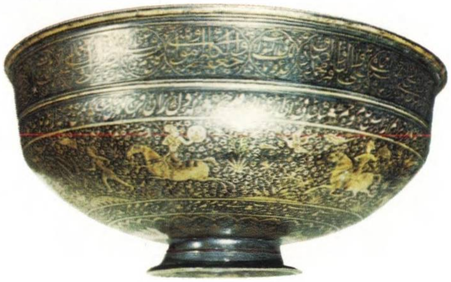


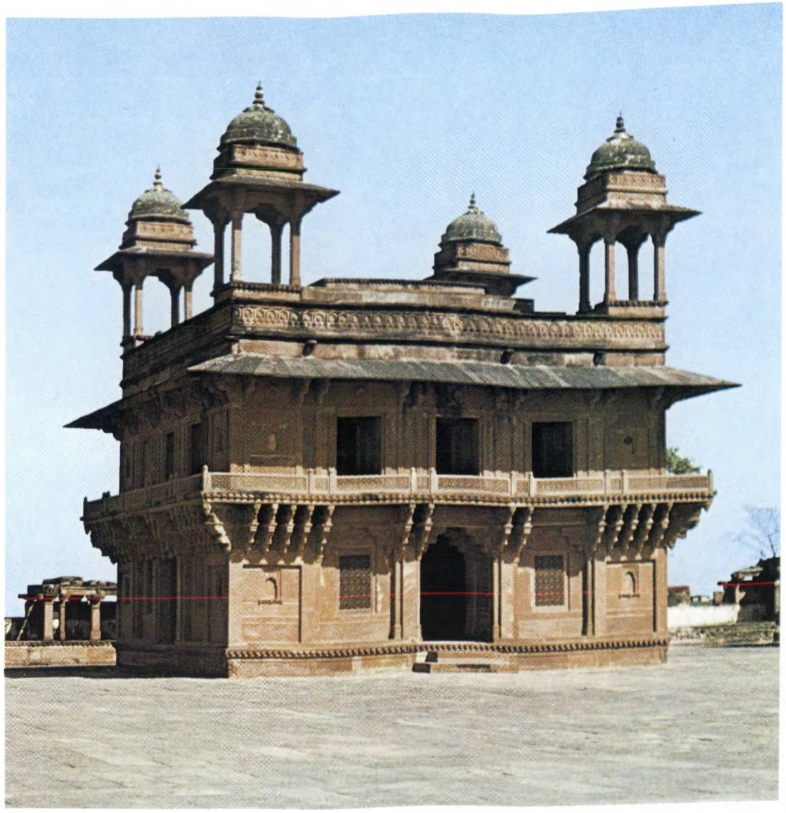
The feast of Nauroz at Jahāngīr's court, Mughal style, *c.* AD 1615. In the collection of the Reza Library, Rāmpur, Uttar Pradesh.



The musical mode Megha-Malāra, Rajasthani style, Būndi, late 17th century AD. In a private collection.

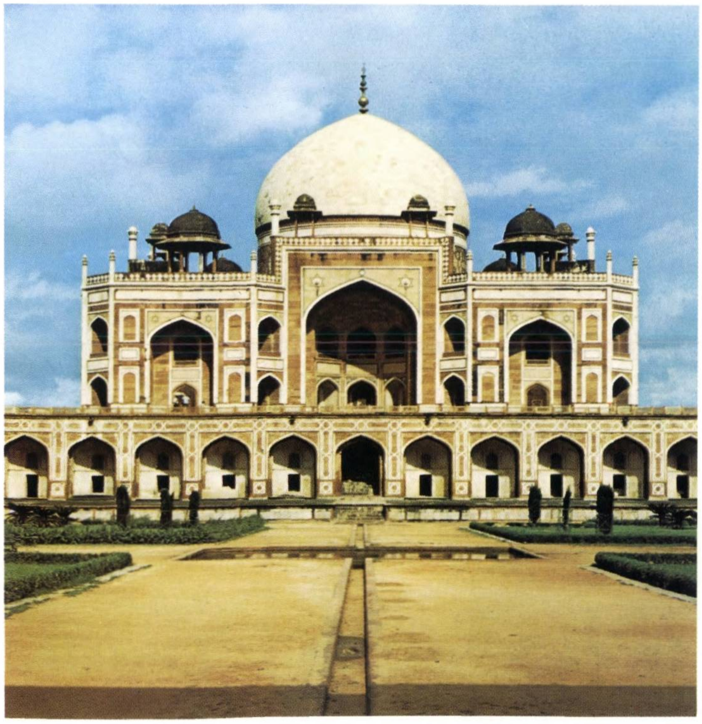
Copper bowl with hunting scene, Mughal, *c.* AD 1583. In the Prince of Wales Museum of Western India, Bombay.



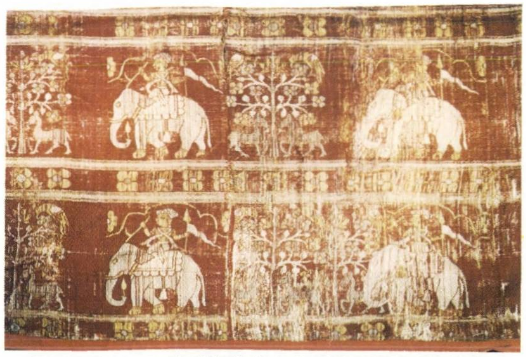


The so-called Dīvān-e Khāss at Fatehpur Sīkri, red sandstone, *c.* AD 1585.

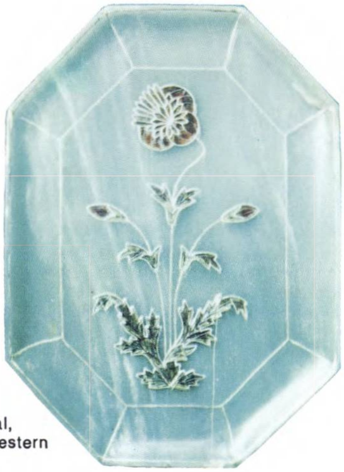
Mughal architecture and decorative arts



The tomb of Humāyūn at Delhi, red sandstone and marble, *c.* AD 1564.



Detail of a silk tapestry, Mughal, early 17th century AD. In the Prince of Wales Museum of Western India, Bombay.



Jade dish inlaid with semiprecious stones and gold, Mughal, mid-17th century AD. In the Prince of Wales Museum of Western India, Bombay.

served no other purpose than to emphasize the importance of the sanctum. The principle of gaining height not by the superimposition of ornamental cornice moldings with *candraśālā* decoration but by the multiplication of stories, each imitating the story below, also distinguished the later architectural style of southern India.

The great Mahābodhi temple at Buddh Gaya, commemorating the spot where the Buddha attained enlightenment, though burdened with later restorations, is essentially a temple of this period. It has a particularly majestic *śikhara,* decorated with ornamental niches and *candraśālā*s, rising over a square sanctum to a great height.

Along with temples, *stūpa*s continued to be built. These also aspired to height, which was achieved by multiplication and heightening of the supporting terraces and elongating the drum and dome. A good example of this new form is the Dhamekh *stūpa* at Sārnāth. Along more conventional lines, but quite elaborate, are the brick *stūpa*s in Sind, notably a fine example at Mīrpur Khās.

The rock-cut temple and monastery tradition also continued in this period, notably in western India, where the excavations—especially at Ajantā—acquire extreme richness and magnificence. The monasteries are characterized by the introduction of images into some of the cells, so that they partake of the nature of temples instead of being simple residences. Temples with an apsidal plan and barrel-vault roofs, however, soon went out of fashion, and are found very rarely in the subsequent period. The early 5th-century cave temples at Udayagiri, Madhya Pradesh, are similar to the simple flat-roofed temples with a hall and are not descended from ancient traditions as preserved in western India.

Medieval temple architecture. Architectural styles initiated during the 5th and 6th centuries found their fullest expression in the medieval period (particularly from the 9th to the 11th centuries), when great stone temples were built. Two main types can be broadly distinguished, one found generally in northern India, the other in southern India. To these can be added a third type, sharing features of both and found in Karnataka and the Deccan. These three types have been identified by some scholars with the *nāgara, drāviḍa,* and *vesara* classes referred to in some Sanskrit texts, though the actual meaning of these terms is far from clear. In spite of the havoc wrought by the destructive Islāmic invasions, particularly in the Indo-Gangetic Plains, an extremely large number of monuments have survived in almost every other part of India, particularly in the south, and these continue to be discovered and recorded to the present day.

Period of the great stone temples

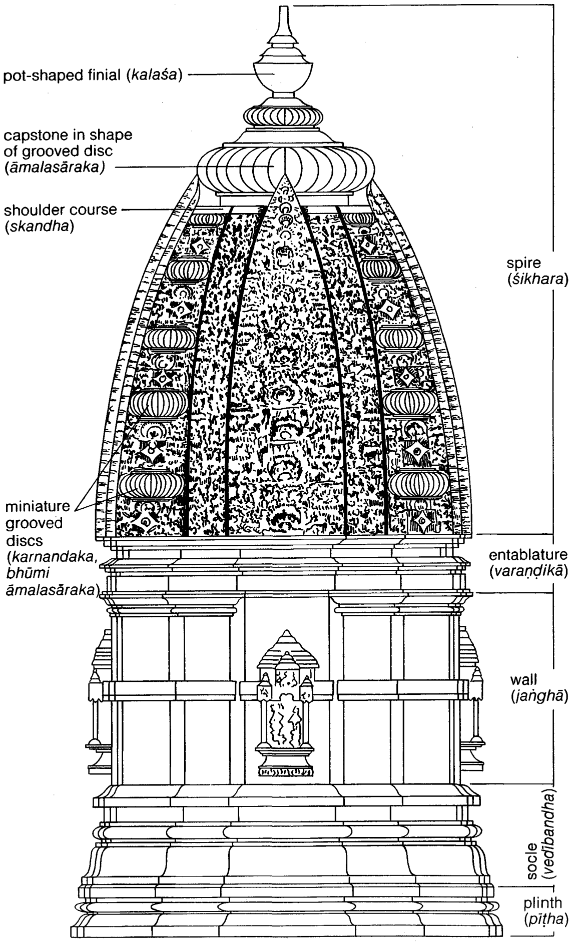
Medieval temple architecture: North Indian style. North Indian temples generally consist of a sanctum enshrining the main image, usually square in plan and shaped like a hollow cube, and one or more halls (called *maṇḍapas*), aligned along a horizontal axis. The sanctum may or may not have an ambulatory, but it is invariably dark, the only opening being the entrance door. The doorway surrounds are richly decorated with bands of figural, floral, and geometrical ornament and with river-goddess groups at the base. A vestibule (*antarāla*) connects the sanctum to the halls, which are of two broad types: the *gūḍhamaṇḍapa*s, which are enclosed by walls, light and air let in through windows or doors; and open halls, which are provided with balustrades rather than walls and are consequently lighter and airier. The sanctum almost invariably, and the *maṇḍapa*s generally, have *śikhara*s; those on the sanctum, appropriately, are the most dominant in any grouping. Internally, the sanctum has a flat ceiling; the *śikhara* is solid theoretically, though hollow chambers to which there is no access must be left within its body to lessen the weight. The ceilings of the halls, supported by carved pillars, are coffered (decorated with sunken panels) and of extremely rich design.

The sanctum is often set on a raised base, or a plinth (*pīṭha*), above which is a foundation block, or socle (*vedībandha*), decorated with a distinct series of moldings; above the *vedībandha* rise the walls proper (*jaṅghā*), which are capped by a cornice or a series of cornice moldings (*varaṇḍikā*), above which rises the *śikhara.* One, three, and sometimes more projections extend all the way from the base of the temple up the walls to the top of the *śikhara.* The central offset (*bhadra*) is the largest and generally carries an image in a niche; the other projections (*rathas*), too, are often decorated with statuary.

The entire temple complex, including sanctum, halls, and attendant shrines, may be raised on a terrace (*jagatī*), which is sometimes of considerable height and size. The attendant shrines—generally four—are placed at the corners of the terrace, forming a *pañcāyatana,* or quincunx, arrangement that is fairly widespread. The temple complex may be surrounded by a wall with an arched doorway (*toraṇa*).

Śikhara types

The *śikhara* is the most distinctive part of the North Indian temple and provides the basis for the most useful and instructive classification. The two basic types are called *latina* and *phāmsanā.* Curvilinear in outline, the *latina* is composed of a series of superimposed horizontal roof slabs and has offsets called *latā*s. The edges of the *śikhara* are interrupted at intervals with grooved discs, each one demarcating a "story." The surface of the entire *śikhara* is covered with a creeper-like tracery, or interlaced work, composed of diminutive ornamental *candraśālās.*

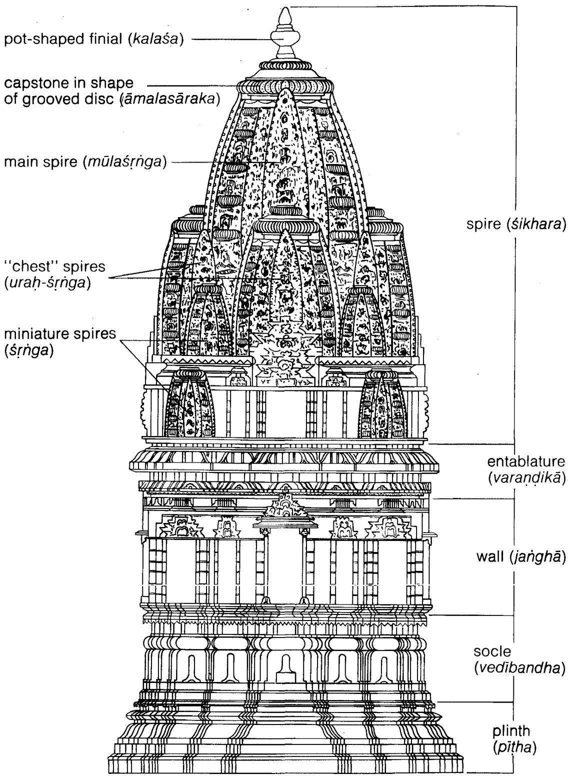

pot-shaped finial (kalaśa)
capstone in shape of grooved disc (āmalasāraka)
shoulder course (skandha)
spire (śikhara)
miniature grooved discs (karnandaka, bhūmi āmalasāraka)
entablature (varaṇḍikā)
wall (jaṅghā)
socle (vedibandha)
plinth (pīṭha)

Elevation of a North Indian temple with the *latina* type of superstructure.

The *śikhara* is truncated at the top and capped by a shoulder course (*skandha*), above which is a circular necking (*grīvā*), carrying a large grooved disc called the *āmalasāraka.* On it rests a pot and a crowning finial (*kalaśa*).

Unlike the *latina,* the *phāmsanā śikhara* is rectilinear rather than curvilinear in outine, and it is lower in height. It is composed of horizontal slabs, like the *latina,* but is capped by a bell-shaped member called the *ghaṇṭā.* The surface of this type of *śikhara* may have projections, like the *latina śikhara,* and be adorned with a variety of architectural ornament.

From the 10th century onward, the *śekharī* type of spire, an elaboration of the *latina* type, became increasingly popular. In its developed form it consisted of a


pot-shaped finial (*kalaśa*)
capstone in shape of grooved disc (*āmalasāraka*)
main spire (*mūlaśṛṅga*)
"chest" spires (*uraḥ-śṛṅga*)
miniature spires (*śṛṅga*)
spire (*śikhara*)
entablature (*varaṇḍikā*)
wall (*jaṅghā*)
socle (*vedībandha*)
plinth (*pīṭha*)

Elevation of a North Indian temple with the *śekharī* type of superstructure.

central *latina* spire (*mūlaśṛṅga*) with one or more rows of half spires added on the sides (*uraḥ-śṛṅga*) and the base strung with miniature spires (*śṛṅgas*). The corners, too, are sometimes filled with quarter spires, the whole mass of carved masonry recalling a mountain with a cluster of subsidiary peaks.

The *latina* and *śekharī* spires are generally found on the sanctum, while the *phāmsanā* and its variants are usually confined to the *maṇḍapas,* or halls. The sanctum spires also have a large and prominent projection in front (*śukanāsā*), generally rising above the vestibule (*antarāla*). These projections are essentially large ogee arches of complex form, which often contain the image of the presiding deity.

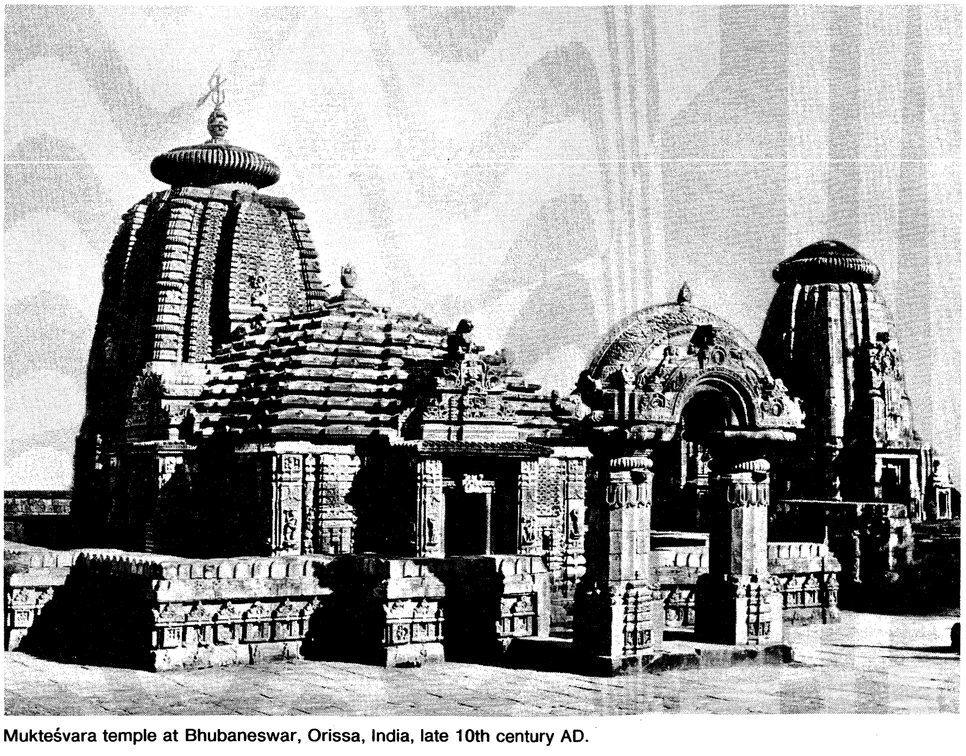
The *bhūmija* *śikhara*

A particularly rich and pleasing variety of North Indian *śikhara,* popular in Mālwa, western India, and northern Deccan, is the *bhūmija* type. It has a central projection on each of the four faces, the quadrants so formed filled with miniature spires in vertical and horizontal rows right up to the top.

Although basically reflecting a homogeneous architectural style, temple architecture in northern India developed a number of distinct regional schools. A detailed elucidation of all has yet to be made, but among the most important are the styles of Orissa, central India, Rājasthān, and Gujarāt. The style of Kashmir is distinct from the rest of northern India in several respects, and hardly any examples of the great schools that flourished in modern Uttar Pradesh, Bihār, and Bengal are left standing. The North Indian style also extended for some time into the Karnataka (formerly Karṇāṭa) territory, situated in the southern Deccan, though the architecture of Tamil Nadu was relatively unaffected by it.

Medieval temple architecture: North Indian style of Orissa. The greatest centre of this school is the ancient city of Bhuvaneśvara, in which are concentrated almost 100 examples of the style, both great and small, ranging in date from the 7th to the 13th century. Among the earliest is the Paraśurāmeśvara temple (7th–8th century), with a heavy, stately *latina śikhara,* to which is attached a rectangular *gūḍhamaṇḍapa* with double sloping roofs. The walls are richly carved, but the interiors, as in almost all examples of the style, are left plain. The Mukteśvara temple (10th century), which has a hall with a *phāmsanā* roof, is the product of the most exquisite workmanship. The enclosing wall and the arched entrance, or *toraṇa,* are still present, giving a clear idea of a temple with all its parts fully preserved. The Brahmeśvara temple, which is dated on the basis of an inscription to the mid-10th century, is a *pañcāyatana,* with subsidiary shrines at all of the corners. The most magnificent building, however, is the great Liṅgarāja temple (11th century), an achievement of Orissan architecture in full flower. The *latina* spire soars to a considerable height (over 125 feet [40 metres]); the wall is divided into two horizontal rows, or registers, replete with statuary; and the

P. Chandra



Mukteśvara temple at Bhubaneswar, Orissa, India, late 10th century AD.

attached hall is exquisitely and minutely carved. The most famous of all Orissan temples, however, is the colossal building at Konārak, dedicated to Sūrya, the sun god. The temple and its accompanying hall are conceived in the form of a great chariot drawn by horses. The *śikhara* over the sanctum has entirely collapsed; and all that survives are the ruins of the sanctum and the *gūḍhamaṇḍapa,* or enclosed hall, and also a separate dancing hall. Of these, the *gūḍhamaṇḍapa* is now the most conspicuous, its gigantic *phāṃsanā śikhara* rising in three stages and adorned with colossal figures of musicians and dancers.

Because the Orissan style usually favours a *latina śikhara* over the sanctum, the *śekharī* spire of the Rājrānī temple (11th century) at Bhuvaneśvara (Bhubaneswar) is quite exceptional. Of particular interest as a late survival of early building traditions is the Vaitāl Deul (8th century), the sanctum of which is rectangular in plan, its *śikhara* imitating a pointed barrel vault. Besides Bhuvaneśvara, important groups of temples are to be found at Khiching and Mukhalingam.

Medieval temple architecture: North Indian style of central India. The area roughly covered by the modern state of Madhya Pradesh was the centre of several vigorous schools of architecture, of which at least four have been identified. The first flourished at Gwalior and adjacent areas (ancient Gopādri); the second in modern Bundelkhand, known in ancient times as Jejākabhukti; the third in the eastern and southeastern parts in the ancient country of Ḍāhala, of which Tripurī, near modern Jabalpur, was the capital; and the fourth in the west, in an area bordering Gujarāt and Rājasthān in the fertile land of Mālava (Mālwa).

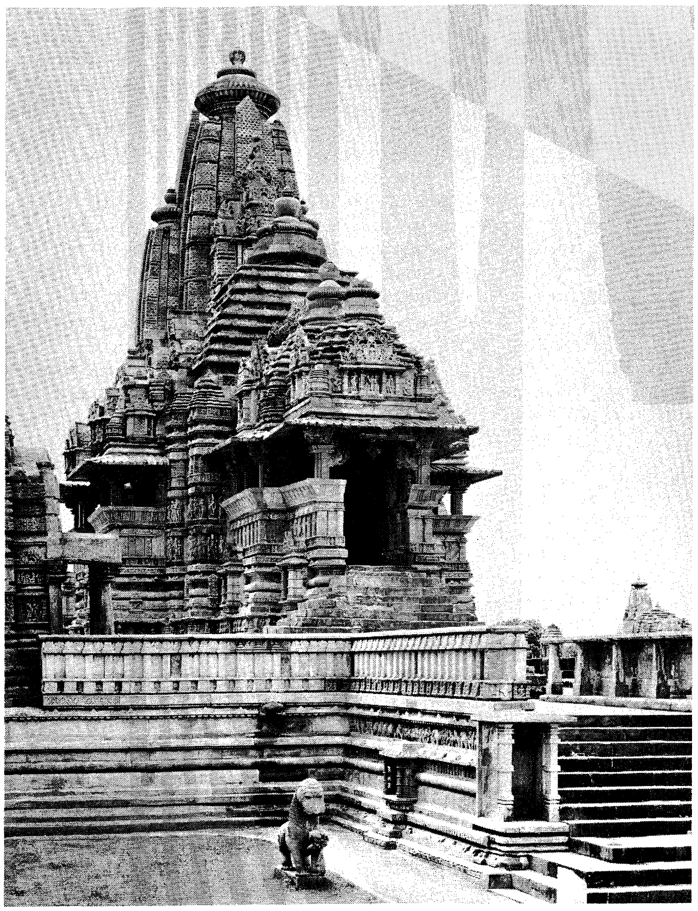
The Gwalior temples

The earliest examples in the Gwalior area are a group of small shrines at Naresar, a few miles from Gwalior proper; dating to the 8th century, the shrines have *latina* spires and sparsely ornamented walls. In the 9th century a series of magnificent temples was built, including the Mālā-de at Gyāraspur, the Śiva temples at Mahuā and Indore, and a temple dedicated to an unidentified mother goddess at Barwa-Sāgar. The period appears to have been one of experimentation, a variety of plans and spires having been tried. The Mālā-de temple is an early example of the *śekharī* type in its formative stages; the Indore temple has a star-shaped plan; and the Barwa-Sāgar example has a twin *latina* spire over a rectangular sanctum. The masonry work is of the finest quality and the architectural ornament is crisply carved. (The figural sculptures are few.) The temple at Umrī, with a *latina* spire, is small and exquisitely finished; but the largest and perhaps the finest temple is the Telī-kā-Mandir on Gwalior Fort, rectangular in plan and capped by a pointed barrel vault, recalling once again the survival of ancient roof forms. The walls are decorated with niches (empty at present) topped by tall pediments (triangular gable ornament).

The style of this region became increasingly elaborate from the 10th century, during the supremacy of the Kacchapaghāṭa dynasty. The many examples from this period are distinguished by a low plinth and rich sculptural decoration on the walls. Outstanding among them are the Kākan-maḍh at Suhāniā (1015–35) and the Sās-Bahū temple (completed 1093) in Gwalior Fort. The several temples at Surwāyā and Kadwāhā, though smaller in size, are distinguished for their extremely rich and elegant workmanship.

The style is best represented by a large group of temples at Khajurāho, the capital of the Candella dynasty, though examples are also to be found in Mahoba and at several other sites in the Jhānsi district of Uttar Pradesh, notably Chāndpur and Dudhai. All of the distinctive characteristics of the fully developed style can be seen in the Lakṣmaṇa temple at Khajurāho (dated 941), which is a *pañcāyatana* placed on a tall terrace enclosed by walls. The sanctum has an ambulatory and, facing it, a series of halls, including the *gūḍhamaṇḍapa,* a porch, and a small intermediate hall. Both the ambulatory and the *gūḍhamaṇḍapa* are provided with lateral, balconied arms, or transepts, which let in light and air. Each hall has its own pyramidal *śikhara,* all skillfully correlated to ascend gradually to the main *śekharī* spire over the sanctum. Extraordinary richness of carving, both in the interior and on the exterior, where the walls carry as many as three rows of sculpture, and a skillful handling of the main spire to suggest ascent are distinguishing features of the style. The largest temple of the group, very similar to the Lakṣmaṇa, is the Kandāriyā Mahādeo; and among the most distinguished are the Viśvanātha and the Pārśvanātha temples. The Dūlādeo temple, which does not have an ambulatory, represents the closing phase of the group and probably belongs to the 12th century.

The Lakṣmaṇa temple



Lakṣmaṇa temple at Khajurāho, Madhya Pradesh, India, *c.* AD 941.

P. Chandra

The earliest temples of the Ḍāhala area, dating from the 8th–9th century, are the simple shrines at Bāndhogarh, which consist of a sanctum with *latina* spire and porch. To the 10th century, when the local Kalacuri dynasty was rapidly gaining power, belong the remarkable Śiva temples at Chandrehe and Masaun, the former being circular in plan, with a *latina* spire covered with rich *candraśālā* tracery. The Golā Math at Maihar has the more conventional square sanctum, with a very elegant *latina śikhara,* the walls of which are adorned with two rows of figural sculpture. There must have existed at Gurgī a large number of temples, though all of them now are in total ruin. Judging from a colossal image of Śiva-Pārvatī and a huge entrance, which have somehow survived, the main temple must have been of very great size. Another important site is Amarkantak, where there are a large group of temples, the most important of which is the Karṇa. Although generally of the 11th century, they are quite simple, lacking the rich sculptural decoration so characteristic of the period. By contrast, the Virāṭeśvara temple at Sohāgpur, with an unusually tall and narrow *śekharī* spire, is covered with sculptural ornamentation as rich as that of Khajurāho.

Temples o the Mālava area

The Mālava region, ruled largely by the Paramāra dynasty, appears to have been the first to develop the *bhūmija* type of *śikhara* (10th century). The finest and most representative group of these structures is at Un. Though, unfortunately, they are considerably damaged, judging from the remains, they must have been very elegant structures. The best preserved and easily the finest *bhūmija* temple is the Udayeśvara (1059–82), situated at Udaipur in Madhya Pradesh. The *śikhara,* based on a stellate plan, is divided into quadrants by four *latās*, or offsets, each one of which has five rows of aediculae. The large hall has three entrance porches, one to the front and two to the sides, and walls that are richly carved. The whole complex, including seven subsidiary shrines, is placed on a broad, tall platform. The Siddheśvara temple at Nemāwar (early 12th century) is even larger than the Udayeśvara, though the proportions are not as well balanced and the quality of the carving is inferior. Structures in the *bhūmija* manner continued to be made in Mālava up to the 15th century; the Malvai temple at Alīrājpur is a good example of the late phase.

From Mālava, the *bhūmija* style spread to the neighbouring regions. To the north in Rājasthān, the Mahānāleśvara temple at Menāl (*c.* 11th century), the Sun temple at Jhālrapātan (11th century), the Śiva temple at Rāmgarh (12th century), and the Uṇḍeśvara temple (12th century) at Bījoliān are important examples. To the west, in Gujarāt, are temples at Limkheda and Sarnāl of the 11th and 12th centuries. The style was particularly favoured in Mahārāshtra, to the south. Among surviving examples, the most impressive is the Ambarnāth temple near Bombay (11th century); Balsāṇe and Sinnar also have pleasing temples. The style continued up to the 16th century, many examples having been found in north Deccan and Berār. The *bhūmija* style also spread to the east of Mālava; the Bhāṇḍ Dewal at Arang (11th century), for example, is a Ḍāhala adaptation.

Medieval temple architecture: North Indian style of Rājasthān. A group of temples at Osiān, dating to about the 8th century, represents adequately the opening phases of medieval temple architecture in Rājasthān. They stand on high terraces and consist of a sanctum, a hall, and a porch. The sanctum is generally square and has a *latina* spire. The walls, with one central and two subsidiary projections, are decorated with sculpture, often placed in niches with tall pediments. The halls are generally of the open variety, provided with balustrades rather than walls, so that the interiors are well lit. The surrounds of the doorway sanctum are quite elaborate, with four or five bands of decoration and the usual river-goddess groups at the base. The pillars, with *ghaṭa-pallava* (vase-and-foliage) capitals, are also decorated, richness of sculpture and architectural elaboration being a characteristic of this group of monuments. The Mahāvīra temple, which is the largest, belongs to the 8th century, though renovated in later times, when the *toraṇa* (gateway) and the *śikhara* were added. Other important temples are Harihara Nos. 1, 2, and 3 and two temples dedicated to Vishnu. The ruined Harshat Mātā temple at Ābānerī, of a slightly later date (*c.* 800), was erected on three stepped terraces of great size and is remarkable for the exquisite quality of the carving. Some of the finest temples of the style date from the 10th century, the most important of which are the Ghaṭeśvara temple at Bāḍolī and the Ambikā Mātā temple at Jagat. The simple but beautiful Bāḍolī temple consists of a sanctum with a *latina* superstructure and an open hall with six pillars and two pilasters (columns that project a third of their width or less from the wall) supporting a *phāmsanā* spire. Only the central projections of the sanctum walls are decorated with niches containing sculpture. A large open hall was built in front of the temple at a later date. The Ambikā-Mātā temple at Jagat, of the mid-10th century, is exceptionally fine. It consists of a sanctum, a *gūḍhamaṇḍapa,* or enclosed hall, and a parapeted porch with projecting eaves. The walls of the sanctum and the hall are covered with fine sculpture, the superstructures being of the *śekharī* and the *phāmsanā* types.

rowing esem-lance etween yles of āja-hān and ujarāt

Temples, too numerous to mention, dating from the 10th and—to an even greater extent—the 11th century onward, are found throughout Rājasthān. The styles of Rājasthān and neighbouring Gujarāt during these centuries grew closer and closer together until the differences between them were gradually obliterated. This coalescence resulted in the emergence of a composite style found throughout Gujarāt and Rājasthān. Temples situated in the two areas are discussed separately here, but this is for the sake of convenience and does not signify any real stylistic difference.

The temples at Kirāḍu in Rājasthān, dating from the late 10th and 11th centuries, are early examples of the style shared by Rājasthān and Gujarāt. The Someśvara temple (*c.* 1020) is the most important and clearly shows the movement toward increasing elaboration and ornamentation. Each of the constituent parts became more complex; the moldings of the plinth, for example, are multiplied to include bands of elephants, horses, and soldiers. The walls are covered with sculpture, and the spire is of the rich *śekharī* type. Situated in Rājasthān, but again in the composite style, are the extraordinarily sumptuous temples known as the Vimala Vasahī (1031) and the Lūṇa Vasahī (1230) at Mt. Ābū. The Vimala Vasahī consists of a sanctum, a *gūḍhamaṇḍapa,* and a magnificent assembly hall added in mid-12th century. The plain, uncarved exterior walls of the rectangular enclosure of the temple have on the inside rows of cells containing images of divinities. The interiors are very richly carved, the coffered ceilings loaded with a wealth of detail. The Lūṇa Vasahī is even more elaborate, though the quality of the work had begun to decline perceptibly.

Traditional architecture continued even after the Islāmic invasions, particularly during the reign of Rāṇā Kumbhā of Mewār (*c.* 1430–69). During this period, the tall nine-storied Kīrttistambha and other temples at Chitor and also the great Chaumukha temple at Ranakpur (1438) were built.

The temple at Gop

Medieval temple architecture: North Indian style of Gujarāt. Gujarāt was the home of one of the richest regional styles of northern India. A temple at Gop (*c.* 600), with a tall terrace and a cylindrical sanctum with high walls capped by a *phāmsanā* roof, and other temples in Saurāshṭra show the formative phases of the style. Its distinctive features are clear in an interesting group of temples from Roḍā (*c.* 8th century). The sanctum is square in plan and has *latina* spires that are weighty and majestic. The walls are relatively plain, with niches, housing images, provided only on the central projection. The masonry work is exceptionally good, a characteristic of Gujarāt architecture throughout its history. The Rāṇakdevī temple at Wadhwān, of the early 10th century, is also characterized by plain walls and a *latina* spire, while the Śiva temple at Kerākoṭ has a *śekharī* spire and also a *gūḍhamaṇḍapa.* The great Sun Temple at Modhera, datable to the early years of the 11th century, represents a fully developed Gujarāt style of great magnificence. The temple consists of a sanctum (now in ruins), a *gūḍhamaṇḍapa,* an open hall of extraordinary richness, and an arched entrance in front of which was the great tank. The Navalakhā temple at Sejakpur continued this tradition. The Rudramāla at Siddhapur, the most magnificent temple of the 12th century, is now in a much ruined condition, with only the *toraṇa* (gateway) and some subsidiary structures remaining. Successively damaged and rebuilt, the Somanātha at Prabhāsa Patan was the most famous temple of Gujarāt, its best known structure dating from the time of Kumārapāla (mid-12th century). It has been now dismantled, but a great temple built at the site in recent years testifies to the survival of ancient traditions in modern Gujarāt.

The hills of Satrunjaya and Girnār house veritable temple cities. Most of the shrines, which are of late date, are picturesque but otherwise of little significance. With the Islāmic conquest, the Gujarāt architect adapted his considerable skills to meet the needs of a patron of different religion and quickly produced a totally successful Indian version of Islāmic architecture.

Medieval temple architecture: North Indian style of Karnataka. The North Indian style was largely confined to India above the Vindhyas, though for a short period it also flourished in a region of southern India known as Karnataka from ancient times and now largely part of Karnataka (formerly Mysore) state. Here, temples of the northern and the southern styles are found next to each other, notably at Aihole and Pattadkal. The earliest appears to be the Lāḍh Khān at Aihole, closely related to the 5th-century temple at Nāchnā Kutharā in northern India. The northern style was also cultivated at Pattadkal, where the most important examples are the Kāśīviśvanātha, the Galaganātha, and the Pāpanātha. Ālampur, now in Andhra Pradesh, has eight temples of the northern style with *latina* spires. These belong to the late 7th and early 8th centuries and are the finest and among the last examples of the northern style in southern India.

Distinct architectural style of Kashmir

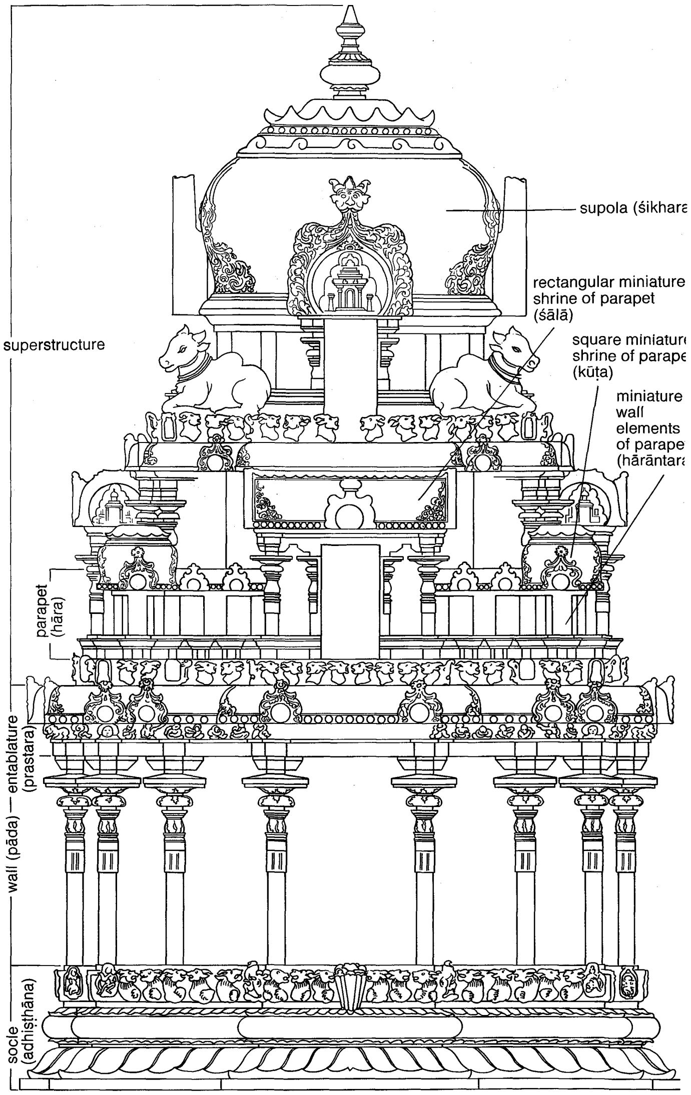
Medieval temple architecture: North Indian style of Kashmir. The architectural style of the Kashmir region is quite distinct: unlike other regions, in which the sanctum usually has a *latina* or *śekharī* spire, the roof of the Kashmir sanctum is of the *phāmsanā* type, with eaves raised in two stages. The greatest example to survive is the ruined Sun Temple at Mārtanḍ (mid-8th century), which, though its *śikhara* is missing, gives a good idea of the characteristic features of the style. The temple is placed in a rectangular court enclosed by a series of columns. Access to the court is through an imposing entrance hall, the walls of which have doorways with gabled pediments and a trefoil (shaped like a trifoliate leaf) recess. The Avantisvāmī temple of the mid-9th century, now quite ruined, must have been similar, though much more richly ornamented. The style continued up to the 12th century; the Rilhaṇeśvara temple at Pāndrenṭhan is a comparatively well-preserved example of this period.

Medieval temple architecture: South Indian style. The home of the South Indian style, sometimes called the drāviḍa style, appears to be the modern state of Tamil Nadu; examples, however, are found all over southern India, particularly in the adjoining regions of Karnataka and Andhradeśa, now largely covered by the states of Karnataka and Andhra Pradesh. Both Andhradeśa and Karnataka developed variants, particularly Karnataka, which evolved a distinct manner, basically South Indian but with features of North Indian origin. The Karnatic style extended northward into Mahārāshtra, where the Kailāsa temple at Ellora is the most famous example.

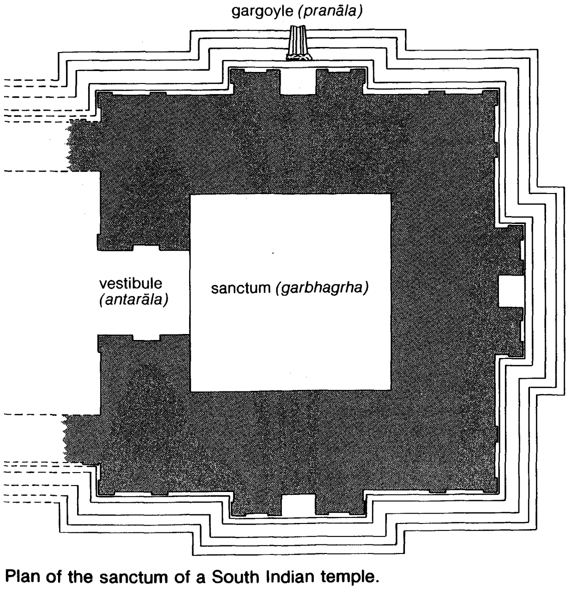
A typical South Indian temple consists of a hall and a square sanctum that has a superstructure of the *kūṭina* type. Pyramidal in form, the *kūṭina* spire consists of stepped stories, each of which simulates the main story and is conceived as having its own "wall" enclosed by a parapet. The parapet itself is composed of miniature shrines strung together: square ones (called *kūṭas*) at the corners and rectangular ones with barrel-vault roofs (called *śālās*) in the centre, the space between them connected by miniature wall elements called *hārāntaras*. (Conspicuous in the early temples, these stepped stories of the superstructure with their parapets became more and more ornamental, so that in the course of time they evolved into more or less decorative bands around the pyramidal superstructure.) On top of the stepped structure is a necking that supports a solid dome, or cupola (instead of the North Indian grooved disc), which in turn is crowned by a pot and finial. The walls of the sanctum rise above a series of moldings, constituting the foundation block, or socle (*adhiṣṭhāna*), that differ from North Indian temples; and the surface of the walls does not have the prominent offsets seen in North Indian temples but is instead divided by pilasters. In the Karnatic version, particularly from the late 10th century onward (sometimes called the *vesara* style), this arrangement of the superstructure is loaded with decoration, thus considerably obscuring the component elements. At the same time, these elements—particularly the central offset with its subsidiaries that carry *candraśālā* motifs—are so manipulated that they tend to form distinct vertical bands, in this respect closely recalling the *śikharas* of northern India.

The design of the hall-temple roofed by a barrel vault, popular in the centuries before and after Christ, was adopted in southern India for the great entrance buildings, or *gopuras*, that give access to the sacred enclosures in which the temples stand. Relatively small and inconspicuous in the early examples, they had, by the mid-12th century, outstripped the main temple in size.

Medieval temple architecture: South Indian style of Tamil Nadu (7th–18th century). The early phase, which, broadly speaking, coincided with the political supremacy of the Pallava dynasty (*c.* 650–893), is best represented by the important monuments at Mahābalipuram. Besides a fine group of small cave temples (early 7th century), among the earliest examples of their type in southern India, there are



Elevation of a South Indian temple with the *kūṭina* type of superstructure.



Plan of the sanctum of a South Indian temple.

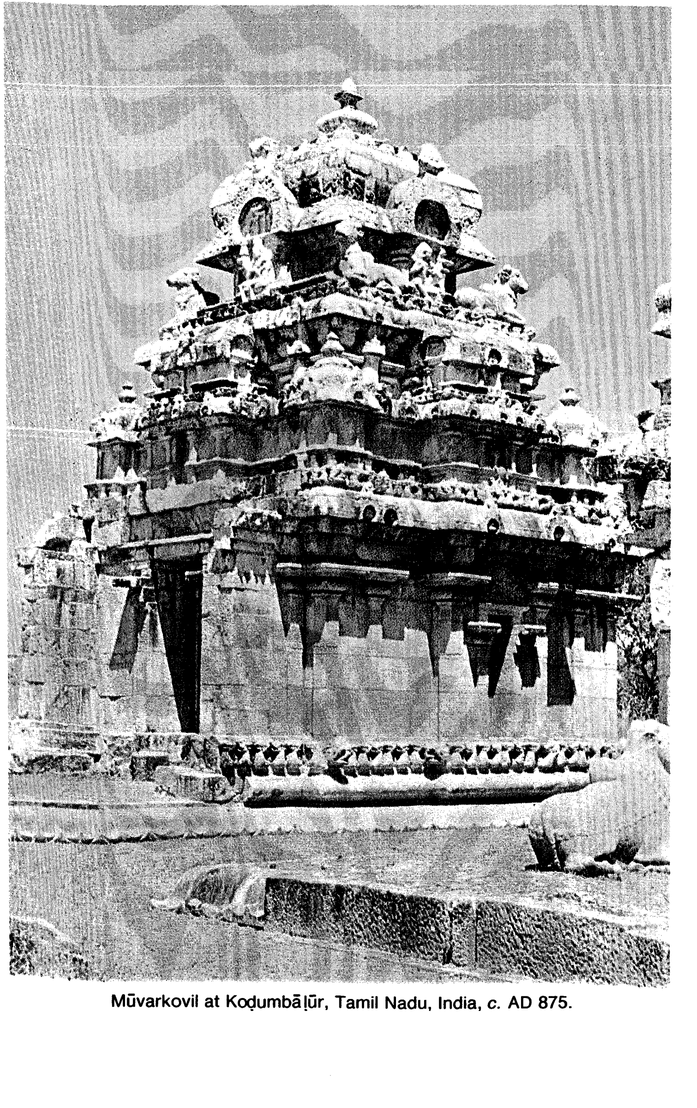
he Shore
emple

here several monolithic temples carved out of the rock, the largest of which is the massive three-storied Dharmarāja-ratha (*c.* 650). The finest temple at this site and of this period is an elegant complex of three shrines called the Shore Temple (*c.* 700), not cut out of rock but built of stone. The Tālapurīśvara temple at Panamalai is another excellent example. The capital city of Kānchipuram also possesses some fine temples—for example, the Kailāsanātha (dating a little later than the Shore Temple), with its stately superstructure and subsidiary shrines attached to the walls. The enclosure wall has a series of small shrines on all sides and a small *gopura.* Another splendid temple at Kānchipuram is the Vaikuṇṭha Perumāl (mid-8th century), which has an interesting arrangement of three sanctums, one above the other, encased within the body of the superstructure.

The 9th century marked a fresh movement in the South Indian style, revealed in several small, simple, but most elegant temples set up during the ascendancy of the Cōḷa and other contemporary dynasties. Most important of a large number of unpretentious and beautiful shrines that dot the Tamil countryside are the Vijayālaya Cōḷīśvara temple at Nārttāmalai (mid-9th century), with its circular sanctum, spherical cupola, and massive, plain walls; the twin shrines called Agastyīśvara and Cōḷīśvara, at Kīḷaiyūr (late 9th century); and the splendid group of two temples (originally three) known as the Mūvarkovil, at Koḍumbāḷūr (*c.* 875).

These simple beginnings led rapidly (in about a century) to the mightiest of all temples in the South Indian style, the Bṛhadīśvara, or Rājarājeśvara, temple, built at the Cōḷa capital of Thanjāvūr. A royal dedication of Rājarāja I, the temple was begun around 1003 and completed about seven years later. The main walls are raised in two stories, above which the superstructure rises to a height of 190 feet (60 metres). It has 16 stories, each of which consists of a wall with a parapet of shrines carved in relatively low relief. The great temple at Gaṅgaikoṇḍacōḷapuram, built (1030–40) by the Cōḷa king Rājendra I, is somewhat smaller than the Bṛhadīśvara; but the constituent elements of its superstructure, whose outline is concave, are carved in bolder relief, giving the whole a rather emphatic plasticity. The Airāvateśvara (1146–73) and Kampahareśvara (1178–1223) temples at Dārāsuram and Tribhuvanam follow the tradition of the 11th century but are smaller and considerably more ornate. They bring to a close a great phase of South Indian architecture extending from the 11th to the 13th century.

P. Chandra



Mūvarkovil at Koḍumbāḷūr, Tamil Nadu, India, *c.* AD 875.

From the middle of the 12th century onward, the *gopuras*, or entrance buildings, to temple enclosures began to be greatly emphasized. They are extremely large and elaborately decorated with sculpture, quite dominating the architectural ensemble. Their construction is similar to that of the main temple except that they are rectangular in plan and capped by a barrel vault rather than a cupola, and only the base is of stone, the superstructure being made of brick and plaster. Among the finest examples are the Sundara Pāndya *gopura* (13th century) of the Jambukeśvara temple at Tiruchchirāppalli and the *gopuras* of a great Śiva temple at Chidambaram, built largely in the 12th–13th century. Even larger *gopuras*, if not of such fine quality, continued to be built up to the 17th century. Such great emphasis was placed on the construction of *gopuras* that enclosure walls, which were not really necessary, were especially built to justify their erection. In the course of time several walls and *gopuras* were successively built, each enclosing the other so that at the present day one often has to pass through a succession of walls with their *gopuras* before reaching the main shrine. A particularly interesting example is the Ranganātha temple at Srīrangam, which has seven enclosure walls and numerous *gopuras*, halls, and temples constructed in the course of several centuries. The *gopuras* of the Mīnākṣī temple at Madurai are also good representative examples of this period.

In addition to the *gopuras*, temples also continued to be built. Although they never achieved colossal size, they are often of very fine workmanship. The Subrahmaṇya temple of the 17th century, built in the compound of the Bṛhadīśvara temple at Thanjāvūr, indicates the vitality of architectural traditions even at this late date.

Rock-cut
cave
temples
of South
India

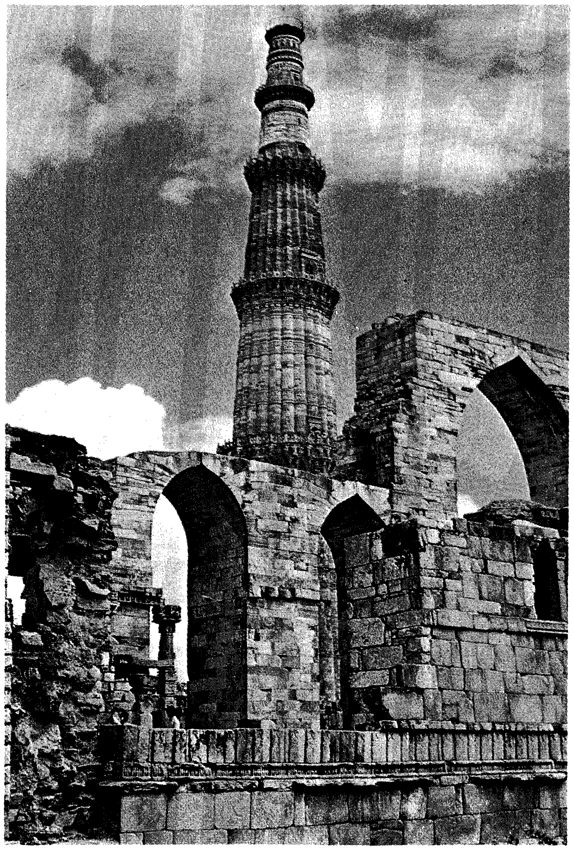
Medieval temple architecture: South Indian style of Karnataka. The early phase, as in Tamil Nadu, opens with the rock-cut cave temples. Of the elaborate and richly sculptured group at Bādāmi, one cave temple is dated 578, and two cave temples at Aihole are early 8th century. Among structural temples built during the rule of the Cālukyas of Bādāmi are examples in the North Indian style; but, because the Karnataka region was more receptive to southern influences, there are a large number of examples that are basically South Indian with only a few North Indian elements. The Durgā temple (*c.* 7th century) at Aihole is apsidal in plan, echoing early architectural traditions; the northern *latina śikhara* is in all probability a later addition. The Mālegitti Śivālaya temple at Bādāmi (early 8th century), consisting of a sanctum, a hall with a parapet of *śālās* and *kūṭas* (rectangular and square miniature shrines), and an open porch, is similar to examples in Tamil Nadu. The Virūpākṣa at Pattadkal (*c.* 733–746) is the most imposing and elaborate temple in the South Indian manner. It is placed within an enclosure, to which access is through a *gopura;* and the superstructure, consisting of four stories, has a projection in the front, a feature inspired by the prominent projections, or *śukanāsā,* of North Indian temples. Belonging to the 9th century is the triple shrine (the three sanctums sharing the same *maṇḍapa,* or hall) at Kambaḍahalli and the extremely refined and elaborately carved Bhoganandīśvara temple at Nandi. The Chāvuṇḍarāyabasti (*c.* 982–995) at Śravaṇa-Beḷgoḷa is also an impressive building, with an elegant superstructure of three stories.

With the 10th century, the Karnatic idiom begins to show an increasing individuality that culminates in the distinctive style of the 12th century and later. The Kalleśvara temple at Kukkanūr (late 10th century) and a large

Jaina temple at Lakkundi (*c.* 1050–1100) clearly demonstrate the transition. The superstructures, though basically of the South Indian type, have offsets and recesses that tend to emphasize a vertical, upward movement. The Lakkundi temple is also the first to be built of chloritic schist, which is the favoured material of the later period and which lends itself easily to elaborate sculptural ornamentation. With the Mahādevā temple at Ittagi (*c.* 1112) the transition is complete, the extremely rich and profuse decoration characteristic of this shrine being found in all work that follows. Dating from the reign of the Hoysaḷa dynasty (*c.* 1141) is a twin Hoysaḷeśvara temple at Halebīd, the capital city. The sanctums are stellate in form but lack their original superstructures. The pillars of the interior are lathe-turned in a variety of fanciful shapes. The exterior is almost totally covered with sculpture, the walls carrying the usual complement of images; the base, or socle, is decorated with several bands of ornamental motifs and a narrative relief. Among other temples that were constructed in this style, the most important are the Chenna Keśava temple at Belūr (1117), the Amṛteśvara temple at Amritpur (1196), and the Keśava temple at Somnāthpur (1268).

Strong traditions of cave architecture in Mahārāshtra

Medieval temple architecture: South Indian style of Mahārāshtra, Andhradeśa, and Kerala. The traditions of cave architecture are stronger in Mahārāshtra than in any other part of India; there, great shrines were cut out of rock right up to the 9th century AD and even later. Of those belonging to the early phase, the most remarkable is a temple at Elephanta (early 6th century); equally impressive are numerous temples at Ellora (6th–9th centuries). The Karnatic version of the South Indian style extended northward into Mahārāshtra, where the Kailāsa temple at Ellora, erected in the reign of the Rāṣṭrakūṭā Krishna I (8th century), is its most stupendous achievement. The entire temple is carved out of rock and is over 100 feet (30 metres) high. It is placed in a courtyard, the three sides of which are carved with cells filled with images; the front wall has an entrance *gopura.* The tall base, or plinth, is decorated with groups of large elephants and griffins, and the superstructure rises in four stories. Groups of important temples in the southern style are also found in the Andhra country, notably at Biccavolu, ranging in date from the 9th to the 11th centuries. The 13th-century temples at Palampet are the counterparts of the elaborate Karnatic style of the same period, but without its overpowering elaboration. The temples of Kerala represent an adaptation of the South Indian style to the great main fall of this region and are provided with heavy sloping roofs of stone that imitate timber originals required for draining away the water.

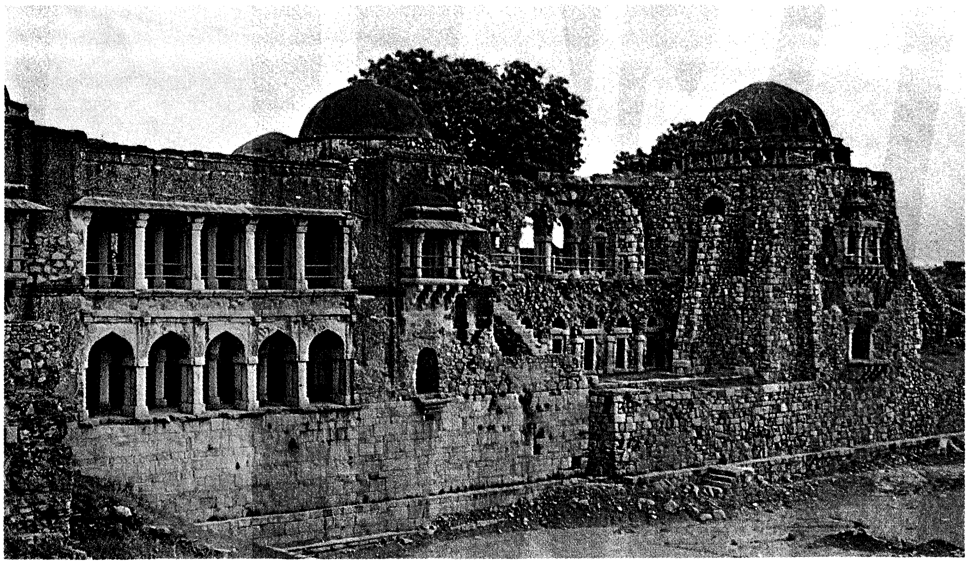


Quṭb Mīnār and the Qūwat-ul-Islām mosque at Delhi, *c.* AD 1196.

P. Chandra

Islāmic architecture in India: period of the Delhi and provincial sultanates. Although the province of Sind was captured by the Arabs as early as 712, the earliest examples of Islāmic architecture to survive in the subcontinent date from the closing years of the 12th century; they are located at Delhi, the main seat of Muslim power throughout the centuries. The Qūwat-ul-Islām mosque (completed 1196), consisting of cloisters around a courtyard with the sanctuary to the west, was built from the remains of demolished temples. In 1198 an arched facade (*maqṣūrah*) was built in front to give the building an Islāmic aspect, but its rich floral decoration and corbelled (supported by brackets projecting from the wall) arches are Indian in character. The Quṭb Mīnār, a tall (288 feet high), fluted tower provided with balconies, stood outside this mosque. The Aṛhāi-dīn-

P. Chandra

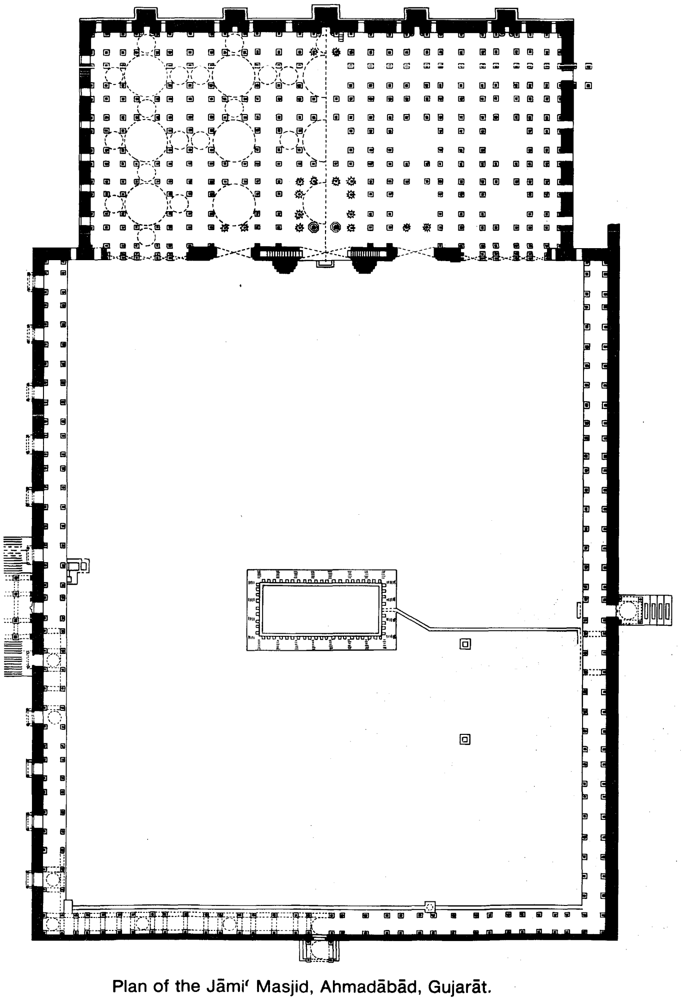


Tomb and palace of Fīrūz Shāh (Hawz-e Khāss) at Delhi, *c.* 1380.

kā-jhomprā mosque (*c.* 1119), built at Ajmer, was similar to the Delhi mosque, the *maqṣūrah* consisting of engrailed (sides ornamented with several arcs) corbel arches decorated with greater restraint than the Quṭb example. The earliest Islāmic tomb to survive is the Sultān Gharī, built in 1231, but the finest is the tomb of Iltutmish, who ruled from 1211 to 1236. The interior, covered with Arabic inscriptions, in its richness displays a strong Indian quality. The first use of the true arch in India is found in the ruined tomb of Balban (died 1287). From 1296 to 1316 ʿAlāʾ-ud-Dīn Khaljī attempted to expand the Qūwat-ul-Islām mosque, which already had been enlarged in 1230, to three times its size; but he was unable to complete the work. All that has survived of it is the Alai Darwāzah, a beautiful entrance.

4th-
ntury
yle
npover-
hed and
ıstere

In contrast to this early phase, the style of the 14th century at Delhi, ushered in by the Tughluq dynasty, is impoverished and austere. The buildings, with a few exceptions, are made of coarse rubble masonry and overlaid with plaster. The tomb of Ghiyās-ud-Dīn Tughluq (*c.* 1320–25), placed in a little fortress, has sloping walls faced with panels of stone and marble. Also to be ascribed to his reign is the magnificent tomb of Shāh Rukn-e ʿĀlam at Multān in Pakistan, which is built of brick and faced with exquisite tile work. The Koṭla Fīrūz Shāh (1354–70), with its mosques, palaces, and tombs, is now in ruins but represents the major building activity of Fīrūz Shāh, who took a great interest in architecture. Many mosques and tombs of this period and of the 15th century are found in Delhi and its environs; the most notable of them are the Begampur and Khiṛkī mosques and an octagonal tomb of Khān-e Jahān Tilangānī. In the early 16th century, Shēr Shāh Sūr refined upon this style, the Qalʿah-e Kuhnah Masjid and his tomb at Sasarām (*c.* 1540) being the finest of a series of distinguished works that were created during his reign.


Plan of the Jāmiʿ Masjid, Ahmadābād, Gujarāt.

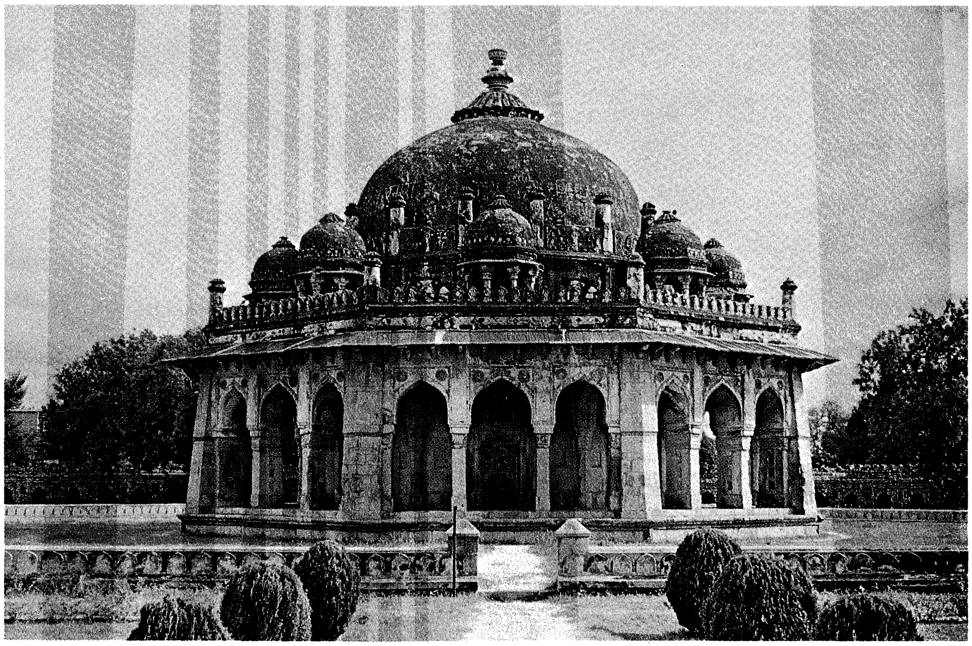
The provinces, which gradually became independent sultanates, did not lag behind in architectural activity. In West Bengal, at Pandua, is the immense Ādīna Masjid (1364–69), which utilized remains of Indian temples. In Jaunpur, Uttar Pradesh, are a group of elegant mosques, notably the Aṭalā Masjid (1377–1408) and the Jāmiʿ Masjid (*c.* 1458–79), characterized by *maqṣūrah*s that have the aspect of imposing gateways. The sultans of Mālwa built elegant structures at Māndu and at Chanderi in the middle of the 15th century. The sultanate of Gujarāt is notable for its great contribution to Islāmic architecture in India. The style, which is basically indigenous, reinterprets foreign influences with great resourcefulness and confidence, producing works notable for their integrity and unity. The city of Ahmadābād (Ahmedabad) is full of elegant buildings; the Jāmiʿ Masjid (*c.* 1424), for example, is a masterly exposition of the style. Fine examples dating from the second half of the 15th century are the small but exquisite mosques of Muḥāfiz Khān (1492) and Rānī Sabraʾi (1514) at Ahmadābād and the handsome Jāmiʿ Masjid at the city of Chāmpāner.

The Deccan was another great centre, but in contrast to Gujarāt it took little from the indigenous building traditions. Among the earliest works is the Jāmiʿ Masjid at Gulbarga (1367), with its extraordinary cloisters consisting of wide arches on low piers, producing a most solemn effect. The city of Bīdar possesses many remains, including a remarkable series of 12 tombs, the most elaborate of which is that of ʿAlā-ud-Dīn Aḥmad Bahmanī (died 1457), which has extremely fine decorations in coloured tile. Some of the finest examples of Islāmic architecture in the Deccan, however, are in Bijāpur. The most important buildings of this city are the great Jāmiʿ Masjid (begun in 1558) with its superb arched cloisters; the ornate Ibrāhīm Rawẓa; and the Gōl Gunbad (built by Muḥammad ʿĀdil Shāh), a tomb of exceptional size and grandeur, with one of the largest domes in existence.

The Hindu kingdoms that managed to retain varying degrees of independence during the period of Islāmic supremacy also produced important works. These structures naturally bore the imprint of what survived of traditional Indian architecture to a greater extent than did those monuments patronized by Muslims. Among the Hindu structures of this period are the extensive series of palaces, all in ruin, built by Rāṇā Kumbhā (*c.* 1430–69) at Chitor, and the superb Mān Mandir palace at Gwalior (1486–1516), a rich and magnificent work that exerted considerable influence on the development of Mughal architecture at Fatehpur Sīkrī.

Striking revival of Islāmic architecture

Islāmic architecture in India: Mughal style. The advent of the Mughal dynasty marks a striking revival of Islāmic architecture in northern India: Persian, Indian, and the various provincial styles were successfully fused to produce works of unusual refinement and quality. The tomb of Humāyūn, begun in 1564, inaugurates the new style. Built entirely of red sandstone and marble, it shows considerable Persian influence. The great fort at Āgra (1565–74) and the city of Fatehpur Sīkri (1569–74) represent the building activities of the emperor Akbar. The former has the massive so-called Delhi gate (1566) and lengthy and immense walls carefully designed and faced with dressed stone throughout. The most important achievements, however, are to be found at Fatehpur Sīkri; the Jāmiʿ Masjid (1571), with the colossal gateway known as the Buland Darwāza, for example, is one of the finest mosques of the Mughal period. Other notable buildings include the palace of Jodhā Bāī, which has a strongly indigenous aspect; the exquisitely carved Turkish Sultānā's house; the Pānch-Maḥal; the Dīvān-e ʿĀmm; and the so-called hall of private audience. Most of the buildings are of post and lintel construction, arches being used very sparingly. The tomb of the emperor, at Sikandarā, near Āgra, is of unique design, in the shape of a truncated square pyramid 340 feet (103 metres) on each side. It consists of five terraces, four of red sandstone and the uppermost of white marble. Begun about 1602, it was completed in 1613, during the reign of Akbar's son Jahāngīr. Architectural undertakings in this emperor's reign were not very ambitious, but there are fine buildings,

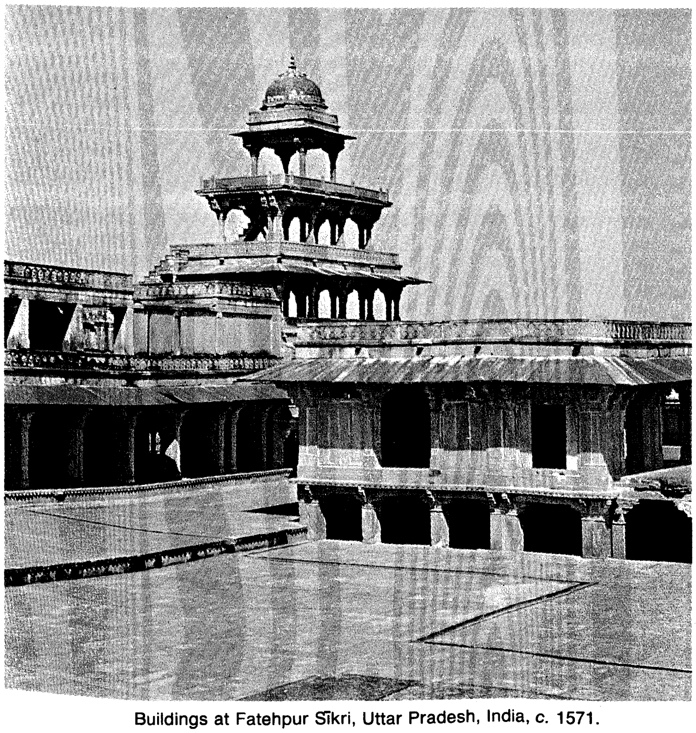


Tomb of 'Isā Khān at Delhi, 1547.

P. Chandra

chiefly at Lahore. The tomb of his father-in-law I'timād-ud-Dawla, at Āgra, is small but of exquisite workmanship, built entirely of delicately inlaid marble. The reign of Shāh Jahān (1628–58) is as remarkable for its architectural achievements as was that of Akbar. He built the great Red Fort at Delhi (1639–48), with its dazzling hall of public audience, the flat roof of which rests on rows of columns and pointed, or cusped, arches, and the Jāmi' Masjid (1650–56), which is among the finest mosques in India. But it is the Tāj Mahal (*c.* 1632–*c.* 1649), built as a tomb for Queen Mumtāz Maḥal, that is the greatest masterpiece of his reign. All the resources of the empire were put into its construction. In addition to the mausoleum proper, the complex included a wide variety of accessory buildings of great beauty. The marble mausoleum rises up from a tall terrace (at the four corners of which are elegant towers, or minārs) and is crowned by a graceful dome.

P. Chandra



Buildings at Fatehpur Sīkri, Uttar Pradesh, India, c. 1571.

Other notable buildings of the reign of Shāh Jahān include the Motī Masjid (*c.* 1648–55) and the Jāmi' Masjid at Āgra (1548–55).

Architectural monuments of the reign Aurangzel

Architectural monuments of the reign of Aurangzeb represent a distinct decline; the tomb of Rābī'ah Begam at Aurangābād, for example (1679), is a poor copy of the Tāj Mahal. The royal mosque at Lahore (1673–74) is of much better quality, retaining the grandeur and dignity of earlier work; and the Motī Masjid at Delhi (1659–60) possesses much of the early refinement and delicacy. The tomb of Ṣafdar Jang at Delhi (*c.* 1754) was among the last important works to be produced under the Mughal dynasty and had already lost the coherence and balance characteristic of mature Mughal architecture.

European traditions and the modern period. Buildings imitating contemporary styles of European architecture, often mixed with a strong provincial flavour, were known in India from at least the 16th century. Some of this work was of considerable merit, particularly the baroque architecture of the Portuguese colony of Goa, where splendid buildings were erected in the second half of the 16th century. Among the most famous of these structures to survive is the church of Bom Jesus, which was begun in 1594 and completed in 1605.

The 18th and 19th centuries witnessed the erection of several buildings deeply indebted to Neoclassic styles; these buildings were imitated by Indian patrons, particularly in areas under European rule or influence. Subsequently, attempts were made by the British, with varying degrees of success, to engraft the neo-Gothic and also the neo-Saracenic styles onto Indian architectural tradition. At the same time, buildings in the great Indian metropolises came under increasing European influence; the resulting hybrid styles gradually found their way into cities in the interior. In recent years an attempt has been made to grapple with the problems of climate and function, particularly in connection with urban development. The influence of the Swiss architect Le Corbusier, who worked on the great Chandīgarh project, involving the construction of a new capital for Punjab, in the early 1950s, and that of other American and European masters has brought about a modern architectural movement of great vitality, which is in the process of adapting itself to local requirements and traditions.

Sculpture. On the Indian subcontinent, sculpture seems to have been the favoured medium of artistic expression. Even architecture and the little painting that has survived from the early periods partake of the nature of sculpture. Particularly is this true of rock-cut architecture, which is

Church of Bom Jesus at Velha Goa, India, 1594–1605.
P. Chandra

often little more than sculpture on a colossal scale. Structural buildings are also profusely adorned with sculpture that is often inseparable from it. The close relationship between architecture and sculpture has to be taken into account when considering individual works that, even if complete in themselves, are also fragments belonging to a larger context. Indian sculpture, particularly from the 10th century onward, thus cannot be studied in isolation but must be considered as part of a larger entity to the total effect of which it contributes and from which it in turn gains meaning.

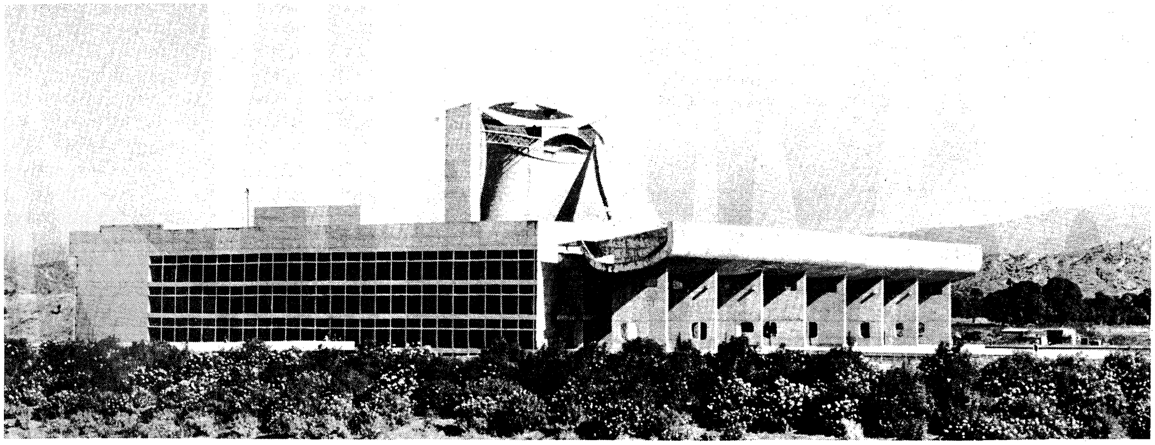
The subject matter of Indian sculpture is almost invariably religious. This does not mean that it cannot be understood as a work of art apart from its religious significance; but, at the same time, an understanding of its motivation and intent enriches one's appreciation. Much of what is represented is the recounting of legend and myth, particularly in the two centuries before Christ, when narrative relief was much in vogue. The work at this time, didactic and edificatory in intent, generally expresses itself in forms that are surprisingly earthy and sensuous. The anthropomorphic representation of the Buddha is avoided, and the subsidiary gods and goddesses are very much creatures of this earth. The Buddha image formulated around the 1st century AD is not what one would expect of the meditative, compassionate, Master of the Law; he is presented rather as an energetic, earthy being radiating strength and power.

ɹddhist
fluences

The foundations of traditional Hindu imagery were also laid about the same time that the Buddha image was first formulated: images with several arms, and sometimes heads, representing the Indian mind's attempt to define visually the infiniteness of divinity. In subsequent periods the image with many arms became a commonplace in Hindu, Buddhist, and Jaina iconography. Although the various pantheons expanded, they continued to share features of common derivation, expressing the belief that beyond the phenomenal multiplicity of forms lay the unity of the Godhead.

In addition to the major religions, there has always existed in India a substratum of folk beliefs and cults dedicated to the worship of powers that preside over the operation of the life processes of nature. These fertility cults, best expressed in the worship of the male and female divinities *yakṣas* and *yakṣīs*, played an important part in the development of Indian art. Among the perennial motifs that spring from the cults, those expressing life and

Paolo Koch—Rapho/Photo Researchers



The Legislative Assembly chambers of the states of Haryana and Punjab in Chandīgarh, India, designed by Le Corbusier, 1952.

abundance—such as the lotus, the pot overflowing with vegetation, water, or the like, the tree, the amorous couple, and above all the *yakṣa*s and *yakṣī*s themselves—are most significant. The images of these divinities, in particular, are the source of a great deal of artistic imagery and played a leading part in the development of iconographic types such as the images of the Buddha, the goddess Śrī, and other divinities. The maternal as the ideal of female beauty, which is manifested artistically in the emphasis on full breasts and wide hips, can be traced to the same beliefs. The very richness and exuberance of much Indian art is an expression of the view of life that equates beauty with abundance.

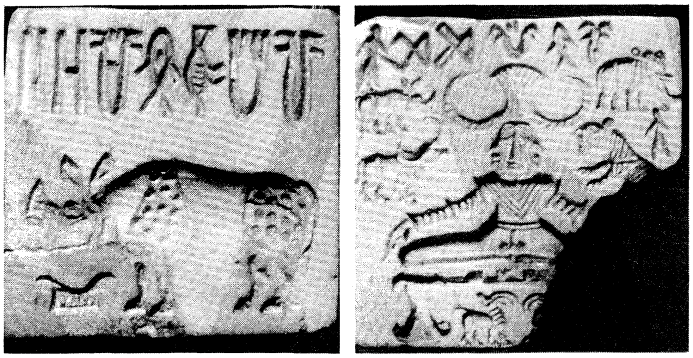
It is difficult to generalize about the style of a sculptural tradition that extended over a period of almost 5,000 years, but it is nevertheless clear that the distinguishing quality of Indian sculpture is its emphatic plasticity so obvious in Sānchi I and Mathurā sculpture from the 1st–3rd century AD. Forms are seen as swelling from within in response to the power of an inner life, the sculptor's function being to make these more manifest. At the same time a vision of form that is carved from without rather than modelled from within is also present, as for example at Bhārhut. The history of much of Indian sculpture, marked by periods of high achievement bursting with creativity followed by periods in which the potentialities so postulated are gradually worked out, is essentially the interaction of these two dominant tendencies.

Indus Valley civilization (c. 2500–1800 BC). Sculpture found in excavated cities consists of small pieces, generally terra-cotta objects, soapstone, or steatite, seals carved for the most part with animals, and a few statuettes of stone and bronze. The terra-cotta figurines are summarily modelled and provided with elaborate jewelry, which was fashioned separately and applied to the surface of the piece. Most of the work is simple, but a small group of human heads with horns are very sensitively modelled. Animal figures are common, particularly bulls, which are often carved with a sure understanding of their bulky, massive form. This plastic quality is also found in the humped bulls engraved on steatite seals, where the modelling is more refined and sensitive. A humpless beast, generally called a "unicorn," is another favourite animal, but it is frequently quite stylized. In addition to bisons, elephants, rhinoceroses, and tigers, seals are carved with images of apparent religious significance, often strongly pictographic.

Terra-cotta sculpture and seals

The terra-cotta sculpture and the seals both show two clear and distinct stylistic trends, one plastic and sensuous, the other linear and abstract. These appear during the same period and are also seen in the small group of stone and bronze sculptures that date from this period (National Museum, New Delhi). Of extraordinarily full and refined modelling is a fragmentary torso from Harappā, barely four inches (10 centimetres) high but of imposing monumentality; the same feeling for massive form is present in a lesser known bronze buffalo. A jaunty bronze dancing girl with head tilted upward (about 4½ inches [11 centimetres] high), from Mohenjo-daro, and a headless figure of a male dancer from Harappā, shoulders twisted in a circular movement, clearly demonstrate, in the attenuated and wiry tension of their forms, the second component of Indus Valley art. Of great interest is a famous bearded figure from Mohenjo-daro wearing a robe decorated with a pattern composed of trefoil motifs. The tight, compressed shape of the body and the expansive modelling of the head demonstrate that the two aspects of form revealed in Indus Valley art were not compartmentalized but interacted with each other. This can also be seen in the interplay of modelled form and textured surface frequently found in works produced by this civilization.

P. Chandra



Steatite seals of the Indus Valley civilization (*c.* 2300–*c.* 1750 BC). In the National Museum of India, New Delhi.

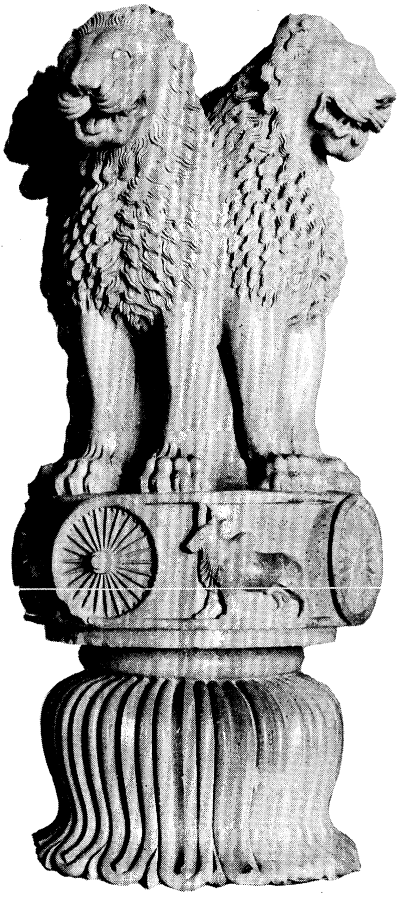
Maurya period (c. 3rd century BC). Little is known of Indian art in the period between the Indus Valley civilization and the reign of the Maurya emperor Aśoka. When sculpture again began to be found, it was remarkable for its maturity, seemingly fully formed at birth. The most famous examples are great circular stone pillars, products of Aśoka's imperial workshop, found over an area stretching from the neighbourhood of Delhi to Bihār. Made of fine-grained sandstone quarried at Chunār near Vārānasi (Benares), the monolithic shafts taper gently toward the top. They are without a base and, in the better preserved examples, are capped by campaniform lotus capitals supporting an animal emblem. The entire pillar was carefully burnished to a bright lustre commonly called the "Maurya polish." The most famous of these monuments is the lion capital at Sārnāth, consisting of the front half of four identical animals joined back to back. There is a naturalistic emphasis on build and musculature, and the modelling is hard, vigorous, and energetic, stressing physical strength and power. Very similar, if not at the same level of achievement, is the quadruple lion capital at Sānchi. Single lions are found at Vaiśālī (Bakhra), Rāmpurvā, and Lauriya Nandangarh. The Vaiśālī pillar is heavy and squat, and the animal lacks the verve of the other animals—features, according to some, designating it as an early work, executed before the Maurya style attained its maturity. By contrast, the Rāmpurvā lion, finished with painstaking and concise artistry, represents the style at its best. His smooth, muscled contours, wiry sinews, rippling, flamboyant mane, and alert stance reveal the work of a superior artist. An example at Lauriya Nandangarh is interesting because the pillar and the lion are both complete and in their original place, giving a clear idea of the column as it appeared to its contemporaries.

The Rāmpurv lion

The lion was the animal most often represented, but figures of elephants and bulls are also known. At Dhauli in Orissa, the fore part of an elephant is carved out of rock on a terrace above a boulder that carries several of Aśoka's edicts. The modelling here is soft and gentle, and the plump, fleshy qualities of the young animal's body, seen as emerging from the rock, are suffused with warmth and natural vitality. Since the contrast with the rather formal, heraldic lions could not be more complete, the sculpture clearly testifies to the simultaneous existence of a style different from that of the lion capitals. The style might very well represent the indigenous tradition of plastic form that appears consistently in later art and also in some of the animal capitals made in the imperial atelier, notably the damaged elephant that once crowned the pillar at Saṅkīsa and, above all, the splendid bull from Rāmpurvā. In this great work of art, the two opposing concepts of form merge in a work of harmonious power. The pronounced naturalism comes from the same source as do the lions, but the tense line and hard modelling yield to a form that wells from within and at the same time is given stability and strength by a vision imposed from without.

The sudden appearance of Maurya art with seemingly no tradition behind it has led to speculation that it was the creation of foreign artists, either Achaemenian or Hellenistic. Persian influence, particularly in the lotus capitals and the figures of lions can hardly be denied, but what is remarkable is the drastic reinterpretation of alien forms by Indian artists. This is a process that is repeatedly seen in the history of Indian art.

Besides the animal sculpture, some human figures, more or less life size, can also be assigned to the Maurya period, though scholarly opinion is by no means unanimous on the point. Among the most important are three images discovered at Patna (ancient Pāṭaliputra, the Maurya



Lion capital from Sārnāth, Uttar Pradesh, India, Chunār sandstone, mid-3rd century BC. Height 2.13 m.

P. Chandra

capital), two of which are representations of *yakṣa*s, the popular male divinities associated with cults of fertility, and the third, found at Dīdarganj (a section of Patna), a representation of a *yakṣī,* or female divinity. Stylistically the images are very similar. The standing *yakṣa*s (Indian Museum, Calcutta) are powerful creatures; the ponderous weight of their bodies, together with a certain refined appreciation of the soft flesh, is admirably rendered. The Dīdarganj *yakṣī* (Patna Museum), a masterpiece, displays the Indian ideal of female beauty, the heavy hips and full breasts strongly emphasizing the maternal aspect. In a nude torso discovered at Lopanipur, the sophisticated and sensitive treatment of the surfaces and the gentle blending planes that avoid all harsh accents produce a work of much refinement.

ıall
ıne
ıcs

Small stone discs (also called ring stones because several of them are perforated in the centre), found from Taxila to Patna, are clearly connected with the cult of a nude mother goddess. They represent Maurya sculpture on a smaller and more intimate scale but characterized by the same refined and exquisite workmanship. They are executed in bas-relief, which became the favourite form of sculpture in the subsequent period.

The terra-cotta art of the Maurya period is best represented by a substantial group of figurines, modelled for the most part, the clay sculptor performing work in his medium at the same level as the artist working in stone. Patna has yielded a large number of such works, but examples are found throughout the Gangetic Plain. The clothing and jewelry on the figurines are heavy and elaborate, the modelling, particularly of the head, is sensitive, and the expression is often one of great charm and refinement. There are also more archaic examples, distinguished by flat bodies, enormous hips, and modelled heads and breasts.

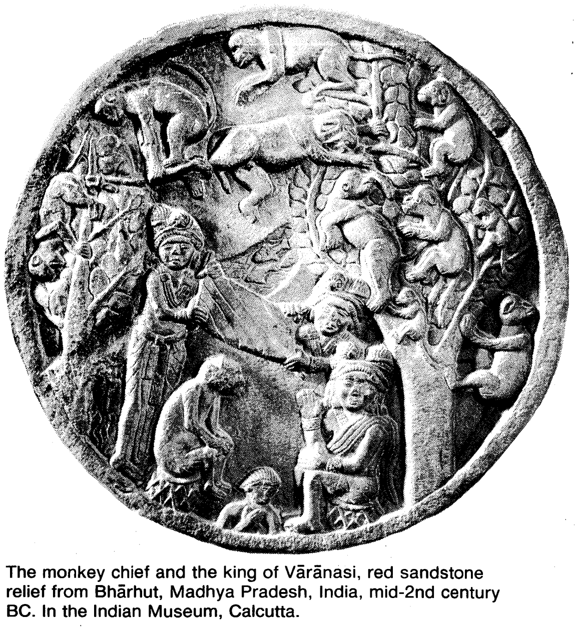
Indian sculpture in the 2nd and 1st centuries BC. The Maurya Empire collapsed in the early years of the 2nd century BC, and with it passed the art with which it was intimately related. The sculpture that is found throughout India from the middle of the 2nd century BC is startlingly different, but the process by which this change took place in a relatively short period of time is not fully understood. Several schools, sharing common features but nevertheless possessing distinct individual characteristics, are known to have existed. The history of the schools of northern India is somewhat obscure, largely due to the great destruction wrought in the Gangetic heartland; but there appears to have flourished there and in adjacent areas a school of great importance represented by the remains discovered at Bhārhut, Sānchi, Mathurā, and Buddh Gaya. Western India had its own school, as revealed in the sculptures decorating the cave temples, notably those of Bhājā, Pītalkhorā, and Kārli. In the southeast, the important school of Andhradeśa flourished in the Krishna River Valley at Amāravatī, Jaggayyapeta, and associated sites; and in eastern India, what is now the modern state of Orissa, made its contribution in the rock-cut sculptures at Udayagiri-Khandagiri. The distinctive schools, though spread over a subcontinent, were not isolated from each other. The contacts fostered by a flourishing trade and by the constant movement of pilgrims were always very close, and it was never long before developments in one part of India were echoed in another.

Judging from extant remains, artists of the earlier period (*c.* 3rd century BC) preferred figures carved in the round, relief sculpture being quantitatively quite insignificant. By contrast, it was sculpture in low relief that was favoured in the first two centuries before Christ; the earlier tradition was not quite forgotten, but figures carved in the round are relatively few. Although there is no stylistic difference, relief sculpture is here considered first according to the various regional schools, and sculpture in the round is treated separately.

Indian sculpture in the 2nd and 1st centuries BC: relief sculpture of northern and central India. Among the most important, and perhaps the earliest, remains in northern India are reliefs from the great *stūpa* at Bhārhut, dating approximately to the middle of the 2nd century BC. The work, suggesting a style imitating wooden sculpture, is characterized by essentially cubical forms, flat planes that meet at sharp angles, and very elaborate and precisely detailed ornamentation of surfaces. Most of the sculpture was confined to the railing of the *stūpa.* Some of the supporting posts bear large image of *yakṣa*s and *yakṣī*s of popular religion, now clearly pressed into the service of Buddhism, while most of the others are decorated with medallions in the centre and crescent-shaped motifs, or lunates, at the top and bottom, all filled with

Reliefs from the *stūpa* at Bhārhut

P. Chandra



The monkey chief and the king of Vārānasi, red sandstone relief from Bhārhut, Madhya Pradesh, India, mid-2nd century BC. In the Indian Museum, Calcutta.

lotus motifs. Some medallions contain amorous couples, the overflowing pot, the goddess Śrī standing on lotuses while being ceremonially bathed by elephants and other symbols of abundance; still others contain the earliest illustrations of events in the Buddha's life and of narratives of his former incarnations as related in the *Jātaka* tales (a collection of tales about the Buddha). Although compositions are crowded, great economy of expression is evident because the artist confines himself to the representation of essentials. Figures are often carved in horizontal rows, sometimes asymmetrically, adapting themselves awkwardly to the circular space of the medallion. Continuous narrative, in which events succeeding in time are shown in the same space, is often resorted to—the first occurrence of what was to become a favourite narrative technique. There is no attempt at establishing any interrelationship, psychological or compositional, between the various figures, each of which is strictly confined within its own space. The faces are masklike, without trace of emotion, lending a solemn and hieratic quality to their expression. Trapped between the background and a frontal plane beyond which they are not allowed to project, the figures are in a sense strictly two-dimensional, more so than in any other style of Indian sculpture. Often, however—particularly in the treatment of animals—the artist is more relaxed, giving glimpses of intimate observation and a natural rendering that anticipates the direction of future development. Like the posts, the top part, or coping, of the stone rail is also carved on both faces; on one of them is a continuous creeper bearing lotus flowers, leaves, and buds; on the other, again the winding stem of a creeper, but bearing other good things of life—such as clothes, jewelry, and fruits—and also scenes illustrating *Jātaka* stories.

Bhārhut is an extremely important monument inasmuch as it seems to mark a new beginning after the refined and naturalistic art of the Maurya Empire. The sophistication, in spite of the archaic, hieratic manner, would indicate that a considerable body of sculptural tradition, particularly in wood, preceded it; but of this no traces have survived. Be that as it may, Bhārhut states for the first time, and at some length, themes and motifs that would henceforth remain a part of Indian sculpture.

Stray finds of sculpture at Mathurā and other sites in modern Uttar Pradesh indicate that the Bhārhut style was spread over a large part of northern India, particularly the region roughly between that city and Vārānasi and Buddh Gaya in the east. A closely related style is also found at Sānchi in eastern Mālava, where a representative example is the sculpture of the railing of Stūpa II. Although the themes and motifs found at Bhārhut occur here, narrative representations are all but absent. The style is almost identical; the stiff and rigid contours are a little softer, but both the scale and richness of Bhārhut are missing.

Principal glory of Sānchi

It is the sculpture of the four gateways (*toraṇas*) of the Great Stūpa (Stūpa I) at Sānchi, however, that is the principal glory of that site, carrying the promise of the Bhārhut style to its fulfillment. The *toraṇas,* four in number, were attached to the plain railing around the middle of the 1st century BC. They consist of square posts with capitals supporting a triple architrave, or molded band, with voluted (turned in the shape of a spiral, scroll-shaped ornament) ends and a top crowned with Buddhist symbols. Bracket figures, in the form of *yakṣīs*, serve as additional supports. All parts of these gates, strongly reminiscent of wooden construction, are covered from top to bottom with the most exquisite sculpture. Subjects and motifs found at Bhārhut are also found here, the same profusely flowering lotus stem and associated motifs, the same compositions with figures basically arranged in horizontal rows, the same love for clear detail; but to all of these are added a truly voluminous sense of form, a smoother and more energetic movement, and a keen appreciation for the forms of nature, all of which endow the sculpture with a naïve and sensuous beauty unparalleled in Indian art.

Departures from the Bhārhut style are particularly striking in the narrative reliefs. Their greater depth, taken together with their crowded composition, results in the background, visible at Bhārhut, being submerged in shadow. The figures, in all their richness and abundance, flow out from the dark ground, secured in place by the frame of the panels. The Bhārhut angular silhouette and the rigid, severe outline of the body yields at Sānchi to a gently swelling plasticity, animated by a soft, breathing quality that molds the contours without strain or tension. There is a pronounced concern with the organization of composition, and the narration is often leisurely and discursive; the artist does not just tell the basic story but also lingers over the details, amplifying them to give a vivid picture of everyday life. The emotional monotone of Bhārhut survives in some Sānchi sculptures, but in others it is superseded by joyous faces and the emotional impact of vivid gesture and movement. Dejection is written large on the faces of the soldiers of Māra's army, who had tried to disturb the Buddha's meditation, as they stagger away from the scene of defeat, and the sensuousness of the amorous scenes is successfully evoked by the tender and intimate gestures of the couples. No longer transfixed in their own space, they turn to look at each other lovingly, responding to each other with a deeply felt understanding.

Long and elaborate bas-reliefs carved on the architraves of the *toraṇas* are the summit of the Sānchi sculptor's art. Among the finest are representations of the wars for the relics, the defeat of Māra, the *Viśvāntara Jātaka,* and the *Ṣaḍḍanta Jātaka.* The compositions are rich and crowded with figures, and are arranged with great skill. Particularly striking is the masterly handling of animals, notably the elephant, whose fleshy body and graceful movement are captured unerringly. Deer, water buffaloes, bulls, monkeys—all of the beasts and birds of the forests—are rendered with a sense of intimacy indicating the artist's sense of the fellowship of man and animal in the world of nature. The lush Indian landscape is often carved with ornamental trees, waterfalls, pools, mountains, and rivers. The Sānchi sculptor also shows a marked preference for architectural settings, filling his compositions with numerous buildings that often provide the spatial context for the action. Entire cities, with surrounding walls, elaborate gate houses, and palatial mansions, are depicted. Depth is achieved by rendering side views, and multiple perspective continues to be the rule.

Images of *yakṣīs*

The several large images of *yakṣīs* serving as brackets supporting the lowermost architraves of the *toraṇas* are unique achievements. Like the same goddesses at Bhārhut, they are shown in association with a tree to which they cling, but the style is remarkably different. The modelling shows a concern for the charms of the body, stressing the tactile nature of its flesh. The heavy jewelry and clothing that conceal the body are drastically reduced, revealing its nudity. The soft, melting sensuousness of the female form is so greatly emphasized that the belly and the folds of flesh at the waist are almost flabby, redeemed only by the smooth, firm breasts and the tender arms and limbs.

By comparison, reliefs adorning the railing around the Mahābodhi temple at Buddh Gaya (of about the same date or a little earlier) are in a somewhat impoverished idiom, lacking the rich proliferation both of Bhārhut and Sānchi. The posts have the usual medallions, lunates filled with lotuses, and reliefs depicting the familiar scenes of Buddhist myth and legend. The artistry of Buddh Gaya, however, is of a lower level of achievement than that at either Bhārhut or Sānchi: the relief is deeper than that at Bhārhut but shallower than that at the Great Stūpa of Sānchi; and crowded compositions are lacking, as are the clear and precise ornament and the rich floral motifs. The Buddh Gaya sculptor, however, though abbreviating even further the iconography of Bhārhut, breaks up, as does the Sānchi sculptor, the spatial isolation that so uncompromisingly separated each individual figure at that site.

The great school of Mathurā, also, seems to have come into existence about the 2nd century BC, though its period of greatest activity falls in the first two centuries after Christ. The city was repeatedly sacked in the course of the centuries, which may account for the paucity of materials, but enough has been discovered to reveal that the style, in its early stages, was very similar to that of Bhārhut, characterized by flat two-dimensional sculpture decorated with abundant and precise ornament. Several fragments discovered at the site show the gradual stages by which this style evolved, leading to the sculpture of the Great Stūpa at Sānchi on the one hand and to Buddh Gaya on the other.

Indian sculpture in the 2nd and 1st centuries BC: relief sculpture of Andhradeśa. Besides the schools of northern India, a very accomplished style also existed in southeast India; the most important sites are Jaggayyapeta and Amarāvatī, activity at the latter site extending well into the 2nd century AD. The early remains are strikingly similar to those at Bhārhut, the relief generally even shallower and the modelling comparatively flat. In contrast to those found in northern India, the proportions of the human body are elongated; but in its flat, cubical modelling, angular, halting contours, and precise, detailed ornamentation, the style is essentially similar to contemporary work elsewhere, right down to the same conventional clothing and jewelry. The nervous, fluid treatment of surfaces, so characteristic of subsequent Andhra sculpture, is already present here. The preferred material is marble rather than the sandstone invariably used in the north.

The style of the Andhradeśa school developed in a manner consistent with other regions of India, becoming more voluminous and shedding the early rigidity fairly rapidly. A group of sculptures at Amarāvatī are characterized by the same qualities that distinguish the work at the Great Stūpa of Sānchi: full and lissome forms, modelling that emphasizes mass and weight, and sensuously rendered surfaces.

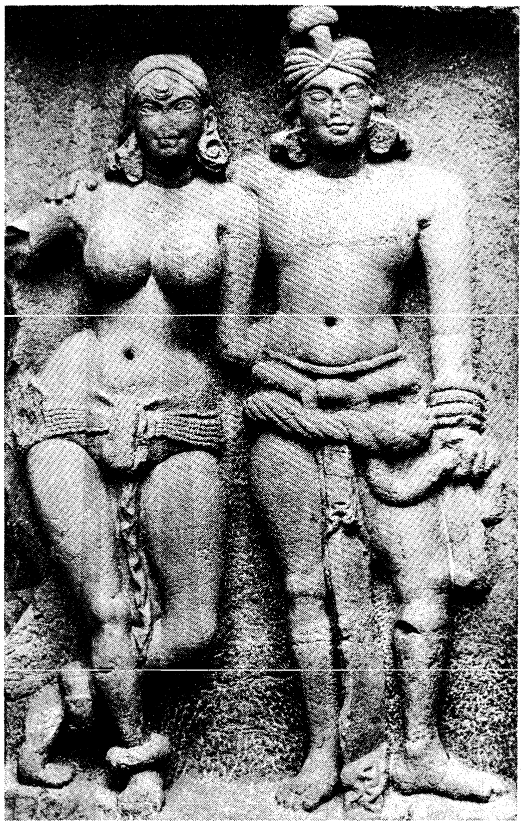
ɔck-cut
ve
mples
the
estern
hāts

Indian sculpture in the 2nd and 1st centuries BC: relief sculpture of western India. The numerous rock-cut cave temples in the Western Ghāts are, comparatively speaking, much less profusely adorned with sculpture than remains from other parts of India. The earliest works are undoubtedly the bas-reliefs on a side wall of the porch of a small monastery at Bhājā. They are commonly interpreted as depicting the god Indra on his elephant and the sun god Sūrya on his chariot but are more probably illustrations of the adventures of the mythical universal emperor Māndhātā. What is immediately evident is that these sculptures are not imitations of wooden prototypes, like those at Bhārhut, but, rather, reflect a tradition of terra-cotta sculpture, abundant examples of which are found in northern India and Bengal, where this medium was very popular because of the easy availability of fine clay. The terra-cotta tradition is reflected in the amorphous, spreading forms of Bhājā and in the fine striations used in depicting ornaments and pleated cloth, techniques natural and appropriate to the fashioning of wet clay. The fact that there are some similarities to the Bhārhut style—the stilted postures of the figures and the flat contours of the body, for example—indicates that the beginnings of the western Indian school would also have to be placed about the middle of the 2nd century BC.

The next major group of sculptures in western India have been found at Pītalkhorā. The colossal plinth of a monastery decorated with a row of elephants, the large figures of the door guardians, and several fragments recovered during the course of excavations are among the more important remains. A great proportion of the work represents an advance over the style of Bhājā, though features derived from terra-cotta sculpture continue to be found: the figures are carved in greater depth and volume, but the texture of the drapery, the soft contours of the body, and the high relief of the jewelry, which sometimes gives the impression of having been fashioned separately and then applied, testify to the continuing strength of the terra-cotta tradition. Although the hard line and sharp cutting of some sculpture is reminiscent of the earlier, wood-carving tradition as seen at Bhārhut, the forms are more appropriate to the stone medium. Moreover, the expression is more explicit; and for the first time, both gently smiling and boldly laughing figures of *yakṣa*s appear, as well as the figure of a lover blissfully drunk on wine offered to him by his beloved. These features are also found in the later sculpture of the Great Stūpa at Sānchi and, to a more pronounced extent, in the sculpture of the Mathurā school of the 1st centuries AD—for example, in the happily smiling *yakṣī*s from Bhutesar.

The cave temple at Kondane has, above the entrance hall, four beautiful panels depicting pairs of dancers. The forms retain the robust and full modelling of the more developed sculpture at Pītalkhorā, but to this is added an ease of movement and considerable rhythmic grace. Traces of the terra-cotta tradition are now totally absent; nor do they occur in the next phase, best represented by a group of sculptures found in the rock-cut temples and monasteries at Beḍsā and Nāsik and in the *caitya,* or temple proper, at Kārli. Sculpture at all these sites shows many affinities to the Great Stūpa at Sānchi and should be approximately contemporary or a little earlier. Easily the most outstanding achievements of this region and period, and for that matter one of the greatest achievements of the Indian sculptor, are the large panels, depicting amorous couples, located in the entrance porch of the Kārli *caitya.* Here the promise of early work achieves its fulfillment, the full weighty forms imbued with a warm, joyous life and a free, assured movement. The resemblance to work at the Great Stūpa of Sānchi is obvious, though these figures at Kārli are on a much larger scale and possess a massiveness and monumentality that is a characteristic of the distinct western Indian idiom.

P. Chandra



Amorous couple, detail of the *caitya* at Kārli, Mahārāshtra, India.

Indian sculpture in the 2nd and 1st centuries BC: relief sculpture of Orissa. Sculpture decorating the monasteries cut into the twin hills of Udayagiri and Khandagiri in Orissa represents yet another early Indian local idiom. The work is not of one period but extends over the first two centuries before Christ; the stages of development roughly parallel the styles observed at Sānchi Stūpa No. II, Buddh Gaya, and the Great Stūpa at Sānchi, but they possess, like other regional schools, fairly distinct and individual features. The earliest sculptures are the few simple reliefs found in the Alakāpurī cave, humble works that recall the bas-reliefs of Sānchi Stūpa II. The Mañcapurī, Tatowā Gumphā, and Anantā cave sculptures—particularly the image of Sūrya riding a chariot—are more advanced and resemble work at Buddh Gaya. The forms are heavy and solid and lack the accomplished movement of the later cave sculpture adorning the Rānī Gumphā monastery.

Sculpture decorating rock-cut monasteries

These, like other sculptures here, are in a poor state of preservation, but they represent the finest achievements at the site. Most remarkable is a long frieze, stretching between the arched doorways of the top story, representing a series of incidents that have not yet been identified. The work parallels that of the Great Stūpa at Sānchi, with the same supple modelling and crowded compositions. At the same time there is a nervous agitation, a fluid, agile movement together with a decided preference for tall, slender human figures. The reliefs on the guard rooms of Rānī Gumphā are also quite remarkable, depicting forested landscapes filled with rocks from which waterfalls flow into lakes that are the sporting grounds of wild elephants. The fine work of this cave strikes a romantic and lyrical note seldom found in Indian art.

Indian sculpture in the 2nd and 1st centuries BC: sculpture in the round and terra-cotta. The most important sculpture in the round are the life-size or colossal images of *yakṣa*s and *yakṣī*s, which reinterpret forms established by the two Patna *yakṣa*s and the Dīdarganj *yakṣī* of the Maurya period—very much as a few animal capitals, particularly the *makaras* (a crocodile-like creature) from Kauśāmbī and Vidiśā (Besnagar), echo the tradition of the superb Maurya animal capitals. It is the *yakṣa* figures, however, that deserve special attention, for they played a significant part in the iconographic developments of the 1st century AD and later and contributed substantially to the imagery of the anthropomorphic Buddha icon.

The most famous of the *yakṣa* images is a colossal figure recovered from the village of Pārkham, near Mathurā (Archaeological Museum). It is about 8⅔ feet (2.6 metres) in height, and, though the two hands are broken and the head is considerably damaged, it is an image of great strength. Its squat neck, its head set close to the body, which tends toward corpulence, its swelling belly restrained by a flat band, and a broad chest adorned with necklaces—all of these features contribute to an image turgid with earthy power. The back is flat and cursively finished, so that the figure has the appearance more of a bifacial relief than of an image carved in the round. Although the forms retain some of the cubical modelling of Bhārhut, the swelling limbs and torso have a massive weightiness that makes the image an appropriate representation of a divinity that presides over the productive processes of nature and endows plenty and abundance on his worshippers.

Mathurā *yakṣa* images

The Mathurā region seems to have been an important centre of *yakṣa* worship, for several images, most of them fragmentary, have been discovered there. Some images have also been found from the ancient city of Vidiśā (Vidisha Museum), one of which is even larger than the Pārkham example and is in a better state of preservation. The god holds a bag in one hand (the other was held below the chest), and the hair is tied in a large top knot over the forehead. The image is accompanied by a female consort (*yakṣī*), wide-hipped and full-breasted, who also emphasizes and personifies the powers of fertility.

The widespread nature of the cult is evidenced by the occurrence of *yakṣa* images throughout India. Fragments in the round (not to speak of the relief representations in a Buddhist context) of the 2nd to 1st centuries BC have been found from Madhyadeśa, Orissa, Rājasthān, Andhradeśa, and Mahārāshtra. At Pītalkhorā there is an exceptionally fine image of a *yakṣa* conceived as a potbellied dwarf carrying a shallow bowl on his head; the features, with a gently laughing mouth, are suffused with good humour. Similar *yakṣa*s, employed as atlantes (male figures used as supporting elements), are also found on the western gateway of the Great Stūpa at Sānchi and at other sites, notably Sārnāth.

The latest in the series of cult images is the image of the Yakṣa Maṇibhadra, from Pawāyā (Gwalior Museum). The sculpture is at present headless, but the rest of the body is well preserved. The right hand holds a fly whisk that flares over the shoulder; the modelling of the legs and torso is sensitive, and the folds of the garment wrapped around the body are full and voluminous, recalling the style of sculpture at Sānchi.

The terra-cotta sculpture of the period consists mainly of relief plaques made from molds found at numerous sites in northern India. These generally depict popular divinities; a richly dressed female figure loaded with profuse jewelry, obviously a mother goddess, is the favoured subject. Scenes from daily life also abound—as well as what appear to be illustrations of current myths and stories. Superb examples have been found from Mathurā, Ahichhatrā, Kauśāmbī, Tāmlūk, and Chandraketugarh. The workmanship is often of the most exquisite clarity and delicacy, the style paralleling that of contemporary stone sculpture.

Indian sculpture from the 1st to 4th centuries AD. This period is characterized by the dominance in northern India of the ancient school of Mathurā. Other schools, such as those that flourished at Sārnāth and Sānchi in the first two centuries before Christ, for example, were markedly restricted in their artistic output. Much of their sculpture was imported from Mathurā, and the few images they produced locally were strongly influenced by Mathurā work. The narrative bas-relief tradition, consisting of elaborate compositions of edificatory character, was on the wane, and the emphasis was on carving individual figures, either in high relief or in the round. For the first time, images appear of the Buddha, *bodhisattva*s, and various other divinities including specifically Hindu images representing the gods Vishnu, Śiva, Varāha, and Devī slaying the buffalo demon; some of these figures begin to feature several arms, a characteristic of later iconography. There are also many images of *yakṣī*s, often in most alluring attitudes and gestures. Their enticing bodies are now presented as unified organic entities, lacking all traces of the stiff, puppetlike aspect that had not been entirely overcome even at the Great Stūpa of Sānchi. During this period, also, a

P. Chandra

A Mathurā image of the Buddha discovered at Sārnāth, Uttar Pradesh, India, red sandstone, *c.* AD 81. In the Sārnāth Museum.

fresh incursion of foreign influence by way of western Asia was received, quickly assimilated, and transformed in the characteristic manner of Indian art.

e
ıool of
ındhāra

The school of Gandhāra, with Taxila in Pakistan as its centre and stretching into eastern Afghanistan, flourished alongside the Kushān school of Mathurā. It is of a startlingly different aspect, stressing a relatively naturalistic rendering of form, ultimately of Greco-Roman origin. The school evolved a distinct type of Buddha image and was also rich in relief sculptures depicting Buddhist myth and legend. Drawing largely on Indian traditions of composition, it nevertheless reinterpreted them in its own manner. The schools of Mathurā and Gandhāra were in close proximity and undoubtedly influenced each other, but essentially each adheres to its own concept of style.

The ancient Indian relief style found its fullest expression and development at neither Mathurā nor Gandhāra but in Andhradeśa, notably at the great sites of Amarāvatī and Nāgārjunikoṇḍa. Railing pillars and other parts of *stūpa*s decorated with *Jātaka* tales and scenes from the Buddha's life are found in great number and are of the most exquisite quality. Free-standing images of the Buddha, on the other hand, are relatively rare, being found only toward the close of the period.

Indian sculpture from the 1st to 4th centuries AD: Mathurā. One of the most important contributions of the school of Mathurā was the development of the cult image of the Buddha, who had been previously represented by aniconic (not made as a likeness) symbols. There is a certain amount of controversy about whether Mathurā or Gandhāra originated the Buddha image, which appears to be insoluble in view of the circumstantial nature of the evidence. It is possible that the two schools independently developed their own separate types of images; but, at least as far as the Mathurā image is concerned, it is clear that it is a natural development from the tradition of large *yakṣa* sculptures found in this region. The development can easily be seen in a famous image (discovered at Sārnāth and now in the Sārnāth Museum) of Mathurā manufactured and dedicated by the monk Bala. Carved in the round, the image is shown in a pose of strict frontality, the left hand held at the waist and the right arm, now damaged, originally raised to the shoulder—a posture immediately recalling that of the *yakṣa* images. The jewelry, however, is appropriately omitted, and the body is clothed in simple monastic garments. The modelling throughout is strong and sensuous, and the radiant energy of the body, its affirmative, outgoing movement, is more appropriate to the personality of a *yakṣa* than to that of the Buddha. This standing Buddha image, as seen in the Bala statue, is the standard Mathurā type, several examples of which are known. Along with this one, a similar, seated type developed, of which the best example is the splendid image known as the Kaṭra Buddha (Archaeological Museum). The modelling of the body is refined, the breasts characteristically heavy and prominent, and the flesh of the torso, with its subtle modulations, as convincingly rendered as the Bala image.

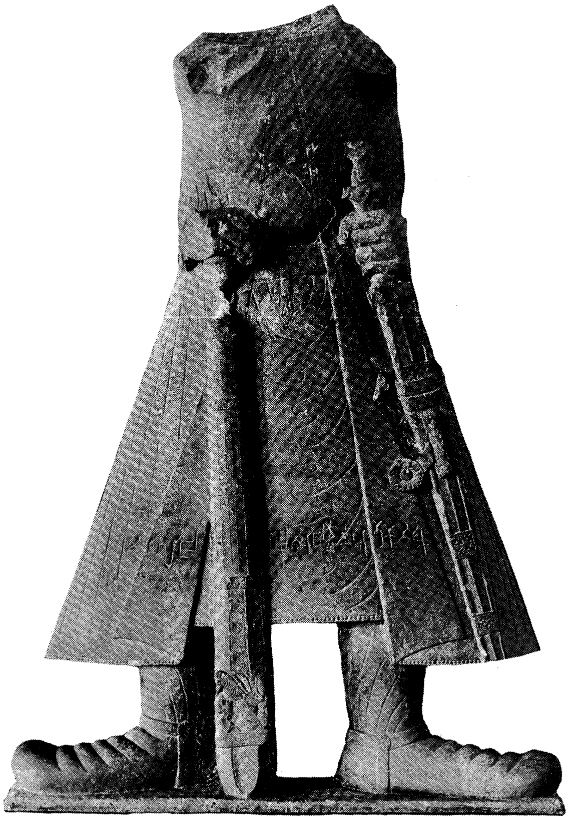
ɔne
ɔlets

The new trends formulated early by the Mathurā school do not indicate a sharp break from the traditions of the earlier schools. This is clear in a series of magnificent *āyāgapaṭas,* or stone tablets originally set up outside *stūpa*s to receive worship and offerings. They are usually square or rectangular and richly decorated with auspicious and religious symbols as well as angelic and mythical beings. The extremely decorative, lavish surface treatment gives the immediate impression of a great profusion of multiple forms, akin in feeling to the sculpture of the Great Stūpa of Sānchi. The organization of these forms, however, has none of the easy freedom of Sānchi. The figures, for example, are often cast in a regular, winding shape imitating the movement of the undulating lotus creeper. The same movement is seen in rows of animals depicted with haunches raised and chests touching the ground, features seen in earlier art but now much more emphatically stylized. The bodies of the animals also begin to be overpowered by vegetal forms, the tails, for example, terminating in foliate tips; in a later age, this tendency results in the almost total disintegration of animal shapes under the pressure of the floral.

It is not to these bas-reliefs, however, that one turns for the most delightful creations of the Mathurā school (for they are in fact the last vestiges of a style rapidly passing out of favour) but to the large number of railing pillars usually carved with representation of *yakṣī*s engaged in playful and enticing activities such as plucking blossoms from trees or leaning on its branches, dancing, bathing under a waterfall, and adorning themselves. Among the most beautiful of these is a group that was recovered from Kaṅkālī Ṭīlā and now in the State Museum at Lucknow. The modelling of the figures is generally heavy, the soft, plump bodies suffused with a slow, languorous movement. What is important, however, is the emotion, which is no longer expressed in the face alone but in the whole attitude of the body. The pensive mood of a woman holding a lamp, for instance, is evoked not only by the serene features of the face but by the gentle sway of the relaxed body. Present throughout is a fresh movement of life, a marked striving for diverse and varied effects of posture, movement, expression, and even dress and ornament that brings about vital changes in the nature of Indian sculpture. A remarkable group of railing posts decorated with *yakṣī* images, which were recovered from Bhūtēsar near Mathurā (Archaeological Museum), represent an even more refined achievement than the Kaṅkālī Ṭīlā figures. The heavy proportions, in spite of the full breasts and the wide hips, have been overcome; the happy faces express carefree joy, and the postures of the body are so alive with rhythm as to give the impression of a dancing figure.

Mathurā, during this period, was ruled by the Kushān (Kuṣāṇa) dynasty. A group of portrait sculptures of these rulers (Archaeological Museum), recovered from a village called Māt in the environs of Mathurā, gives an interesting glimpse of the foreign influences entering India at the time. One of them (unfortunately lacking the head) represents the emperor Kaniṣka wearing heavy boots, a tunic, and a coat, and leaning on a mace. The image is quite

P. Chandra



Headless portrait statue of the emperor Kaniṣka from Mathurā, Uttar Pradesh, India, *c.* late 1st century AD. In the Archaeological Museum, Mathurā, Uttar Pradesh, India. Height 1.85 m.

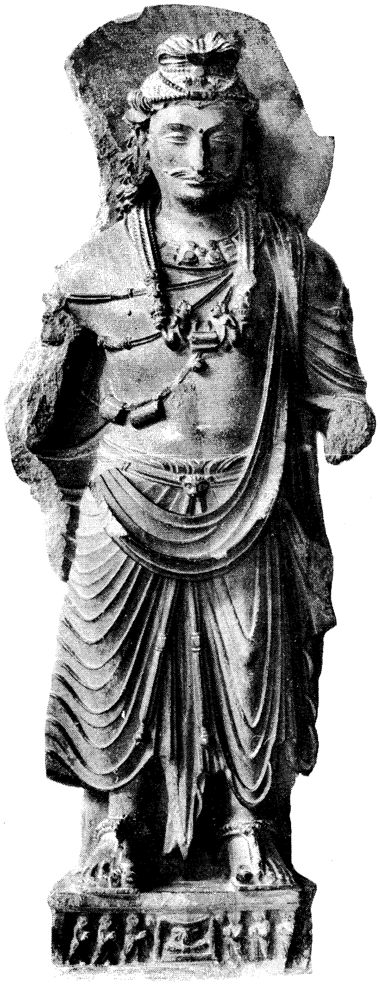
different not only in dress but also in style from other contemporary works, being essentially linear, with the forms entirely set into the surface. The surfaces have little ornamentation and are marked by extreme simplicity; they are also uncompromisingly stiff and rigid. It is possible that these images represent attempts by a Mathurā artist to imitate a style preferred by his imperial masters; but it was not long before the foreign elements were assimilated into the Mathurā style proper, for later images of Kushān chiefs have the same expanding and voluminous form that characterizes other sculptures of this school. A large number of ornamental motifs that now appear in India for the first time undergo a similar process of transformation.

The extent of Mathurā influence on Indian art of this period can be gauged by the sculpture of the school found at several sites in different parts of northern India, notably Ahichhatrā, Kauśāmbī, Sārnāth, and Sānchi. Most of these sites had been flourishing centres earlier, but only a very limited amount of sculpture was produced during the ascendancy of the Mathurā school; and whatever local sculpture was produced at this time was heavily influenced by the Mathurā style. At Sārnāth, for example, both the Bala Buddha imported from Mathurā and its local imitations have been found.

Ivory plaques of Bagrām

Ivory plaques discovered at Bagrām (Begrām) in Afghanistan are closely related to the school of Mathurā. These are of great importance; for, though ivory must have been a favourite medium of sculpture, little has been preserved of the early work. Most of it is in very low engraved relief, with fluent, sweeping outlines. The figures are depicted in easy and elegant postures, and the workmanship often attains considerable virtuosity.

P. Chandra



Bodhisattva of the Gandhāra school, schist sculpture, *c.* 2nd century AD. In the Municipal Museum, Allāhābād, Uttar Pradesh, India.

Indian sculpture from the 1st to 4th centuries AD: Gandhāra. Contemporary with the school of Mathurā, and extending almost into the 6th century, is the Gandhāra school, whose style is unlike anything else in Indian art. It flourished in a region known in ancient times as Gandhāra, with its capital at Taxila in the Punjab, and in adjacent areas including the Swāt Valley and eastern Afghanistan. The output of the school was very large; numerous images, mostly of Buddhas and *bodhisattvas*, and narrative reliefs illustrating scenes from the Buddha's life and legends have been found. The favoured material is gray slate or blue schist and, particularly during the later phases, stucco. Except for objects excavated at a few well-known sites (such as Taxila, Peshāwar, and the Swāt Valley, in Pakistan, and Jalālābād, Haḍḍā, and Bāmiān, in Afghanistan), most of the finds have been the result of casual discovery or clandestine treasure hunts and plunder, so that their correct provenance is not known. If to this are added the large variety of idioms that appear to have existed simultaneously and the total absence of securely dated images, the wide divergence of scholarly opinion with regard to the schools' evolution can be understood. In the present day, there is general agreement, however, that its most flourishing period probably coincided with Kushān rule, particularly the reigns of the emperor Kaniṣka and his successors, and that the school did not long outlast the growth of the Gupta school in the 5th century.

The origins of the Gandhāra style are ultimately Greco-Roman, though, recently, emphasis has been placed on Roman art as the more immediate source. It has also been suggested that the school was created by foreign craftsmen imported into India and by their Indian pupils.

The Gandhāra school is also credited by some scholars with the invention of the anthropomorphic Buddha image. Whether this is correct or not, the Gandhāra image is quite different from that of Mathurā and illustrates the difference between the two schools. Instead of the powerful images directly descended from *yakṣa* prototypes, the Gandhāra version is an adaptation of an Apollo figure, with rather sweet and sentimental features. The definite volume and substance given to the pleated folds of the monastic robes make this image more naturalistic than anything found in Indian art. At the same time, the iconographical features are of Indian origin. Large numbers of *bodhisattva* images conceived in the image of royalty, some with strongly individualized facial features, have also been found.

Narrative relief sculpture

In contrast to Mathurā, narrative relief sculpture was very popular in Gandhāra art. Again, in composition and iconography these reliefs are largely dependent on the earlier Indian schools, but the style is quite distinct. Instead of continuous narrative, incidents separated in time are separately represented, though often arranged in sequence. Violent emotions are realistically rendered. The compositions range from simple horizontal placement of figures to rich and complex arrangements, which often attempt to render space illusionistically.

In the course of time, Indian influence was increasingly felt in the art of Gandhāra, and an abstract vision began to obscure the Greco-Roman naturalism of the earlier forms. In spite of the new influence (and the many graceful but cloying stucco sculptures that are representative of this late phase) the style shows no signs of vital change. This conservatism, together with the large artistic production, gives an overall impression of considerable monotony. Without any real roots in India and with marked foreign features, the avenues of natural development seem to have been closed to the school, which thus finally disappeared. Nevertheless it made vital contributions to the art of Central and eastern Asia, and several features, drastically transformed, were incorporated in Gupta art.

Indian sculpture from the 1st to 4th centuries AD: Andhradeśa. Besides the schools of Mathurā and Gandhāra, a most accomplished school of sculpture flourished in Andhradeśa during the three centuries after Christ, the most important centres being Amarāvatī and Nāgārjunakoṇḍa. The remains consist mainly of carved railings and rectangular slabs that decorated the great Buddhist *stūpas*, which have largely disappeared. The finds are thus fragmentary and belong to several phases of construction or to separate monuments spanning the 1st, 2nd, and 3rd centuries AD.

Unlike the school of Mathurā, which concentrates on the carving of single figures, the Amarāvatī school carried to the fullest limit of its development the ancient tradition of relief sculpture, which flourished in the two centuries before Christ at sites such as Bhārhut, Sānchi, and Amarāvatī itself. The marble railing posts are decorated with central medallions and lunates at the top and bottom, all filled with lotus flowers of a very rich design. Often the medallions also contain reliefs illustrating scenes from the Buddha's life and from the *Jātaka* stories, and these are the principal glory of the site.

Two broad phases in the development of narrative relief can be distinguished. In the first, the artist builds on the achievements of early relief sculpture as seen on the Great Stūpa of Sānchi. The forms are still comparatively heavy, the figures increasingly soft and fleshy, the movement freer but still pervaded by a sense of calm repose. This type of work, represented by relatively few examples, is followed by a phase in which the compositions achieve an extraordinary elaboration and complexity. Most striking is the restless, energetic movement, often nervous and flurried, that possesses the participants in any given scene. Complex relationships and patterns are established between the figures; and space is so articulated that the eye participates in the swirling inner movement of the composition that effectually dissolves the ground on which the figures are carved, while the figures themselves flow out in an endless movement from the ground. The setting is dramatic in the extreme. The loving workmanship, reminiscent of ivory carving, and the superb technical proficiency mark the Amarāvatī reliefs as the culminating point of the entire relief style.

The figures, of both men and women, are of unprecedented suppleness and plasticity, the forms rendered in every variety of torsion and flexion. A fluent, gliding line, often more appropriate to painting than to sculpture, encloses the figures, and pervading the whole is a subtle voluptuousness. The reliefs are often only nominally religious, a pretext for the sculptor's pleasure in representing the leisured and sophisticated life of the time.

Nāgārjunakoṇḍa sculpture marks the last phase of the relief style. The figures become stiffer and puppet-like, the patterns of movement frozen and mechanical but still possessing the energy and richness that always characterize this style.

ddha
ages of
ıdhra-
ıa

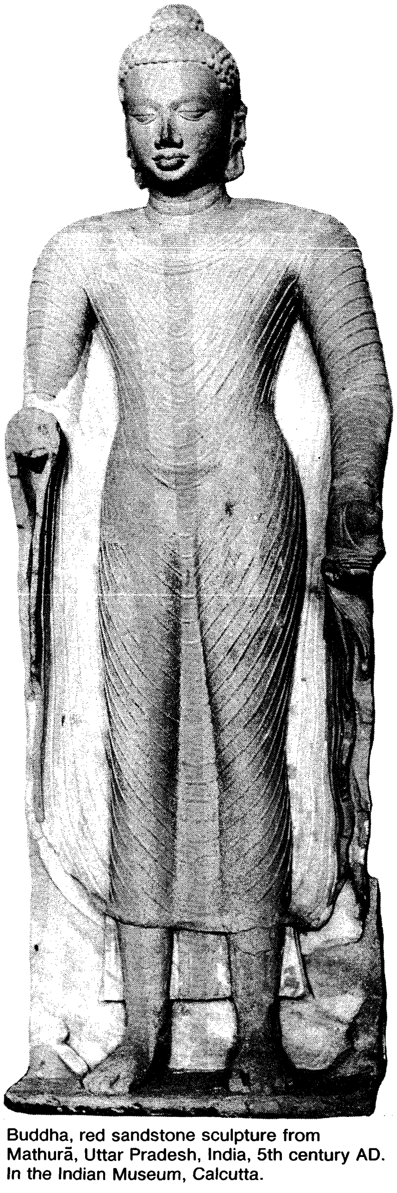
The Buddha is represented in Andhradeśa by both symbolic and anthropomorphic forms. The iconographic formula developed shows him clad in a rather thick garment with stylized folds, and the postures are not as formal and hieratic as the Mathurā. This type of Buddha exercised considerable influence in the development of the Buddha image in Ceylon. In several other features as well, the Andhra style also contributed to the development of early sculpture in Southeast Asia.

Indian sculpture from the 1st to 4th centuries AD: terracotta. The quality of terra-cotta figurines of this period is generally inferior to work produced in the first two centuries BC. Many heads of crude workmanship, with protruding eyes, apparently representing foreigners, were found at sites such as Mathurā, Ahichhatrā, and Kauśāmbī. At the same time, there are some well-modelled heads that imitate the style of stone sculpture and are equally expressive.

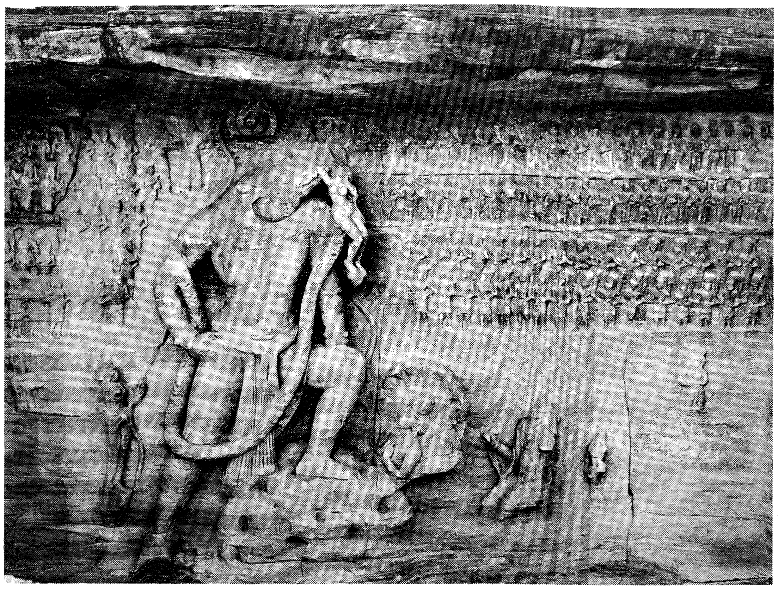
Gupta period (c. 4th–6th centuries AD). During the 4th and the 5th centuries, when much of northern India was ruled by the Gupta dynasty, Indian sculpture entered what has been called its classic phase. The promise of the earlier schools was now fully realized, and at the same time new forms and artistic ideals were formulated that served as the source for development in succeeding centuries. The more or less sensuous and earthy rendering of form was drastically transformed, so that artistic expression closely conformed to the religious vision. The forms are refined and treated with sure and unsurpassed elegance. The volumes, impelled by an inner life, still swell from within but are restrained and controlled, made to flow in smooth and abstract rhythms in an organic and unified concept in which the sensual and the spiritual are inextricably blended. The edificatory, didactic intent of early relief sculpture is abandoned; instead, the works produced are pronouncedly meditative; and the repose and calm that settles on the images of the Buddha, the master of the inner contemplative life, is also seen on images of other divinities. Decorative ornament is in perfect harmony with the volumes it adorns, each emphasizing the other, so that in every respect this classic style of the Gupta period is one of great composure and perfect balance.

Gupta period: Mathurā. The impetus for the new schools seems to have come from Mathurā, which is hardly surprising in view of the preponderant role played by the city in the preceding period. The transformation into the new idiom is best illustrated by a splendid image of the Buddha which is dated AD 384 (Indian Museum, Calcutta). Memories of the rather massive and ponderous weight of the earlier style are present, but the calm face no longer looks out at the world; rather, the vision is turned within, the mood being one of serene contemplation. The style, which consistently uses the local red sandstone, undergoes further refinement, seen in a series of magnificent life-size Buddha images of the 5th century (now scattered in museums throughout the world). The more delicate face radiates a feeling of calm inner bliss, and the body is most subtly modelled by smoothly flowing planes that both suggest the swelling force of life and subordinate it to

P. Chandra



Buddha, red sandstone sculpture from Mathurā, Uttar Pradesh, India, 5th century AD. In the Indian Museum, Calcutta.

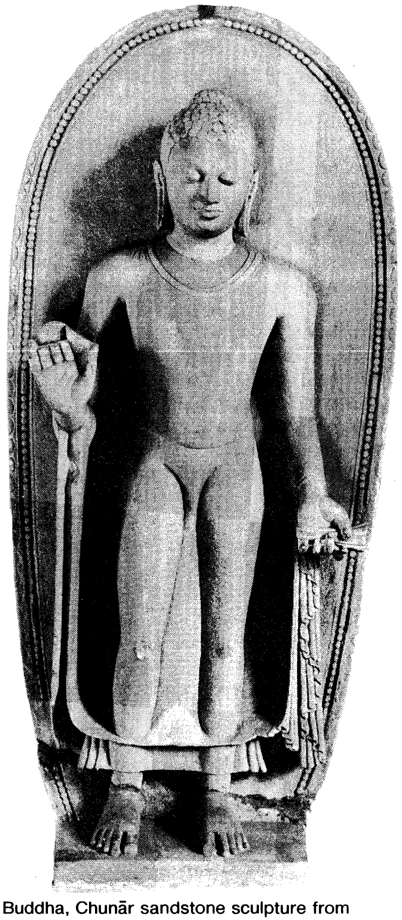


Vishnu rescuing the Earth goddess, sandstone relief panel from a cave at Udayagiri, Madhya Pradesh, India, early 5th century AD.

P. Chandra

the spiritual vision of the whole. Mathurā images generally show the Buddha wearing a diaphanous robe, the folds of which are rendered by stringlike ridges in a reinterpretation of a Gandhāra convention. The gestures of the hand are delicate and varied. The hair is usually rendered by rows of small curls that conceal the conical protuberance. These Mathurā images established an iconographical type that became the norm for the Buddha image.

P. Chandra



Buddha, Chunār sandstone sculpture from Sārnāth, Uttar Pradesh, India, 5th century AD. In the Indian Museum, Calcutta.

Mathurā Hindu images

In addition to the Buddha figure, Mathurā has yielded large numbers of images of the various Hindu divinities, particularly Vishnu-Krishna. This is in keeping with the increasing strength of the various Hindu cults and the intimate association of Mathurā with the god Krishna. The famous image of Vishnu from Kaṭrā Keśavadeva in Mathurā is one of the finest (National Museum, New Delhi). The god is conceived as a royal figure, wearing a crown and appropriate jewelry, his features imbued with a dignified calm that is suitable to his function as the preserver and is also characteristic of most Gupta art.

Gupta period: Sārnāth. This famous centre of Indian art developed a sweeter and more elegant version of the Buddha image than Mathurā's. Instead of the rather strict frontal posture, the weight of the body is thrown more on one leg, resulting in a very subtle contrapposto position, in which the hips, shoulders, and head are turned in different directions. This lends a certain movement to the figure, so that it does not quite possess the static, steadfast quality of Mathurā. The robes are no longer ridged with folds but are plain, and the surface of the stone is even more abstractly handled than is the Mathurā. The faces are heart-shaped, the transitions from one part of the body to another smoother, so that the images have great refinement even if they do not possess the strength of Mathurā. The characteristic Sārnāth style, the preferred material of which is the local buff Chunār sandstone, seems to have developed in the late 5th century, the few earlier works being closer to the Mathurā school. The most famous image from the site and one of the masterpieces of Indian art is that of the seated Buddha preaching (Sārnāth Museum). It is exceptionally well preserved and delicately carved. The face, with serene features and a gentle smile playing on the lips, suggests the joy of supreme spiritual achievement. The halo behind the Buddha is also very beautifully carved, with exquisite floral patterns. Large numbers of Buddha and *bodhisattva* images have been excavated at Sārnāth and are to be found in the museum at the site and in major collections throughout the world.

Gupta period: central India. In addition to the major schools of Sārnāth and Mathurā, important sculpture of the 5th and 6th centuries is found at several sites in central India. The sculptures here are often in their original locations, surviving not as isolated images torn from their architectural context but in association with the temples of which they formed a part. At Udayagiri, near Vidiśā, are a series of simple rock-cut caves of the opening years of the 5th century. The sculpture, made of soft stone, has suffered greatly, but whatever has survived reveals a style that stresses strength and power. Perhaps the most magnificent work is a great relief panel depicting the boar incarnation of Vishnu lifting the earth goddess from the watery deeps into which she had been dragged by a demon. The massive figure of the god, with the body of a man and the head of a boar, is carved in a surging movement across the face of the rock, the goddess resting easily on his shoulder, while a host of beings, human and divine, celebrate this great triumph.

The Śiva temple at Bhumarā has also yielded some sculpture of fine quality. The stone is carved with great precision and skill, nowhere more evident than in the handling of exuberant floral ornament. Little in Indian decorative sculpture can match the brilliance of the large panels filled with lotus stems and floriated scrolls discovered at this site and at Nāchnā Kutharā.

Some of the finest Gupta sculpture adorns the walls of the Vishnu temple at Deogarh. Particularly striking are three large relief panels depicting Vishnu lying on the serpent Śeṣa, the elephant's rescue, and the penance of Nara-Nārāyaṇa. The compositions tend to be dramatic; the carving and decoration, sumptuous, the sturdy forms recalling Mathurā rather than the attenuated grace of Sārnāth. The doorframe of the sanctum of this temple is an especially fine example of architectural decoration popular in this period. Bands of floral scrolls, amorous couples, and flying angels of great elegance are carved around the entrance. Particularly impressive are groups of worshippers at the base, their swaying bodies related to each other with an easy rhythm.

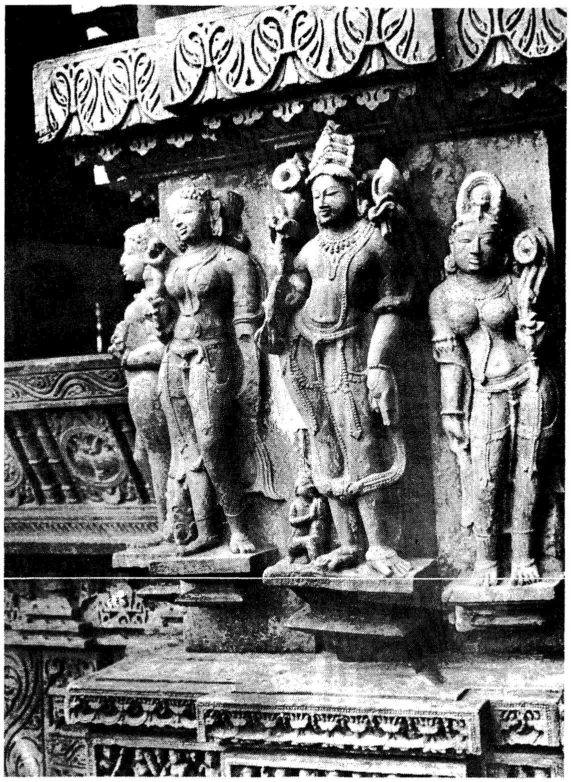
Gupta period: Mahārāshtra. A great revival of artistic activity seems to have taken place in this region during the reign of the Vākāṭaka dynasty and its successors, best expressed in the splendid sculpture decorating the cave temples of Ajantā and Elephanta. The idioms established in the North were adapted here to the needs of a style that conceived figures on a massive scale, as determined by the demands of the great expanses of rock out of which they were carved. Although the sculpture at Ajantā (mostly of the late 5th century) combines the old weightiness with the new restraint and elegance, the style finds its supreme expression in the magnificent cave temple at Elephanta. The central image of this great temple is of immense size and in deep relief. It represents Śiva in his cosmic aspect, the central head clam, introspective, self-sufficient, and transcending time, the heads to the sides, in their sensuous beauty and awesome terror, reflecting the creative and the destructive aspects of the supreme divinity.

ulpture
the
antā and
ephanta
ves

Gupta period: other regions. The impact of the Gupta style of the 5th and 6th centuries was felt in many parts of India, though actual remains thus far discovered are more abundant in some parts than in others. There appears to have been, in Bihār, a distinct school characterized by rather heavy, compact forms; and Gujarāt and southern Rājasthān developed an individual style of considerable voluptuousness and plasticity. Among the notable sculpture of the Idar region are groups of mother goddesses whose massive forms are rendered with an easy grace and intimacy. In the Karnataka country, to the south, the cave temples of Bādāmi reveal yet another distinct idiom, somewhat direct and elemental but nevertheless belonging to the same general style, with local variations, that prevailed over the greater part of India.

Gupta period: terra-cotta. Terra-cotta sculpture, like art in other mediums, was greatly developed. Fairly large and elaborate plaques were used to adorn brick *stūpa*s and Hindu temples from Sind to Bengal. The polychrome relief images of the Buddha from Mīrpur Khās are delicate and slender, with traces of Gandhāra feeling. Representations of divinities and mythological scenes from temples in Bīkaner, Ahichhatrā, Bhītargaon, and Śrāvastī are works on a more popular level, possessing an earthy ponderousness. A large number of figurines, particularly fragments of heads with elaborate coiffures and delicate, smiling features, have been found at Rājghāt in Vārānasi (Benares) and at other sites.

Medieval Indian sculpture. Indian sculpture from the 7th century onward developed, broadly speaking, into two styles that flourished in northern and southern India, respectively. In each of these regions there also developed additional local idioms, so that there was a wide variety of schools. All, however, evolved in a consistent manner, the earlier phase marked by relatively plastic forms, the later phase by a style that emphasizes a more linear rendering. The sculpture was used mainly as a part of the architectural decor, and the quantity required was vast. This often entailed a mechanical production, with the result that works of quality are few in proportion to the numbers.



Detail of a wall of the Lakṣmaṇa temple at Khajurāho, Madhya Pradesh, India, sandstone, *c.* AD 941.

P. Chandra

Besides the two main idioms, the local schools of Mahārāshtra and Karnataka are of particular interest because they possess considerable individuality and often show both northern and southern features.

Sculpture in bronze was also produced in fairly large quantities in this period. Again, several local schools can be distinguished, the most important of which are those of eastern and southern India.

Medieval Indian sculpture: North India. The history of North Indian sculpture from the 7th to the 9th centuries is one of the more obscure periods in Indian art. Two trends, however, are clear: one exhibits the decline and disintegration of classical forms established during the 5th and 6th centuries; and the other, the evolution of new styles that began to possess overall unity and stability only in the 10th century.

A breakdown of the Gupta formula is observable from at least the 7th century onward, if not a little earlier: harmonious proportion, graceful movement, and supple modelling begin to yield to squat proportions, a halting movement, and a more congealed form. Toward the 8th century, signs of a new movement become evident in a group of sculptures that departs from the progressively lifeless working out of the Gupta idiom. The modelling emphasizes breadth but with a pronounced feeling for rhythm, and the delineation of decorative detail is fairly restrained. In the 9th century, particularly during the second half, a distinct change came over the styles of all of northern India. A new elegance, a richer decorativeness, and a staccato rhythm so characteristic of the medieval styles of the 10th and 11th centuries begin to be clearly seen and felt. Sculpture of this period reaches a standard of elegance never surpassed in the medieval period: the grace and voluminousness of earlier work are modified

Signs of
a new
movement

but not lost; the harsh angularity of later work, avoided. An idea of the style can be formed from an important group of sculptures at Ābānerī, the Śiva temple at Indore, and the Telī-kā-Mandir temple at Gwalior, as well as from individual works in various North Indian museums.

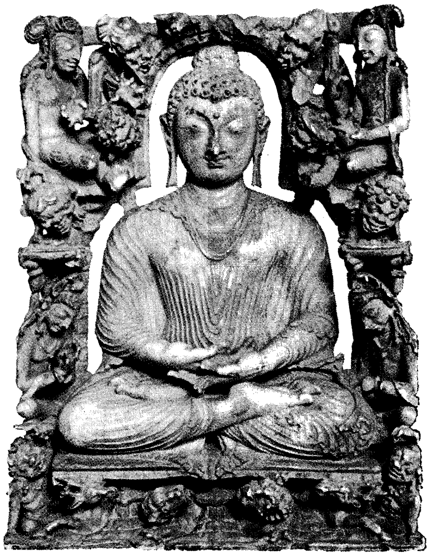
With the 10th century, the conventions of North Indian sculpture became fairly well established. The style is represented by examples from such monuments as the Lakṣmaṇa temple at Khajurāho (dated 941), the Harasnāth temple at Mt. Harsha (*c.* mid-10th century), in Rājasthān, and numerous other sites scattered all over northern India. These works are executed in a style that has become harder and more angular, the figures covered with a profusion of jewelry that tends to obscure the forms it decorates. These features are further accentuated in the 11th century, when many temples of great size, adorned with prodigious amounts of sculpture, were erected all over northern India. There is a decline in the general level of workmanship: the carving is often entirely conventional and lifeless, the features rigid and masklike, and the contours stiff and unyielding. The ornamentation, consisting of a profusion of beaded jewelry, is for the most part as dull, repetitive, and lifeless as the rest of the sculpture. This phase of artistic activity is represented at important centres from Gujarāt to Orissa; one of them is Khajurāho, with a vast amount of sculpture, all in a good state of preservation but conceived and executed as perfunctory architectural ornamentation. Not all sculpture, however, is of inferior quality; the hard, metallic carving and angular, stylized line sometimes result in works possessing a cold brilliance.

The 12th century marks the end of traditional sculpture all over northern India, except for a few pockets not yet penetrated by the Islāmic invasions. A rigid line imposed itself on the forms, which in turn became desiccated and hard, so that whatever unity of surface may have existed was entirely shattered. A brief revival took place in parts of Gujarāt and Rājasthān in the 15th century, but the sculpture merely imitated the work of the late medieval phase. The pure geometry of their forms, however, sometimes results in works possessing a curious archaistic power.

Sculpture in eastern India

Sculpture in eastern India (consisting of Bangladesh and the modern Indian states of Bihār, West Bengal, and Orissa), though sharing in the broad pattern of development of the rest of northern India, nevertheless represents a distinct idiom. The flatness of planes and angularity of contours are less pronounced, the figures retaining a sense of mass and weight for a greater period of time and to a greater degree. This can be clearly seen in sculpture from Konārak, in Orissa. Dating to the 13th century, the style retains a considerable semblance of plasticity at a period when sculpture in other parts of northern India had assumed a very wooden appearance. In Bihār and Bengal a flourishing school of bronze sculpture also developed, as evidenced by the large number of finds, notably from the sites of Nālandā and Kurkihār. The style generally parallels works in stone, emphasizing plastic values to a great degree. The most flourishing period was the 9th century, when a series of magnificent images representing the gods and goddesses of the Buddhist pantheon were made at Kurkihār and Nālandā. The work of the 10th and 11th centuries is more decorative and often very skillfully and elaborately cast. Of relatively small size and therefore easily transportable, bronze sculpture from this area played an important part in the diffusion of Indian influence in Southeast Asia.

Kashmir sculpture tends to be weightier and more massive than works in other parts of India. Some Gandhāra memories surivive, particularly in the fleshy rendering of the body and the drapery, but the sculpture is very much a part of the stylistic developments in northern India. Representative examples of the style, dating to around the mid-9th century, have been found from Avantipura. A flourishing school of bronze sculpture also existed, numerous examples having come to light in recent years. One of the finest, discovered at Devsar (Sir Pratap Singh Museum, Srinagar), is a large 9th-century ornamental frame, 6½ feet (two metres) high, decorated with various incarnations of Vishnu, all filled with great energy and movement. A good number of ivory images of Kashmir workmanship have also been preserved. These are generally of miniature size, polychromed, and of extremely fine and delicate workmanship. Influences of the Kashmir style of sculpture were strongly felt in the neighbouring Himalayan region, including both Tibet and Nepal.

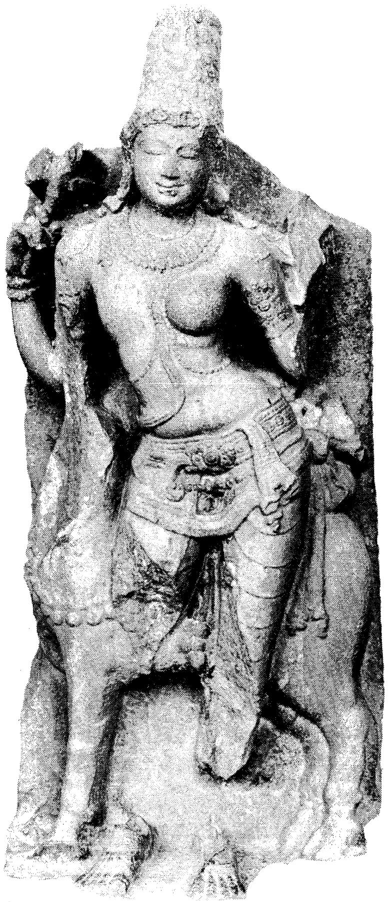


Seated Buddha with attendants, carved ivory sculpture from Kashmir, *c.* 8th century AD. In the Prince of Wales Museum of Western India, Bombay. Height 10 cm.

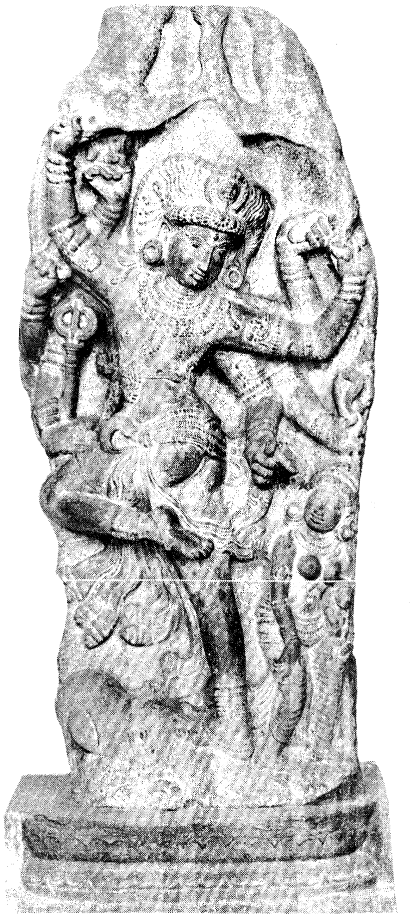
P. Chandra

Medieval Indian sculptures: southern India. The medieval phase in southern India opened with elegant 7th-

P. Chandra



Śiva-Ardhanārīśvara, granite sculpture from the Ugrakaliamman temple, Thanjāvūr, Tamil Nadu, India. In the Thanjāvūr Museum and Art Gallery, Tamil Nadu. Height 1.14 m.



Śiva slaying the elephant demon, granite sculpture from Dārāsuram, Tamil Nadu, India, 13th century AD. In the Thanjāvūr Museum and Art Gallery, Tamil Nadu.

P. Chandra

century sculptures at Mahābalipuram, by far the most impressive of which is a large relief depicting the penance of Arjuna (previously identified as an illustration of the mythical descent of the Ganges). It is carved on the face of a granite boulder with a deep cleft in the centre, representing a river, down which water actually flowed from a reservoir situated above. On both sides are carved numerous figures of divinities, human beings, and animals that crowd the hermitage where Arjuna, practicing penance, is visited by Śiva. The tall, slender figures, with supple tubular limbs, remotely recall the proportions of Amarāvatī, now greatly transformed; and the numerous animals, including the elephant herd with its young, show the same intimate feeling for animal life that characterizes all Indian sculture, but in a manner that has seldom been surpassed.

The light, aerial forms gained stability and strength in subsequent centuries, culminating in superb sculptures adorning small, elegant shrines built during the late 9th century when the Cōḻa dynasty was consolidating its power. The temples at Tiruvalīśvaram, Kodumbālūr, Kilaiyur, Śrīnivāsanālūr, Kumbakonam, and a host of other sites of this period are only sparingly adorned with sculpture, but it is of superb quality. With the 10th and 11th centuries, South Indian sculpture, like its counterpart in the north though to a lesser degree, was carved in flatter planes and more angular forms, and the fresh, blooming life of earlier work is gradually lost. This can be seen, for example, in the sculpture of the numerous temples of Thanjāvūr and Gaṅgaikoṇḍacōḻapuram. The subsequent phase, extending up to the 13th century, is represented by work at Dārāsuram and Tribhuvanam; although the forms become increasingly congealed, brittle works of fine quality—often capturing outer movement with great skill—continue to be produced. Sculpture in southern India continued when artistic activity was interrupted in the north by the Islāmic invasions but, in spite of technical virtuosity, became progressively lifeless. Artistic activity continued in the south into the 17th century, the elaborately sculptured halls at Madurā and the masses of stucco sculpture adorning the immense entrances, or *gopuras*, testifying to the prodigious output and the undistinguished quality of the work produced.

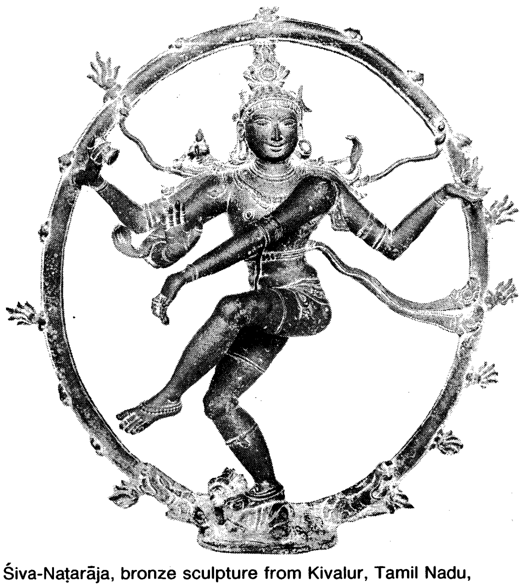
South Indian bronze sculpture

South Indian bronze sculpture has a special place in the history of Indian art. A large number of images were made (some of them still in worship in the mid-20th century and others unearthed from the ground by chance), but examples before the 8th century are quite rare. In bronze, as in stone, the 9th and 10th centuries were periods of high achievement, and many images of excellent quality have survived. They are all cast by the lost-wax, or cire perdue, process (in which a wax model is used) and technically are very accomplished. In the early stages the forms were smooth and flowing, with a fine balance maintained between the body and the complex jewelry, the lines of which follow and reinforce every movement of the plastic surface. The bronzes of the later period lose this cohesiveness, the ornament, by virtue of its hardness, tending to divide and fragment the body it covers. The modelling also became flatter and sharper, though not quite as rapidly in bronze sculpture as in stone. Ancient traditions of workmanship survive to the present day, and a few guilds of craftsmen continue to make competent if somewhat lifeless images.

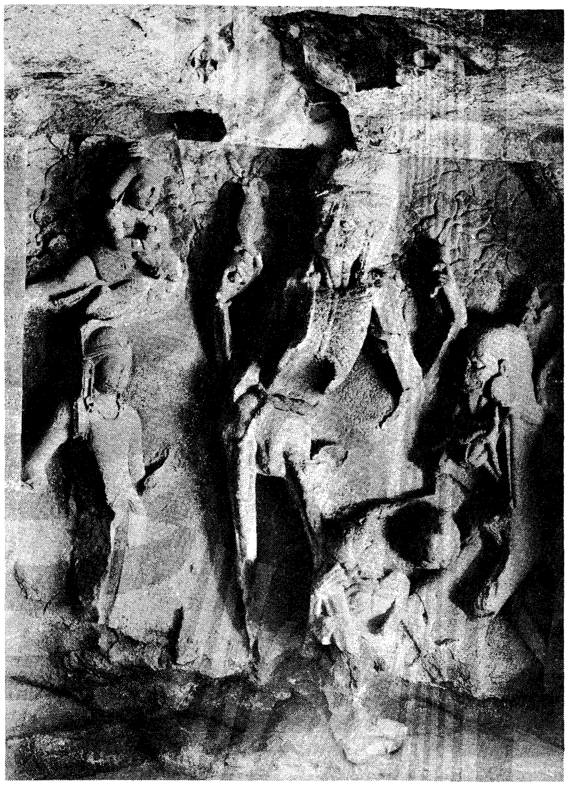
Most South Indian bronze images are representations of Hindu divinities, notably Vishnu and Śiva. One particular form deserves special notice as a striking southern contribution to Indian iconography. It is that of a four-armed Śiva as Lord of the Dance (Naṭarāja), shown within a flaming halo, or aureole, one hand holding the doubleheaded drum symbolizing sound, or creation, and the other holding the fire that puts an end to all that is created. The palm of the third hand faces the devotee, assuring him of freedom from fear, while the fourth hand points to the raised foot, the place of refuge from ignorance and delusion, which are symbolized by the dwarf demon crushed beneath the other foot. Several splendid images are known, the finest being, perhaps, the great image still worshipped in the Bṛhadīśvara temple at Thanjāvūr.

Medieval Indian sculpture: Mahārāshtra and Karnataka. The Karnataka country possessed a flourishing school of sculpture in the 7th and 8th centuries, as seen in examples from Aihole, Pattadkal, and Ālampur. As in architecture, influences from the north are discernible, but the style is basically southern, emphasizing rugged strength and power compared to the more elegant and delicate forms of the Tamil country. In Mahārāshtra, cave temples at

P. Chandra



Śiva-Naṭarāja, bronze sculpture from Kivalur, Tamil Nadu, India, 11th century AD. In the Thanjāvūr Museum and Art Gallery, Tamil Nadu. Height 81 cm.



Dancing Kālī, relief from the Rameśvara cave, Ellora, Mahārāshtra, India, *c.* 7th century AD.
P. Chandra

Ellora carry the most important examples of this phase of sculpture. Here the tradition is continued of images of great size that, in their primitive strength, partake of the nature of the rock out of which they are carved. A series of large, splendid panels (6th century AD) depicting incidents from Hindu mythology in high relief are to be found in the Rameśvara cave; notable among them is a fearsome representation of the dancing Kālī, goddess of death. The Kailāsa temple (*c.* 757–783) has a remarkable group of elephants struggling with lions all around the plinth. Of the several large reliefs, also at Kailāsa, the depiction of Rāvana shaking Kailāsa is a composition of considerable grace and charm.

Distinctive style of Karnataka

Toward the 13th and 14th centuries, a very distinctive style developed in the Karnataka country, which was then largely ruled by kings of the Hoysaḷa dynasty. The materials employed are varieties of stone that are soft when freshly quarried but harden on exposure, which may account partially for the extreme richness of the work. The sculpture is in very high relief, often undercut and literally covered with the most elaborate ornaments and jewelry from top to toe. This unrestrained extravaganza is unique even for Indian art, which shows a preference for intricate and elaborate ornament at all stages of its history.

Painting. Literary works testify to the eminence of painting as an art form in India, particularly in the decoration of walls, but climate has taken a devastating toll, leaving behind only a few tantalizing examples. By far the bulk of the preserved material consists of miniature painting, initially done on palm leaf but later on paper. The subject matter is generally religious (illustrating divinities, myths, and legends) and literary (illustrating poetry and romances, for example), though the Mughal school was also concerned with historical and secular themes. The styles were rich and varied, often closely connected with one another and sometimes developing and changing rapidly, particularly from the 16th century onward. The work also shows a surprising vitality under strained circumstances, surviving up to the very eve of the modern period when the other arts had deteriorated greatly.

Prehistoric and protohistoric periods. Painting in India should have a history stretching as far back as any of the other arts but, because of its perishable nature, little has survived. None of the examples found in rock shelters over almost all of India, and chiefly representing scenes of hunting and war, appears to be earlier than the 8th century BC, and all may be as late as the 10th century AD. A faint idea of the painter's art in the Indus Valley civilization can be had from the pottery, elaborately decorated with leaf designs and geometrical patterns.

Paintings at Ajantā

Ancient wall painting. The earliest substantial remains are those found in rock-cut cave temples at Ajantā, in western India. They belong to the 2nd or 1st century BC and are in a style reminiscent of the relief sculpture at Sānchi. Also found at Ajantā are the most substantial remains of Indian painting of about the 5th century AD and a little later, when ancient Indian civilization was in full flower. The paintings, the work of several ateliers, decorate the walls and ceilings of the numerous cave temples and monasteries at the site. They are executed in the tempera technique on smooth surfaces, prepared by application of plaster. The themes, nominally Buddhist, illustrate the major events of the Buddha's life, the *Jātaka* tales, and the various divinities of the expanding Buddhist pantheon. The ceilings are covered with rich motifs, based generally upon the lotus stem and the world of animals and birds. The style is unlike anything seen in later Indian art, expansive, free, and dynamic. The graceful figures are painted by a sweeping and accomplished brush; and they are given body and substance by modelling in colour and by a schematic distribution of light and shade that has little to do with scientific chiaroscuro. The narrative compositions, handled with utmost dexterity, are a natural outgrowth of the long traditions of relief sculpture and reflect the splendour and maturity of contemporary sculpture. The large images of the *bodhisattvas* in Cave 1, combining rich elegance with spiritual serenity, reflect a vision that sees the shifting world of matter and the transcendental calm of Nirvāṇa as essentially one.

Except for a large and magnificent painting of a dance scene found at the rock-cut cave at Bāgh—a painting executed in a style closely resembling Ajantā—hardly any other work of this great period survives. Cave temples at Bādāmi, in the Karnataka country, and Sittānavāsal, in Tamil Nadu, probably of the late 6th and 7th centuries AD are already but echoes of the style of the 5th century, which appears to have died out around this time.

Eastern Indian style. Small illustrations on palm leaf, chiefly painted at the great Buddhist establishments of eastern India, appear to have conserved some elements of this ancient style; but they have lost its dramatic impact, which is replaced by a studied preciosity and an inhibited meticulousness. The surviving paintings date from the 11th and 12th centuries and are conventional icons of the numerous Buddhist gods and goddesses, narrative representations having largely disappeared. With the destruction of these Buddhist centres by the Islāmic invader, the east Indian style seems to have come to an end.

Western Indian style. The style of Ajantā is succeeded in western India by what has been appropriately named the western Indian style. Among the earliest examples are a few surviving wall paintings of the Kailāsa temple (mid-8th century) at Ellora and the Jaina temples, built at the same site a hundred years later. The plastic sense of form, so evident at Ajantā, is emphatically replaced by a style that even at this early stage is heavily dependent on line. The contours of the figures are sharp and angular, the forms dry and abstract; and the fluent, stately rhythms of Ajantā have become laboured and halting.

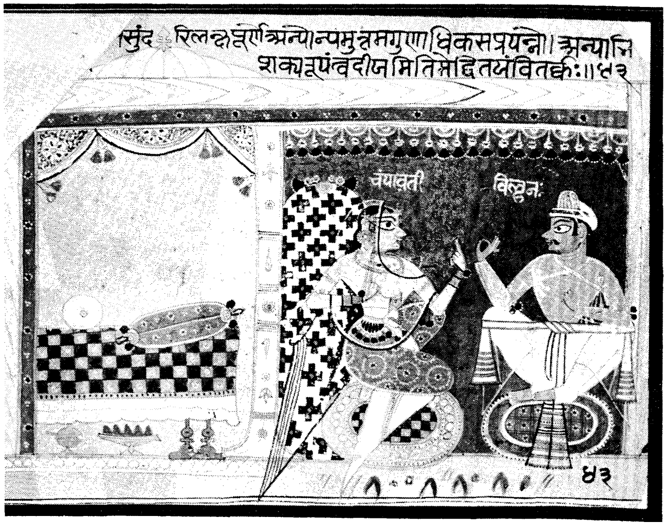
Jaina manuscripts

The most copious examples of this style, however, have survived not on the walls of temples but in the large number of illustrated manuscripts commissioned by members of the Jaina community. The earliest of these are contemporary with eastern Indian manuscripts and are also painted on palm leaf; but the style, instead of attempting to cling to ancient traditions, moves steadily in the direction already established at Ellora. It is a perfect counterpart of contemporary sculpture in western India, relying for its effect on line, which progressively becomes more angular and wiry until all naturalism has been deliberately erased. The figures are almost always shown in profile, the full-face view generally reserved for representations of the *tīrthaṅkaras*, or the Jaina saviours. A convention that appears unfailingly for the duration of the western Indian style is the eye projecting beyond the face shown in profile, meant to represent the second eye, which would not be visible in this posture. The colours are few and pure: yellow, green, blue, black, and red, which was preferred for the background. In the beginning, the illustrations are simple icons in small panels; but gradually they become more elaborate, with scenes from the lives of the various Jaina saviours as told in the *Kalpa-sūtra* and from the adventures of the monk Kālaka as related in the *Kālakācāryakathā* the most favoured.

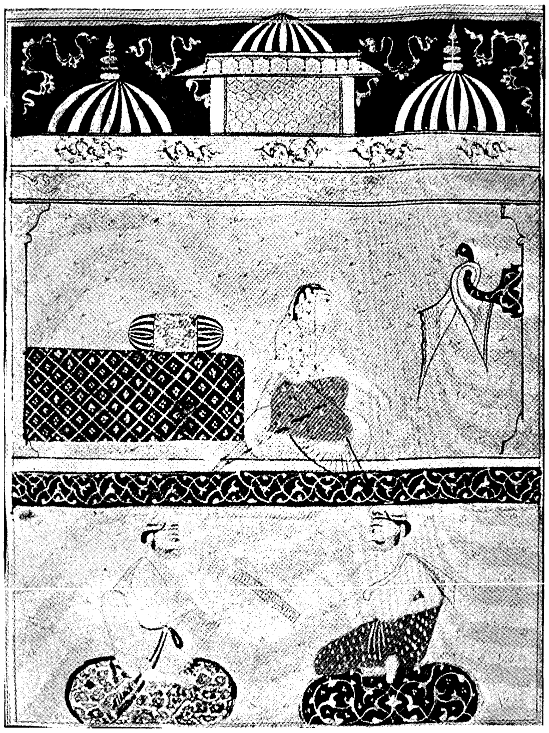
Even greater elaboration was possible with the increasing availability of paper from the late 14th century; with larger surfaces to paint on, by the middle of the 15th century artists were producing opulent manuscripts, such as the *Kalpa-sūtra* in the Devasanopaḍā library, Ahmadābād. The text is written in gold on coloured ground, the margins gorgeously illuminated with richest decorative and figural patterns, and the main paintings often occupying the entire page. Blue and gold, in addition to red, are used with increasing lavishness, testifying to the prosperity of the patron. The use of such costly materials, however, did not necessarily produce works of quality, and one is often left with the impression of cursive and hasty workmanship. With some variations—but hardly any substantial departures from the bounds that it had set for itself—the style endured throughout the 16th century and even extended into the 17th. The political subjugation of the country by the forces of Islām may have contributed to the conservatism of the style but did not result in its total elimination, as seems to have been the case in eastern India. Indeed, in the course of its long life, the western Indian school became a national style, painting at other centres in India interpreting and elaborating its forms in their own individual manner. In the province of Orissa, painting on palm leaf and in a manner entirely dependent on the western Indian style has continued up to the present day.

Transition to the Mughal and Rajasthani styles. The belief held earlier by scholars that the new Islāmic rulers of India did not patronize any painting until the rise of the Mughal dynasty in the 16th century is being abandoned in the face of the literary testimony and the discovery or recognition of illustrated manuscripts that were painted at Indian courts. Nor should this be surprising, as the Muslim kings of India had before them the example of other rulers of the Islāmic world who were great patrons of painting in spite of the injunctions of orthodox Islām against the portrayal of living beings. The taste of these Indian rulers, however, did not turn to the western Indian style but to the flourishing traditions of Islāmic painting abroad, notably neighbouring Iran. As many painters as architects had in all probability been invited from foreign countries; and illustrated manuscripts, handily transported, must have been easily available. As a result there appears to have developed what can only be called an Indo-Persian style, based essentially on the schools of Iran but affected to a greater or lesser extent by the individual tastes of the Indian rulers and by the local styles. The earliest known examples are paintings dating from the 15th century onward. The most important are the *Khamseh* ("Quintet") of Amīr Khosrow of Delhi (Freer Gallery of Art, Washington, D.C.), a *Bostān* painted in Mandu (National Museum, New Delhi), and, most interesting of all, a manuscript of the *Neʿmat-nāmeh* (India Office Library, London) painted for a sultan of Mālwa in the opening years of the 16th century. Its illustrations are derived from the Turkmen style of Shīrāz but show clear Indian features adapted from the local version of the western Indian style.

P. Chandra



Folio from a series illustrating the *Caurapañcāśikā* of Bilhaṇa, *c.* mid-16th century AD. In the Municipal Museum, Ahmadābād, Gujarāt, India.



Folio from an illustrated manuscript of *Candāyana, c.* first half of the 16th century. In the Prince of Wales Museum of Western India, Bombay.

P. Chandra

Fifteenth-century changes in western Indian style

Though the western Indian style was essentially conservative, it was not unfailingly so. It began to show signs of an inner change most notably in two manuscripts from Mandu, a *Kalpa-sūtra* and a *Kālakācāryakathā* of about 1439, and a *Kalpa-sūtra* painted at Jaunpur in 1465. These works were done in the opulent manner of the 15th century, but for the first time the quality of the line is different, and the uncompromisingly abstract expression begins to make way for a more human and emotional mood. By the opening years of the 16th century, a new and vigorous style had come into being. Although derived from the western Indian style, it is clearly independent, full of the most vital energy, deeply felt, and profoundly moving. The earliest dated example is an *Āraṇyaka Parva* of the *Mahābhārata* (1516; The Asiatic Society of Bombay), and among the finest are series illustrating the *Bhāgavata-Purāṇa* and the *Caurapañcāśikā* of Bilhaṇa, scattered in collections all over the world. A technically more refined variant of this style, preferring the pale, cool colours of Persian derivation, a fine line, and meticulous ornamentation, exists contemporaneously and is best illustrated by a manuscript of the ballad *Candāmyana* by Mullā Dāūd (*c.* first half of the 16th century; Prince of Wales Museum of Western India, Bombay). The early 16th century thus appears to have been a period of inventiveness and set the stage for the development of the Mughal and Rājput schools, which thrived from the 16th to the 19th century.

Mughal style: Akbar period (1556–1605). Although the Mughal dynasty came to power in India with the great victory won by Bābar at the Battle of Pānīpat in 1526, the Mughal style was almost exclusively the creation of Akbar. Trained in painting at an early age by a Persian master, Khwāja ʿAbd-uṣ-Ṣamad, who was employed by his father, Humāyūn, Akbar created a large atelier, which he staffed with artists recruited from all parts of India. The atelier, at least in the initial stages, was under the superintendence of Akbar's teacher and another great Persian master, Mīr Sayyid ʿAlī; but the distinctive style that evolved here owed not a little to the highly individual tastes of Akbar himself, who took an interest in the work, inspecting the atelier frequently and rewarding painters whose work was pleasing.

The technique of Mughal painting

The work of the Mughal atelier in this early formative stage was largely confined to the illustration of books on a wide variety of subjects: histories, romances, poetic works, myths, legends, and fables, of both Indian and Persian origin. The manuscripts were first written by calligraphers, with blank spaces left for the illustrations. These were executed largely by groups of painters, including a colourist, who did most of the actual painting, and specialists in portraiture and in the mixing of colours. Chief of the group was the designer, generally an artist of top quality, who formulated the composition and sketched in the rough outline. A thin wash of white, through which the initial drawing was visible, was then applied and the colours filled in. The colourist's work proceeded slowly, the colour being applied in several thin layers and frequently rubbed down with an agate burnisher, a process that resulted in the glowing, enamel-like finish. The colours used were mostly mineral but sometimes consisted of vegetable dyes; and the brushes, many of them exceedingly fine, were made from squirrel's tail or camel hair.

The earliest paintings (*c.* 1560–70) of the school of Akbar are illustrations of *Ṭūṭī-nāmeh* ("Parrot Book; Cleveland Museum of Art) and the stupendous illustrations of the *Dāstān-e Amīr Ḥamzeh* ("Stories of Amīr Ḥamzeh"; Österreichisches Museum für Angewandte Kunst, Vienna), which originally consisted of 1,400 paintings of an unusually large size (approximately 25 inches by 16 inches [65 by 40 centimetres]), of which only about 200 have survived. The *Ṭūṭī-nāmeh* shows the Mughal style in the process of formation: the hand of artists belonging to the various non-Mughal traditions is clearly recognizable, but the style also reveals an intense effort to cope with the demands of a new patron. The transition is achieved in the *Dāstān-e Amīr Ḥamzeh,* in which the uncertainties are overcome in a homogeneous style, quite unlike Persian work in its leaning toward naturalism and filled with swift, vigorous movement and bold colour. The forms are individually modelled, except for the geometrical ornament used as architectural decor; the figures are superbly interrelated in closely unified compositions, in which depth is indicated by a preference for diagonals; and much attention is paid to the expression of emotion. One of the last manifestations of this bold and vigorous early manner is the *Dārāb-nameh (c.* 1580) in the British Museum.

Immediately following were some very important historical manuscripts, including the *Tārīkh-e Khāndān-e Tīmūrīyeh* ("History of the House of Timūr," *c.* 1580–85; Khuda Baksh Library, Patna) and other works concerned with the affairs of the Tīmūrid dynasty, to which the Mughals belonged. Each of these manuscripts contains several hundred illustrations, the prolific output of the atelier made possible by the division of labour that was in effect. Historical events are recreated with remarkable inventiveness, though the explosive and almost frantic energy of the *Dāstān-e Amīr-Ḥamzeh* has begun to subside. The scale was smaller and the work began to acquire a studied richness. The narrative method employed by these Mughal paintings, like that of traditional literature, is infinitely discursive; and the painter did not hesitate to provide a fairly detailed picture of contemporary life—both of the people and of the court—and of the rich fauna and flora of India. Like Indian artists of all periods, the Mughal painter showed a remarkable empathy for animals, for through them flows the same life that flows through human beings. This sense of kinship allowed him to achieve unqualified success in the illustration of animal fables such as the *Anwār-e Suhaylī* ("Lights of Caropus"), of which several copies were painted, the earliest dated 1570 (School of Oriental and African Studies, London). It was in the illustrations to Persian translations of the Hindu epics, the *Mahābhārata* and the *Rāmāyaṇa,* that the Mughal painter revealed to the full the richness of his imagination and his unending resourcefulness. With little precedent to rely on, he was nevertheless seldom dismayed by the subject and created a whole series of convincing compositions. Because most of the painters of the atelier were Hindus, the subjects must have been close to their hearts; and, given the opportunity by a tolerant and sympathetic patron, they rose to great heights. It is no wonder, therefore, that the *Razm-nāmeh* (City Palace Museum, Jaipur), as the *Mahābhārata* is known in Persian, is one of the outstanding masterpieces of the age.

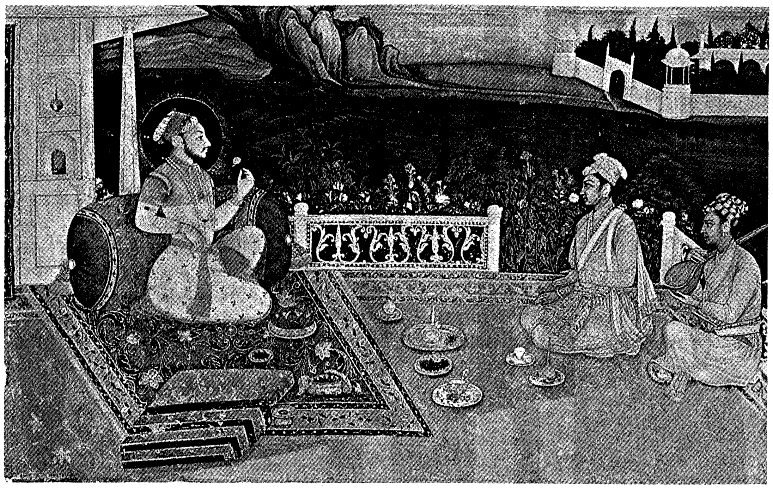
Illustrations for Persian translations of Hindu epics

In addition to large books containing numerous illustrations, which were the products of the combined efforts of many artists, the imperial atelier also cultivated a more intimate manner that specialized in the illustration of books, generally poetic works, with a smaller number of illustrations. The paintings were done by a single master artist who, working alone, had ample scope to display his virtuosity. In style the works tend to be finely detailed and exquisitely coloured. A *Dīvān* ("Anthology") of Anwarī (Fogg Art Museum, Cambridge, Massachusetts), dated 1589, is a relatively early example of this manner. The paintings are very small, none larger than five inches by $2^1/_2$ inches (12 by 6 centimetres) and most delicately executed. Very similar in size and quality are the miniatures illustrating the *Dīvān* of Ḥāfeẓ (Reza Library, Rāmpur). On a larger scale but in the same mood are the manuscripts that represent the most delicate and refined works of the reign of Akbar: the *Bahāristān* of Jāmī (1595; Bodleian Library, Oxford), a *Khamseh* of Neẓāmī (1593; British Museum, London), a *Khamseh* of Amīr Khosrow (1598; Walters Art Gallery, Baltimore and Metropolitan Museum of Art, New York), and an *Anwār-e Suhaylī* (1595–96; Bharat Kala Bhavan, Vārānasī).

Also prepared in the late 1590s were magnificent copies

P. Chandra

Painter at work, detail from a folio of the Muraqqah-e Gulshan, Mughal style, early 17th century AD. In the Staatliche Museen Preussischer Kulturbesitz, Berlin.



Nobleman seated on a terrace, Mughal style painting, mid-18th century. In the National Museum of India, New Delhi.

P. Chandra

of the *Akbar-nāmeh* ("History of Akbar"; Victoria and Albert Museum, London) and the *Kitāb-e Changīz-nāmeh* ("History of Genghis Khan"; Gulistan Library, Tehrān). These copiously illustrated volumes were produced by artists working jointly, but the quality of refinement is similar to that of the poetic manuscripts.

Of the large number of painters who worked in the imperial atelier, the most outstanding were Dasvant and Basāvan. The former played the leading part in the illustration of the *Razm-nāmeh.* Basāvan, who is preferred by some to Dasvant, painted in a very distinctive style, which delighted in the tactile and the plastic, and with an unerring grasp of psychological relationships.

Mughal style: Jahāngīr period (1605–27). The emperor Jahāngīr, even as a prince, showed a keen interest in painting and maintained an atelier of his own. His tastes, however, were not the same as those of his father, and this is reflected in the painting, which underwent a significant change. The tradition of illustrating books began to die out, though a few manuscripts, in continuation of the old style, were produced. For Jahāngīr much preferred portraiture; and this tradition, also initiated in the reign of his father, was greatly developed. Among the most elaborate works of his reign are the great court scenes, several of which have survived, showing Jahāngīr surrounded by his numerous courtiers. These are essentially large-scale exercises in portraiture, the artist taking great pains to reproduce the likeness of every figure.

ortraiture

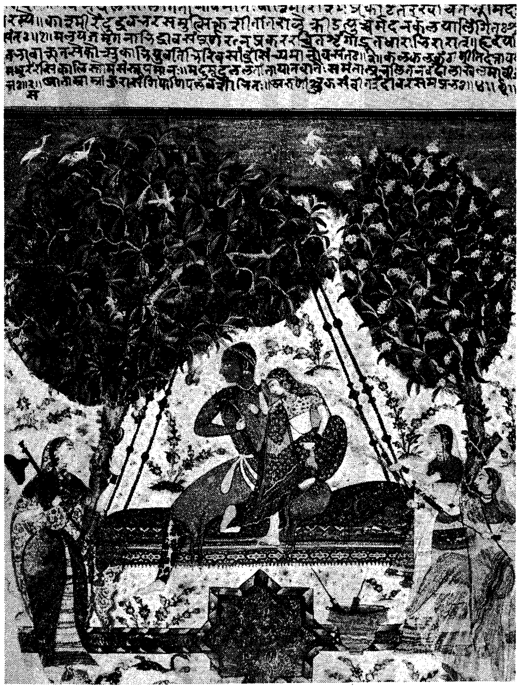
The compositions of these paintings have lost entirely the bustle and movement so evident in the works of Akbar's reign. The figures are more formally ordered, their comportment in keeping with the strict rules of etiquette enforced in the Mughal court. The colours are subdued and harmonious, the bright glowing palette of the Akbarī artist having been quickly abandoned. The brushwork is exceedingly fine. Technical virtuosity, however, is not all that was attained, for beneath the surface of the great portraits of the reign there is a deep and often spiritual understanding of the character of the person and the drama of human life.

Many of the paintings produced at the imperial atelier are preserved in the albums assembled for Jahāngīr and his son Shāh Jahān. The Muraqqah-e Gulshan is the most spectacular. (Most surviving folios from this album are in the Gulistan Library in Tehrān and the Staatliche Museen Preussischer Kulturbesitz, Berlin; a section is temporarily housed in Tübingen.) There are assembled masterpieces from Iran, curiosities from Europe, works produced in the reign of Akbar, and many of the finest paintings of Jahāngīr's master painters, all surrounded by the most magnificent borders decorated with a wide variety of floral and geometrical designs. The album gives a fairly complete idea of Jahāngīr as a patron, collector, and connoisseur of the arts, revealing a person with a wide range of taste and a curious, enquiring mind.

Jahāngīr esteemed the art of painting and honoured his painters. His favourite was Abū al-Hasan, who was designated Nādir-uz-Zamān ("Wonder of the Age"). Several pictures by the master are known, among them a perceptive study of Jahāngīr looking at a portrait of his father. Also much admired was Ustād Manṣūr, designated Nādir-ul-ʿAṣr ("Wonder of the Time"), whose studies of birds and animals are unparalleled. Bishandās was singled out by the emperor as unique in the art of portraiture. Manohar, the son of Basāvan, Govardhan, and Daulat are other important painters of this reign.

Mughal style: Shāh Jahān period (1628–58). Under Shāh Jahān, attention seems to have shifted to architecture, but painting in the tradition of Jahāngīr continued. The style, however, becomes noticeably rigid. The portraits resemble hieratic effigies, lacking the breath of life so evident in the work of Jahāngīr's time. The colouring is jewel-like in its brilliance, and the outward splendour quite dazzling. The best work is found in the *Shāhjahānnāmeh* ("History of Shāh Jahān") of the Windsor Castle Library and in several albums assembled for the emperor. Govardhan and Bichitra, who had begun their careers in the reign of Jahāngīr, were among the outstanding painters; several

P. Chandra



The musical mode *vasanta,* Deccani school painting, Bijāpur, late 16th century. In the National Museum of India, New Delhi.

works by them are quite above the general level produced in this reign.

Genre scenes

Mughal style: Aurangzeb and the later Mughals (1659–1806). From the reign of Aurangzeb (1659–1707), a few pictures have survived that essentially continue the cold style of Shāh Jahān; but the rest of the work is nondescript, consisting chiefly of an array of lifeless portraits, most of them the output of workshops other than the imperial atelier. Genre scenes, showing gatherings of ascetics and holy men, lovers in a garden or on a terrace, musical parties, carousals, and the like, which had grown in number from the reign of Shāh Jahān, became quite abundant. They sometimes show touches of genuine quality, particularly in the reign of Muḥammad Shāh (1719–48), who was passionately devoted to the arts. This brief revival, however, was momentary, and Mughal painting essentially came to an end during the reign of Shāh ʿĀlam II (1759–1806). The artists of this disintegrated court were chiefly occupied in reveries of the past, the best work, for whatever it is worth, being confined to copies of old masterpieces still in the imperial library. This great library was dispersed and destroyed during the uprising of 1857 against the British.

Company school. Rising British power, which assumed political supremacy in the 19th century, resulted in a radical change of taste brought about by the Westernization of important segments of the population. Heavily influenced by Western ideas, a style emerged that represented the adjustment of traditional artists to new fashions and demands. Rooted at Delhi and the erstwhile provincial Mughal capitals of Murshidābād, Lucknow, and Patna, it ultimately spread all over India. Most of the works produced were singularly impoverished, but occasionally there were some fine studies of natural life.

Deccani style. In mood and manner, Deccani painting, which flourished over much of the Deccan Plateau from at least the last quarter of the 16th century, is reminiscent of the contemporary Mughal school. Again, a homogeneous style evolved from a combination of foreign (Persian and Turkish) and Indian elements, but with a distinct local flavour. Of the early schools, the style patronized by the sultans of Bijāpur—notably the tolerant and art-loving Ibrāhīm ʿĀdil Shāh II of Bijāpur, famous for his love of music—is particularly distinguished. Some splendid portraits of him, more lyrical and poetic in concept than contemporary Mughal portraits, are to be found. A wonderful series depicting symbolically the musical modes (*rāgamālā*) also survives. Of illustrated manuscripts, the most important are the *Nujūm-ul-ʿulūm* ("The Stars of the Sciences," 1590; Chester Beatty Library, Dublin) and the *Tārīf-e Ḥuseyn-Shāhī* (Bharata Itihasa Samshodhaka Mandala, Pune), painted around 1565 in the neighbouring state of Ahmadnagar. The sultanate of Golconda also produced work of high quality—for example, a manuscript of the *Dīvān* of Muḥammad Qulī Quṭb Shāh in the Salar Jang Library, Hyderābād, and a series of distinguished portraits up to the end of the 17th century (dispersed in various collections). The state of Hyderābād, founded in the early 18th century and headed by a grandee of the Mughal Empire, was a great centre of painting. The work that was produced there reflects both Golconda traditions and increasing Mughal and Rajasthani influences.

The Hindu basis of Rajasthani style

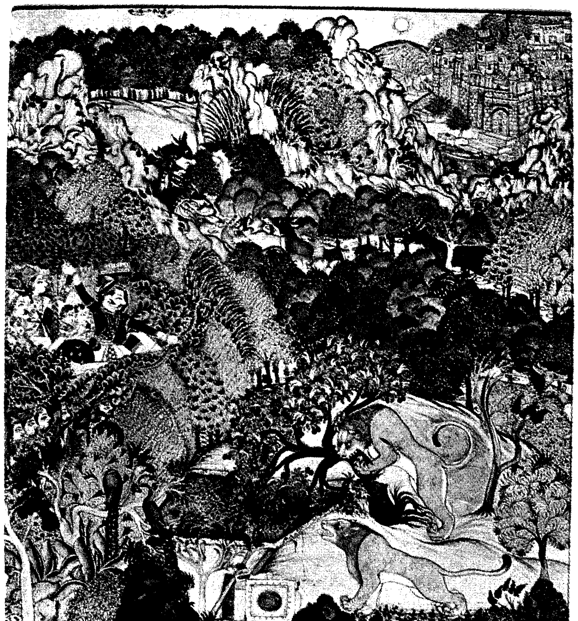
Rajasthani style. This style appears to have come into being in the 16th century, about the same time the Mughal school was evolving under the patronage of Akbar; but, rather than a sharp break from the indigenous traditions, it represented a direct and natural evolution. Throughout the early phase, almost up to the end of the 17th century, it retained its essentially hieratic and abstract character, as opposed to the naturalistic tendencies cultivated by the Mughal atelier. The subject matter of this style is essentially Hindu, devoted mainly to the illustration of myths and legends, the epics, and above all the life of Krishna; particularly favoured were depictions of his early life as the cowherd of Vraja, and the mystical love of Vraja's maidens for him, as celebrated in the *Bhāgavata-Purāṇa,* the *Gītagovinda* of Jayadeva, and the Braj Bhasa verses written by Sūrdās and other poets. The style of the painting, no less than the literature, is a product of the new religious movements, all of which stressed personal devotion to Krishna as the way to salvation. Related popular themes were pictorial representations of the musical modes (*rāgamālā*) and illustrations of poetical works such as the *Rasikapriyā* of Keśavadāsa, which dealt with the sentiment of love, analyzing its varieties and endlessly classifying the types of lovers and beloveds and their emotions. Portraits, seldom found in the early phase, became increasingly common in the 18th century—as did court scenes, scenes of sporting and hunting events, and other scenes concerned with the courtly life of the great chiefs and feudal lords of Rājasthān.

The Rajasthani style developed various distinct schools, most of them centring in the several states of Rājasthān, namely Mewār, Būndi, Kotah, Mārwār, Bīkaner, Kishangarh, and Jaipur (Amber). It also had centres outside the geographical limits of present-day Rājasthān, notably Gujarāt, Mālwa, and Bundel Khand. The study of Rajasthani painting is still in its infancy, for most of the material has been available for study only since the mid-1940s.

The Mughal and Rajasthani styles were always in contact with each other, but in general the Rajasthani schools were not essentially affected by the work produced at the Mughal court during the greater part of the 17th century. This became less true in the 18th century, when the sharp distinction between the two became progressively obscured, though each retained its distinctive features right up to the end.

Rajasthani style: Mewār. The Mewār school is among the most important. The earliest dated examples are represented by a *rāgamālā* series painted at Chawand in 1605 (Gopi Krishna Kanoria Collection, Patna). These simple paintings, filled with bright colour, are only a step removed from the pre-Rajasthani phase. The style became more elaborate in the first quarter of the 17th century when another *rāgamālā,* painted at Udaipur in 1628 (formerly in the Khajanchi Collection, Bīkaner; now dispersed in various collections), showed some superficial acquaintance with the Mughal manner. This phase, lasting until around 1660, was one of the most important for the development of painting all over Rājasthān. Ambitious and extensive illustrations of the *Bhāgavata,* the *Rāmāyaṇa,* the poems of Sūrdās, and the *Gītagovinda* were completed, all full of strength and vitality. The name of Sāhabadī is intimately connected with this phase; another well-known painter is Manohar. The intensity and richness associated with their

P. Chandra



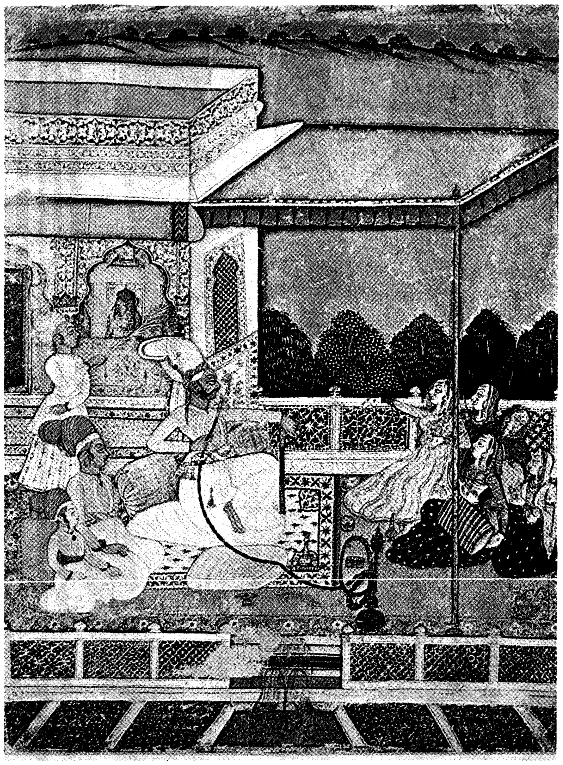
Lion hunt, Rajasthani style painting, Kotah, late 18th century. In a private collection.

atelier began to fade toward the close of the 17th century, and a wave of Mughal influence began to affect the school in the opening years of the 18th century. Portraits, court scenes, and events in the everyday world of the ruling classes are increasingly found. Although the emotional fervour of the 17th century was never again attained, this work is often of considerable charm. The 19th century continued to create work in the tradition of the 18th, one of the most important centres being Nāthdwāra (Rājasthān), the seat of the Vallabha sect. Large numbers of pictures, produced here for the pilgrim trade, were spread over all parts of Rājasthān, northern India, Gujarāt, and the Deccan.

arliest xamples f Būndi hool

Rajasthani style: Būndi and Kotah. A school as important as that of Mewār developed at Būndi and later at Kotah, which was formed by a partition of the parent state and ruled by a junior branch of the Būndi family. The earliest examples are represented by a *rāgamālā* series of extraordinarily rich quality, probably dating from the end of the 16th century. From the very beginning the Būndi style seemed to have found Mughal painting an inspiring source, but its workmanship was as distinctively Rajasthani as the work of Mewār. The artists of this school always displayed a pronounced preference for vivid movement, which is unique in all of Rājasthān. Toward the second half of the 17th century, work at Būndi came unmistakably under the influence of Mewār; many miniatures, including several series illustrating the *Rasikapriyā,* indicate that this was a period of prolific activity. The sister state of Kotah also appears to have become an important centre of painting at this time, developing a great fondness for hunting and sport scenes, all filled with great vigour and surging strength. This kind of work continued well into the 19th century, and if the workmanship is not always of the highest quality, the style maintained its integrity against the rapidly increasing Western influence right up to the end.

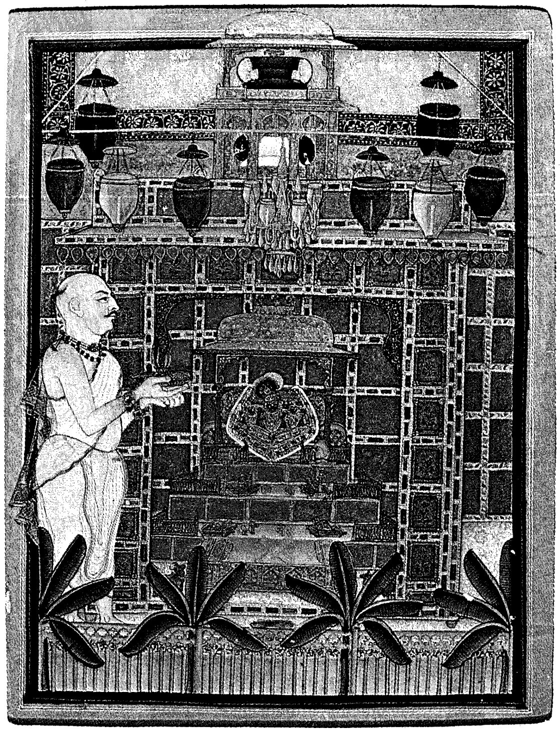
Rajasthani style: Mālwa. It has been suggested but not definitely determined that the school itself does not belong to Mālwa but to some other area, probably Bundelkhand. In contrast to the Būndi school, miniatures generally thought to have been painted in Mālwa are quite archaistic, with mannerisms inherited from the 16th century retained until the end of the 17th. The earliest work is an illustrated version of the *Rasikapriyā* (1634), followed by a series illustrating a Sanskrit poem called the *Amaru Śataka* (1652). There are also illustrations of the musical modes (*rāgamālā*), the *Bhāgavata-Purāṇa,* and other Hindu devotional and literary works. The compositions of all of these pictures is uncompromisingly flat, the space divided into registers and panels, each filled with a patch of colour and occupied by figures that convey the action. This conservative style disappeared after the close of the 17th century. The course of Mālwa painting in the 18th century and later is not known.



A Rājput king listening to music, Rajasthani style painting, Mārwār, mid-18th century. In a private collection.

P. Chandra

P. Chandra



A priest at worship, Rajasthani style painting, Nāthdwāra, mid-19th century. In a private collection.

Rajasthani style: Mārwār. A *rāgamālā* series dated 1623 reveals that painting in this state shared features common to other schools of Rājasthān. Miniatures of the second half of the 17th century are distinguished by some splendid portraits that owe much to the Mughal school. A very large amount of work was done in the 19th century, all of which is highly stylized but strong in colour and often of great charm.

Rajasthani style: Bīkaner. Of all Rajasthani schools, the Bīkaner style, from its very inception in the mid-17th century, shows the greatest indebtedness to the Mughal style. This is due to the presence in the Bīkaner atelier of artists who had previously worked in the Mughal manner at Delhi. They and their descendants continued to paint in a style that could only be classed as a provincial Mughal manner had it not been for the quick absorption of influences from the Rajasthani environment and a sympathy for the religious and literary themes favoured by the royal Hindu patrons. Delicacy of line and colour are consistent characteristics of Bīkaner painting even when, toward the end of the 18th century, it assumed stylistic features associated with the more orthodox Rajasthani schools.

Characteristics of the Bīkaner style

Rajasthani style: Kishangarh. The Kishangarh school, which came into being toward the mid-18th century, was also indebted to the contemporary Mughal style but combined a rich and refined technique with deeply moving religious fervour. Its inspiring patron in the formative phase was Sāvant Singh, more of a devotee and a poet than a king. The style established by him, characterized by pronounced mystical leanings and a distinctive facial type, continued to the middle of the 19th century, though at a clearly lower level of achievement.

Rajasthani style: Jaipur (Amber). The rulers of the state

Month of summer, Rajasthani-style painting, Bīkaner, early 18th century. In a private collection.
P. Chandra

were closely allied to the Mughal dynasty, but paintings of the late 16th and early 17th centuries possessed all of the elements of the Rajasthani style. Little is known about the school until the opening years of the 18th century, when stiff, formal examples appear in the reign of Savāī Jai Singh. The finest works, dating from the reign of Pratāp Singh, are sumptuous in effect and include some splendid portraits and some large paintings of the sports of Krishna. Although the entire 19th century was extremely productive, the work was rather undistinguished and increasingly affected by Western influences. Of the Rajasthani styles of this period, the Jaipur school was the most popular, examples having been found all over northern India.

Pahari style. Closely allied to the Rajasthani schools both in subject matter and technique is the Pahari style, so-named because of its prevalence in the erstwhile hill states of the Himalayas, stretching roughly from Jammu to Gaṛhwāl. It can be divided into two main schools, the Basohlī and the Kāngra, but it must be understood that these schools were not confined to the centres after which they are named but extended all over the area. Unlike Rājasthān, the area covered by the Pahari style is small, and the probability of artists travelling from one area to another in search of livelihood was much greater. Thus, attempts to distinguish regional schools are fraught with controversy, and it has been suggested that a classification based upon ateliers and families is likely to be more tenable than those presently current among scholars. Because the Basohlī and the Kāngra schools show considerable divergences, scholars have postulated the existence of a transitional phase, named the pre-Kāngra style.

Pahari style: Basohlī school. The origins of this remarkable style are not yet understood, but it is clear that the style was flourishing toward the close of the 17th century. The earliest dated paintings are illustrations to the *Rasamañjarī* of Bhānudatta (a Sanskrit work on poetics), executed for a ruler of Basohlī (1690; Boston Museum of Fine Arts). Bold colour, vigorous drawing, and primitive intensity of feeling are outstanding qualities in these paintings, quite surpassing the work of the plains. In addition to other Hindu works such as the *Gītagovinda* and the *Bhāgavata-Purāṇa,* a fairly large number of idealized portraits have also been discovered.

Pahari style: Kāngra school. The Basohlī style began to fade by the mid-18th century, being gradually replaced by the Kāngra style, named after the state of Kāngra but, like the Basohlī style, of much wider prevalence. A curvilinear line, easy flowing rhythms, calmer colours, and a mood of sweet lyricism easily distinguish the work from that of the Basohlī style. The reasons for this change are to be sought in strong influences from the plains, notably the Mughal styles of Delhi and Lucknow. These influences account for the more refined technique; but whatever was borrowed was transmuted and given a fresh and tender aspect. Among the greatest works are large series illustrating the *Bhāgavata-Purāṇa* (National Museum, New Delhi), the *Gītagovinda,* and the *Satsaī* of Bihārī (both in the collection of the maharaja of Tehrī-Garhwāl), all painted in 1775–80. The corpus of work produced is very large and, although it seldom fails to please, works of high achievement are rare. The school flourished from about 1770 to almost the end of the 19th century, but the finest work was produced largely around 1775–1820.

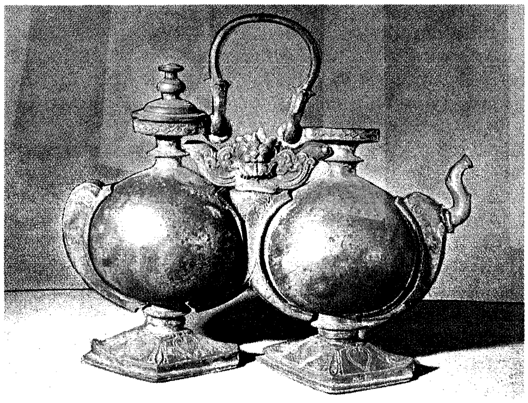
Modern period. Toward the late 19th century traditional Indian painting was rapidly dying out, being replaced by feeble works in a variety of idioms, all strongly influenced by the West. A reaction set in during the early 20th century, symbolized by what is called the Bengal school. The glories of Indian art were rediscovered, and the school consciously tried to produce what it considered a truly Indian art inspired by the creations of the past. Its leading artist was Rabindranath Tagore and its theoretician was E.B. Havell, the principal of the Calcutta School of Art. Nostalgic in mood, the work was mainly sentimental though often of considerable charm. The Bengal school did a great deal to reshape contemporary taste and to make Indian artists aware of their own heritage. Amrita Sher-Gil, who was inspired by the Postimpressionists, made Indian painters aware of new directions. Mid-20th-century Indian painting is very much a part of the international scene, the artists painting in a variety of idioms, often attempting to come to terms with their heritage and with the emergence of India as a modern culture.

The Beng
school

Decorative arts. Fragmentary ivory furniture (*c.* 1st century AD) excavated at Begrām is one of the few indications of the existence in ancient India of a secular art concerned with the production of luxurious and richly decorated objects meant for daily use. Objects that can be clearly designated as works of decorative art become much more extensive for the later periods, during which Islāmic traditions were having a profound effect on Indian artistic traditions. The reign of the Mughal emperors, in particular, produced works of the most elaborate and exquisite craftsmanship; the decorative tradition is clearly preserved in architectural ornament, though surviving decorative objects themselves, particularly before the 17th century, are far fewer than might be expected. Economic conditions, including competition with machine-made goods imported from English factories, and a change in taste from increasing European influence had disastrous consequences for traditional craftsmanship, especially in the late 19th and 20th centuries.

Pre-Islāmic period. Of the very few objects surviving from the pre-Islāmic period, the most important are fragments of ivory caskets, chairs, and footstools found at Begrām, in eastern Afghanistan, but obviously of Indian origin and strongly reminiscent of the school of Mathurā in the 1st century AD. The work is profusely decorated with carved panels and confirms the wide reputation for superb ivories that India had in ancient times. Nothing as spectacular has come down from the succeeding periods, but stray examples such as the so-called Charlemagne chessman (*c.* 8th century; Cabinet des Medailles, Paris) and two magnificent throne legs, of Orissan workmanship, carved in the shape of griffins with elephant heads (13th century; Freer Gallery of Art, Washington D.C., and Philadelphia Museum of Art), indicate that ivory craftsmanship was always vital. Ancient traditions, relatively unaffected by Islāmic influence, continued in southern India up to modern times. An exquisitely carved box from Vijayanagar (16th century; Prince of Wales Museum of Western India, Bombay) is representative; many other exquisite objects of this period and later are among the treasures of South Indian temples.

Work in
ivory



Ritual vessel from Kollur, Mysore, India, copper, *c.* 14th century. In the Prince of Wales Museum of Western India, Bombay.

P. Chandra

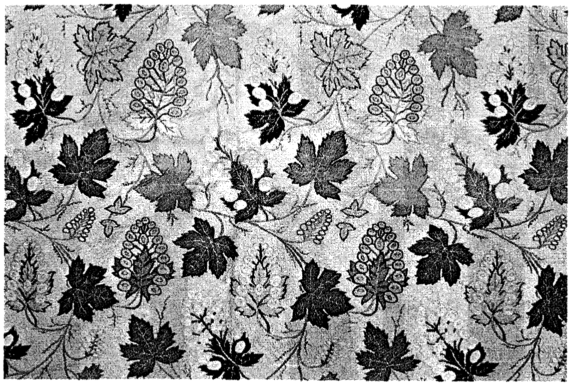
There is even less evidence of what the decorative work in metal was like. The practice of re-using the metal by melting unserviceable items may account for the paucity of objects, for there is little doubt that the craft was always flourishing. A hoard found at Kolhāpur, consisting of plates, various kinds of vessels, lamps and *objets d'art,* including a superb bronze elephant with riders, constitutes the most important surviving group of metal objects and is datable to about the 2nd century AD. Some fine examples of ritual utensils, notably elaborate incense burners, of the 8th–9th century have been excavated at Nālandā; and a large number of 14th-century ceremonial vessels of complex design and excellent workmanship, and apparently belonging to the local temple, were discovered at Kollur, in Mysore state.

Gold played an extremely important role in the manufacture of jewelry, but once again the finds are hardly commensurate with tradition. Small amounts of gold jewelry have been excavated at Mohenjo-daro and Harappā (3rd millennium BC); and, in the historical period, a very important group, of delicate workmanship, has been excavated at Taxila (*c.* 2nd century AD).

From earliest times, India has been famous for the variety and magnificence of its textiles. In this case, however, the Indian climate has been particularly destructive; virtually nothing has survived the heat and moisture. Besides the testimony of literature and the evidence of figural sculpture, only a few fragments of printed textiles are preserved—at Fusṭāṭ in Egypt, where they had been exported. These date approximately to the 14th century.

Islāmic period. Traditions of craftsmanship established during the Islāmic period came to full flower during the reign of the Mughal dynasty. Surviving works of decorative art are more abundant, though once again there are hardly as many examples as might be expected, particularly from the 16th and 17th centuries. According to literary testimony and the few available examples, the finest objects were undoubtedly made in the imperial workshops set up in large number at the capital and in the great cities of the empire, where they were nourished by local traditions. Well-organized, these shops specialized in particular items, such as textiles, carpets, jewelry, ornamental arms and armour, metalware, and jade. Textile manufacture must have been enormous, considering the demands of court and social etiquette and ritual. Contributing to the popularity of tapestries, curtains, draperies, canopies, and carpets in contemporary architecture were the nomadic tenting traditions of the Mughal rulers.

P. Chandra



Detail of a Kashmir shawl, wool embroidery, 19th century. In the Prince of Wales Museum of Western India, Bombay.

Popularity of textiles

The variety of techniques employed in the manufacture of textiles was infinite, ranging from printed and painted patterns to the exquisite embroidery decoration of woolen shawls and the costly figured brocading of silk. An important contribution to carpet weaving was the landscape carpet that reproduced pictorial themes inspired by miniature painting. Much of the surviving textile work dates from the 18th century or later, though the 16th and 17th centuries produced works of the most outstanding quality.

In response to growing European trade, a considerable amount of furniture (chairs, cabinets, chests of drawers, and the like) was produced, mostly wood inlaid with ivory. Many of these pieces have been preserved in the kinder European climate. Although the furniture made for export gives some idea of the craft in India, it must be emphasized that only the ornamental and figural work was Indian, while the form was European. Also in a hybrid Indō-European style were the Christian objects produced by a local school of ivory carvers at Goa.

Metal objects of sumptuous quality were also made, a unique example of which is a splendid, elaborately chiselled 16th-century cup in the Prince of Wales Museum of Western India in Bombay. This tradition was continued in the 17th and particularly the 18th century, when vessels made of a variety of metals and adorned with engraved, chiselled, inlaid, and enamelled designs were very popular. Arms and armour, in particular, were decorated with the skill of a jeweler. Particularly striking are the carved hilts, often done in animal shapes.

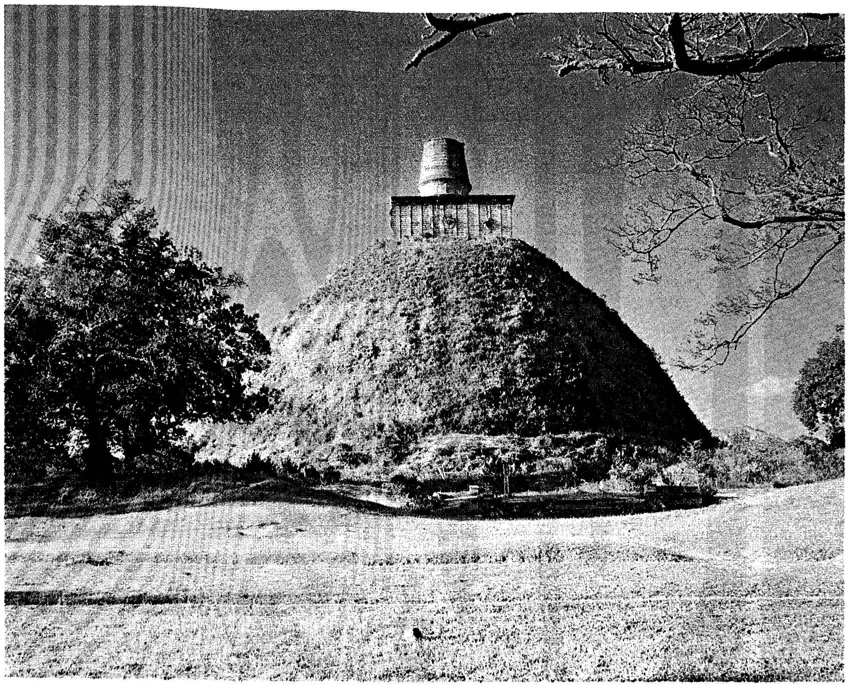
Jade or jadeite was much fancied by the rich and was used together with crystal to make precious vessels as well as sword and dagger hilts. A rather large number of 18th- and 19th-century objects have survived, but they are often of nondescript quality. The greatest period for jade carving seems to have been the 17th century; a few outstanding examples associated with the emperors Jahāngīr and Shāh Jahān are of singular delicacy and perfection. The practice of inlaying jade, and also stone, with precious or semiprecious stones became more popular with the reign of Shāh Jahān and increasingly characteristic of Indian jade craftsmanship from that time on.

Architectural decoration provides a clear idea of the range of ornamental patterns used by the Mughal artist. They consisted mainly of arabesques (intricate interlaced patterns made up of flower, foliage, fruit, and sometimes animal and figural outlines) and infinitely varied geometric patterns—motifs shared with the rest of the Muslim world—together with floral scrolls and other designs adapted from Indian traditions. As a whole, the Mughal decorative style tends to endow ornamental patterns with a distinctive plasticity not seen in the more truly two-dimensional Iranian and Arab work. From the 17th century, a type of floral spray became the most favoured motif and was found on almost every decorated object. The motif, symmetrical but relatively naturalistic at the beginning, became progressively stiff and stylized, but never lost its importance in the ornamental vocabulary.

VISUAL ARTS OF SRI LANKA (CEYLON)

Distinction between art of Sri Lanka and that of India

The art of Sri Lanka is closely allied to that of India but presents several distinctive features that make a separate treatment convenient. There is, first, the considerable transformation of Indian influences, resulting in an idiom of great power and individuality. Sri Lanka also often served as a geographical pocket in which styles that had disappeared in India were preserved, which accounts for the anachronistic features of some phases of Sinhalese art. It also appears that although predominant influences were from neighbouring southern India, this was not exclusively

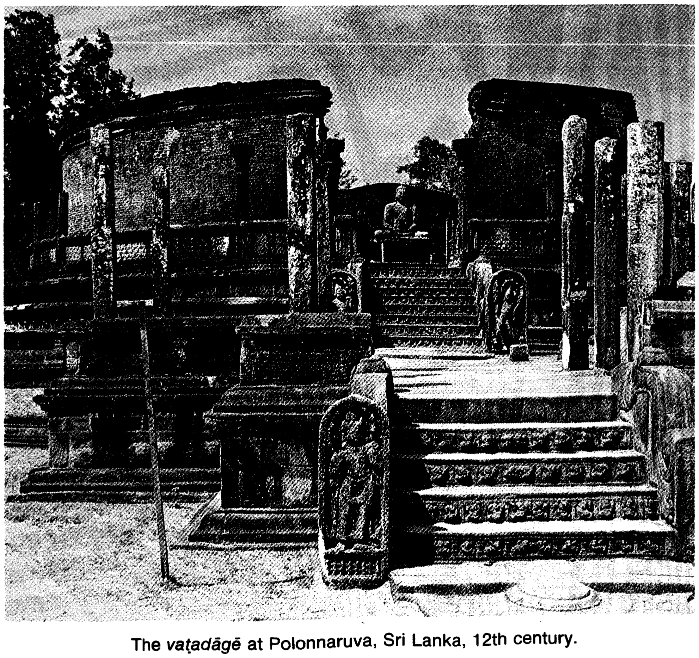


Abhayagiri *dagāba* in the Jetavana at Anurādhapura, Sri Lanka, *c.* 4th century.
ZEFA

so, and styles flourishing in western and northern India, too, contributed to the formulation of Sinhalese art. The difficulties in the study of the art are considerable: the long, unbroken Buddhist traditions and the piety of the rulers and the people have led to the successive renovation of monuments; and in the absence of firmly dated monuments, one of the few relatively reliable tools of study is comparison with Indian art, an approach full of pitfalls and shortcomings.

Architecture. The most impressive monuments are the great *stūpa*s, some of gigantic size and considerable antiquity but often reconstructed in the course of the centuries. They generally have a triple circular base, and as in early Indian *stūpa*s, a hemispherical dome with a miniature railing on top, and a multiple parasol that tends to solidify into a conical structure in the course of time. The material is brick, sometimes covered with plaster and white paint. An important feature are the platforms (*vāhalakaḍa*s) at the cardinal points, often adorned with sculpture. There are many *stūpa*s at the ancient capital of Anurādhapura, at Polonnaruva, and at other sites; of these the Jetavana at Anurādhapura is the largest, though now largely ruined.

ZEFA



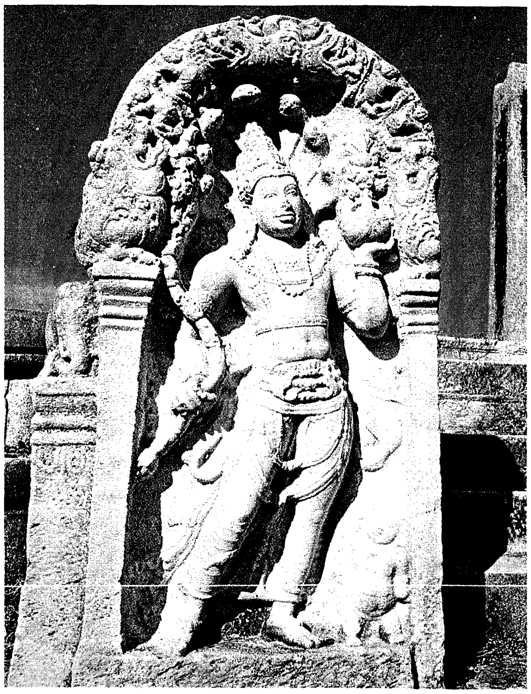
The *vaṭadāgē* at Polonnaruva, Sri Lanka, 12th century.

Small *stūpa*s were often placed in a circular building with a domical metal and timber roof supported by concentric rows of stone pillars. This type of building, known in ancient India as the *caityagṛha,* was very popular in Sri Lanka, though it had disappeared at a fairly early period in the country of its origin. A famous example is the *vaṭadāgē* at Polonnaruva, a structure of great elegance. The dome itself, being of perishable material, has not survived. The *geḍigē,* or large rectangular hall with a corbelled brick vault, housing a Buddha image, is first found in Sri Lanka from the 8th century AD; the most impressive example is the Laṅkātilaka at Polonnaruva built by Parākramabāhu I in the 12th century.

Literature testifies to the existence of elaborate royal and priestly residences of wood, which have largely disappeared. The Lohapāsāda at Anurādhapura, traditionally ascribed to Duṭṭhagāmaṇi (101–77), was originally a nine-story building, now destroyed except for the large number of stone pillars that supported the upper floors. Sigiriya, a 6th-century fortress city with extensive remains, is another notable example of secular architecture.

Sculpture. The earliest sculpture, perhaps, is from the platforms, or *vāhalakaḍa*s, of the Kanṭaka Cetiya, at Mihintalē, and reveals an archaistic style indebted to 1st-century-BC Indian sculpture of Sānchi and Amarāvatī regions. A certain simplicity and restraint characteristic of most Sinhalese work is present even at this early stage. The first Buddha images show a pronounced relationship to examples from Andhradeśa of the 2nd–3rd century AD but often possess considerable vigour, revealing the contribution of the local sculptor. Several fine images are known, one of the best of which is at Ruanveli, Anurādhapura, now very badly restored.

Dated monuments are absent from the 5th to the 12th centuries, but an approximate idea of stylistic development can be obtained by a comparative study of Indian examples. An outstanding image, rather hideously repaired in recent years, is a great seated Buddha in Anurādhapura, the smooth and abstract modelling of which recalls the school of Sārnāth of the 5th–6th century. At Isurumuni, near Anurādhapura, are some marvellous reliefs carved on rocks. One of these depicts elephants at play, and another, a seated man with the head of a horse carved in the background. These fine sculptures recall the South Indian style of the 7th century. A radiant amorous couple carved



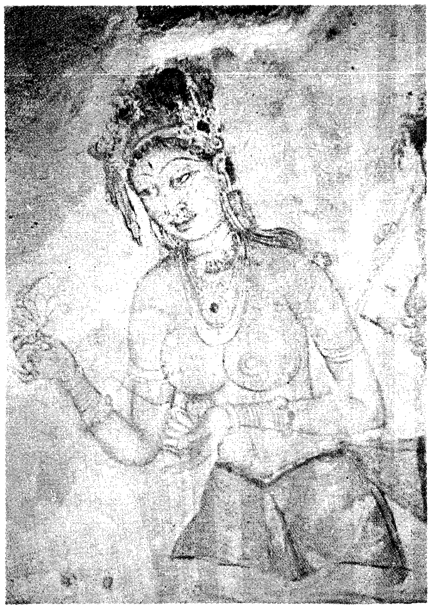
Nāga stele from Anurādhapura, Sri Lanka, 10th century.
George Holton—Photo Researchers

in relief on a stone slab, also at Isurumuni, represents Sinhalese sculpture at its most joyous.

culptured
:aircases

Of about the same period or a little later, are exquisitely sculptured staircases decorated with moonstones, and stelae, or commemorative pillars, carved with a guardian *nāga*, a spirit with combined superhuman and serpent qualities. The latter are among the finest examples of Sinhalese sculpture, the full and weighty modelling relieved by the skillful movement of clearly chiselled ornament. The Ratnapāsāda at Anurādhapura and the eastern staircase of the *vaṭadāgē* at Polonnaruva possess particularly superb specimens. Moonstones—decorated with bands of floral motifs, geese, and a row of animals consisting of a lion, bull, elephant, and horse—placed at the bottom of the staircase, testify to the great taste and elegance that mark Sinhalese decorative carving. At Anurādhapura and related sites a certain freedom characterizes the work, while the slightly later examples at Polonnaruva are stiffer but technically brilliant.

P. Chandra



Painted figure of an *apsaras* from Sigiriya, Sri Lanka, 6th century.

A colossal Buddha, 42 feet (13 metres) high, at Avukana, testifies to the increasing hardness of the Sinhalese style, which, even so, never ceases to be moving. Large images of the Buddha at the Gal Vihāra and a figure supposedly representing Parākramabāhu at Potgal Vihāra, both at Polonnaruva, are of the 12th century. They are figures of great majesty and surpass contemporary work in southern India. After the 13th century, Sinhalese sculpture began to decline, though work of some decorative value was produced up to the 19th century.

Painting. The rock at Sigiriya is adorned with a series of exquisitely painted *apsaras*es (nymphs) showering flowers, their torsos emerging from clouds. The paintings are dated to the 6th century AD; in their plastic resiliency they are reminiscent of contemporary work in India. The next important group of wall paintings come from Tivaṁkapatimā-ghara at Polonnaruva. Although dated to the 12th or 13th century, the figures continue to be modelled, relatively unaffected by the linear distortions of the western Indian style that was flourishing in India. Eighteenth-century paintings, with their flat figures arranged in horizontal rows, reflect contemporary styles of southern India. (P.Ch.)

BIBLIOGRAPHY

Literature: (Sanskrit, Pāli, and Prākrit): The old literature of Southeast Asia has been more extensively described than any of the modern literatures. Though antiquated, MORIZ WINTERNITZ, *Geschichte der indischen litteratur,* 3 vol. (1908–22; Eng. trans., *A History of Indian Literature,* 2nd ed. 1959–67), is still very useful. So is ARTHUR B. KEITH, *A History of Sanskrit Literature* (1928, reprinted 1956), which, however, does not include the Sanskrit theatre. For the old literature of the Veda, ARTHUR A. MACDONNELL, *A History of Sanskrit Literature* (1900), remains very helpful as a survey. A full inventory of the literature is given by S.N. DAS GUPTA and S.K. DE, *A History of Sanskrit Literature, Classical Period,* 2nd ed. (1962). The epochal material is best treated by EDWARD W. HOPKINS, *The Great Epic of India* (1901). On the Sanskrit play specifically, the study by the French scholar SYLVAIN LEVI, *Le Théâtre indien* (1890), remains an important contribution. The best introduction to the narrative literature is to be found in CHARLES H. TAWNEY (trans.), *The Ocean of Story,* ed. by NORMAN M. PENZER, 10 vol. (1923–28, reprinted 1968). GEORGE L. HART, *The Relation Between Tamil and Classical Sanskrit Literature* (1976), is an argument that the two literatures stem from a common source.

Modern Indian literatures: The literatures in the modern Indian languages are bibliographically underrepresented in English. For some of them the best sources are to be found written in those languages themselves. The following bibliography must confine itself to works written in more accessible languages. (*Hindi*): A good though incomplete inventory of the older literature in Hindi and Hindustani is found in JOSEPH GARCIN DE TASSY, *Histoire de la littérature hindouie et hindoustanie,* 2nd ed., 3 vol. (1870–71, reprinted 1968). A useful guide up to its date is EDWIN GREAVES, *A Sketch of Hindi Literature* (1918). A survey of the more modern literature in Hindi is R.A. DWIVELDI, *A Critical Survey of Hindi Literature* (1966). (*Assamese*): Among the few works available is B.K. BARUA, *Assamesè Literature* (1941) and *A History of Assamese Literature* (1965). (*Bengali*): The best extensive introduction to the literature in Bengali is that of SUKUMAR SEN, *History of Bengali Literature* (1960). Another useful source is J.C. GHOSH, *Bengali Literature* (1948). (*Marathi*): Many works on Marathi literature are written in Marathi itself. A good guide is G.C. BHATE, *History of Modern Marathi Literature, 1800–1938* (1939). (*Gujarati*): Little is written in English on Gujarati literature. To be recommended is K.M. MUNSHI, *Gujarāt and Its Literature from Early Times to 1852,* 3rd ed. (1967). A useful older study is K.M. JHAVERI, *Milestones in Gujarati Literature* (1914). (*Punjabi*): A good survey of Punjabi is given by MOHAN SINGH, *An Introduction to Panjabi Literature* (1951). For information about literature in the smaller Indo-Aryan languages, see the symposium of the All-India Writers' Conference, *Writers in Free India* (1950). (*Muslim contributions*): The best short introduction to the cultural and intellectual life of the Moslems of the subcontinent is AZIZ AHMAD, *An Intellectual History of Islam in India* (1969). A comprehensive survey of the literature in Urdu produced in South India, especially in the medieval kingdoms of Bijāpur and Golconda and in the later state Hyderābād is the Urdu work *Dekan men Urdu* by NASIRRUDDIN HASHMI (1963). A study of the convention of classical Urdu poetry in the light of the writings of three Urdu poets of 18th-century Delhi is RALPH RUSSELL and KHURSHIDUL ISLAM, *Three Mughal Poets: Mir, Sauda, Mir Ḥasan* (1968). The latest and most comprehensive survey of literature produced in Persian in various countries, including India and Pakistan, is JAN RYPKA, *History of Iranian Literature* (1968). A useful introduction is MUHAMMAD SADIQ, *A History of Urdu Literature* (1964). (*Tamil*): KAMIL V. ZVELEBIL, *Tamil Literature* (1974), is an introduction to and critical study of the literature. Much useful information is found in XAVIER S. THANI NAYAGAM, *A Reference Guide to Tamil Studies* (1966). Fuller treatment is given in C. and H. JESUDASAN, *A History of Tamil Literature* (1961). Especially recommended are also T.P. MEENAKSHISUNDARAM, *A History of Tamil Literature* (1965); and J.M. SOMASUNDARAM PILLAI, *A History of Tamil Literature with Texts and Translations from the Earliest Times to 600 A.D.* (1968). A good account of Telugu literature is P. CHENCHAYYA and R.M. BHUJANGA RAO BHADUR, *A History of Telugu Literature* (1928). More recent is P.T. RAJU, *Telugu Literature* (1944); and GIDUGU VENKAJA SITAPATI, *History of Telugu Literature* (1968). Very little has been written on Kannada literature. Mention can be made of the older EDWARD P. RICE, *A History of Kanarese Literature*, 2nd ed. rev. and enl. (1921); and of H. THIPPERUDA SWAMY, *The Vīraśaiva Saints* (1968). The literature in Malayalam is sketched in K.M. GEORGE, *A Survey of Malayalam Literature* (1968). Further mention can be made of K.K. NAIR, *A History of Malayalam Literature* (1971); and of P.K. PARAMESWARAN NAIR, *History of Malayalam Literature* (Eng. trans. 1967). Finally, among the very few books in English on the literature of Ceylon (Sri Lanka), prominent mention can be made of CHARLES EDMUND GODAKUMBURA, *The Literature of Ceylon* (1963).

Music: ARNOLD A. BAKE, "The Music of India," in *The New Oxford History of Music*, vol. 1 (1957), a general chapter on Indian music dealing with the philosophical background, Vedic chant, the ancient musical system, the modern classical system, and musical instruments; ELISE B. BARNETT, "Special Bibliography: The Art Music of India," *Ethnomusicology*, 14:278–312 (1970), a listing of books and articles on Indian music, published since 1959, which includes some publications on folk and religious music, dance, and drama; SUDHIBHUSHAN BHATTACHARYA, *Ethnomusicology and India* (1968), a synchronic approach attempting to relate folk, tribal, religious, and classical music in terms of the cultural background—includes outline notations in Indian *sargam;* ALAIN DANIELOU, *The Rāga-s of Northern Indian Music* (1968), an individualistic interpretation of modern North Indian *ragas,* based partly on ancient theory; B.C. DEVA, *Psychoacoustics of Music and Speech* (1967), a collection of articles on various aspects of music and speech, with emphasis on acoustics and the scientific study of Indian music; ARTHUR H. FOX-STRANGWAYS, *The Music of Hindostan* (1914, reprinted 1965), a work of wide scope including discussion of Vedic chant, folk music, and modern Indian classical music (includes notations in Western staff and analogies with Western music); O.C. GANGOLY, *Rāgas and Rāginīs*, 2 vol. (1934–35, reprinted 1948), a historical study that traces the systems of classifying *ragas* in Indian musical treatises, including a discussion of the time-theory of *ragas* as well as their iconography; NAZIR A. JAIRAZBHOY, *The Rāgas of North Indian Music* (1971), a technical work dealing with the structure and evolution of North Indian *ragas* (includes a 45 r.p.m. record of *raga* demonstrations performed on the *sitar* by VILAYAT KHAN, and extensive notations in Western staff and Indian *sargam*); BABURAO JOSHI and A. LOBO, *Introducing Indian Music* (n.d.), a series of four long-playing records, including musical examples and commentary, illustrating the main features of North Indian classical music, with accompanying booklet; WALTER KAUFMANN, *The Rāgas of North India* (1968), description and notations, in Western staff, of about 230 *ragas* of modern North Indian music; ALLEN KEESE, *The Sitar Book* (1968), an elementary guide to the *sitar,* with some description of playing techniques, musical exercises, and *gats* in ten *ragas;* HERBERT A. POPLEY, *The Music of India,* 3rd ed. (1970), a general work including discussion of historical background, scale, *raga, tala,* musical form, and instruments, with notations in Western staff and Indian *sargam;* HAROLD S. POWERS, "Indian Music and the English Language: A Review Essay," *Ethnomusicology,* 9:1–12 (1965), a survey of the most important literature on Indian music written in the English language; and "An Historical and Comparative Approach to the Classification of Rāgas (with an Appendix on Ancient Indian Tunings)," in *Selected Reports* of the Instiute of Ethnomusicology, UCLA, 1:1–78 (1970), a scholarly monograph that attempts to show the relationship between North and South Indian *ragas* that bear the same name but now differ in scale—includes material on ancient Indian music and gives many musical examples in Western staff; SWAMI PRAJNANANDA, *A History of Indian Music,* vol. 1 (1963), a technical work dealing with the origins and the music of the ancient period; P. SAMBAMOORTHY, *South Indian Music,* 5 vol. (1958–69), a comprehensive work covering many aspects of South Indian musical theory, both synchronically and diachronically; RAVI SHANKAR, *My Music, My Life* (1969), a general book combining autobiographical and biographical material with a discussion of Indian music history, theory, and instruments, including a manual for the *sitar.* BONNIE C. WADE, *Music in India: The Classical Traditions* (1979), a discussion of performance, theory, and basic instruments of both the North and South; M.R. GUATAM, *The Musical Heritage of India* (1981), a history with emphasis on classical traditions of both the North and South; WIM VAN DER MEER, *Hindustan: Music in the Twentieth Century* (1980), an introduction with emphasis on vocal music; DANIEL M. NEUMAN, *The Life of Music in North India* (1980), an analysis of the place in society of the musician.

Dance and theatre: The main source book of Indian classical dance and theatre is the *Nātya-śāstra,* ascribed to BHARATA MUNI, trans. by MANMOHAN GHOSH, 2 vol. (1950–61), which deals with ceremonies, gesture language, architecture, production styles, makeup, and costumes. A.K. COOMARASWAMY, *The Dance of Shiva,* rev. ed. (1957), brings alive the philosophy and aesthetics of Hindu dance. For general understanding of classical dance forms and techniques, see RINA SINGHA and REGINALD MASSEY, *Indian Dances* (1967), which includes semiclassical styles, modern ballet, and biographical notes on dancers and gurus. FAUBION BOWERS, *The Dance in India* (1953), is a fascinating description of the four major classical dances. K. BHARATHA IYER, *Kathakali* (1955), is perhaps the best work on this dance-drama, with detailed descriptions of characters, historical background, and interpretation of dramatic symbols, with photographs and line drawings. KAPILA VATSYAYAN, *Classical Indian Dance in Literature and the Arts* (1968), surveys dance as found in temple sculpture, the plastic arts, and literature—a scholarly work with photos. For tribal dances and ceremonies of Central India, see VERRIER ELWIN, *The Muria and Their Ghotul* (1947; abridged ed., *The Kingdom of the Young,* 1968). For dramatic forms, production techniques, and classical rules the best source book is also the *Nātya-śāstra,* ascribed to BHARATA MUNI, *op. cit.* ARTHUR B. KEITH, *Sanskrit Drama in Its Origin, Development, Theory and Practice* (1924, reprinted 1959), remains a standard scholarly critical work for comparative analysis of Sanskrit plays. P. LAL, *Great Sanskrit Plays* (1964) are actable transcreations in crisp modern English. A delightfully-written survey is FAUBION BOWERS, *Theatre in the East* (1956), which puts Indian dance and theatre in the larger perspective of South Asia with vivid and perceptive descriptions. For a general survey of classical, folk, and modern drama with sidelights on opera and ballet, illustrated by photographs and sketches, see BALWANT GARGI, *Theatre in India* (1962). HEMENDRA NATH DAS GUPTA, *The Indian Stage,* 4 vol. (1934–44), deals mainly with the growth of Bengali theatres, actors, and productions in exhaustive detail with reproductions of period notebooks and papers. BALWANT GARGI, *Folk Theatre in India* (1966), is an eyewitness account of religious, secular, and masked dramas in villages, with over 100 photographs and line sketches. See also J.C. MATHUR, *Drama in Rural India* (1964). BERYL DE ZOETE, *Dance and Magic Drama in Ceylon* (1957), is a vivid account of rituals and magical masked dances in a diary form of day-to-day performances. E.R. SARACHCHANDRA, *The Folk Drama of Ceylon,* 2nd ed. (1966), is a scholarly work on the development and background of Sinhalese folk dramas and cults of exorcism. For puppets and masks, see two small pamphlets: J. TILAKASIR, "Puppetry in Ceylon" and SIRI GUNASINGHE, "Masks of Ceylon" (both published by the Department of Cultural Affairs, Ceylon, 1962). For the history of Kolam and description of various characters see O. PERTOLD, "The Ceremonial Dances of the Sinhalese," *Archiv Orientální,* vol. 2 (1930). For Devil Dances and their interpretation in the light of witchcraft rituals, see DANDRIS DE SILVA GUNARATNA, "Demonology and Witchcraft in Ceylon," *J. Ceylon Brch. R. Asiat. Soc.,* vol. 4, no. 13 (1861); PAUL WIRZ, *Exorzismus und Heilkunde auf Ceylon* (1941); RACHEL VAN M. BAUMER and JAMES R. BRANDON (eds.), *Sanskrit Drama in Performance* (1981), essays on Ancient Indian theatrical performance.

Visual arts (General works): VINCENT A. SMITH, *A History of Fine Art in India and Ceylon,* 3rd ed. rev. and enl. (1962); A.K. COOMARASWAMY, *History of Indian and Indonesian Art* (1927, reprinted 1965); and BENJAMIN ROWLAND, *Art and Architecture of India,* 3rd ed. rev. (1967), are general introductions with good bibliographies. A.K. COOMARASWAMY, *Figures of Speech or Figures of Thought* (1946), and *Tranformation of Nature in Art* (1934, reprinted 1956), contain important essays which discuss Indian aesthetic theories from the traditional point of view. A clear classification of Hindu images with particular reference to south India has been made in T.A. GOPINATHA RAO, *Elements of Hindu Iconography,* 2 vol. in 4 (1914–16, reprinted 1968); and J.N. BANERJEA, *The Development of Hindu Iconography,* 2nd ed. rev. and enl. (1956), is an analytical introduction to the subject. N.K. BHATTASALI, *Iconography of Buddhist and Brahmanical Sculptures* *in the Dacca Museum* (1929); BENOYTOSH BHATTACHARYYA, *The Indian Buddhist Iconography,* 2nd ed. (1958); and ALFRED FOUCHER, *Étude sur l'iconographie bouddhique de l'Inde,* 2 vol. (1900–05), are important studies of Buddhist iconography. A.K. COOMARASWAMY, *Yaksas,* 2 vol. (1928–31), is a masterly study of early iconography. HEINRICH ZIMMER, *Myths and Symbols in Indian Art and Civilization* (1946, reprinted 1963), discusses some important symbols of Indian art. Good collections of photographs are in JAMES BURGESS, *The Ancient Monuments, Temples and Sculptures of India,* 2 vol. (1897–1911); A.K. COOMARASWAMY, *Viśvakarma* (1912); STELLA KRAMRISCH, *The Art of India* (1954); and HEINRICH ZIMMER, *The Art of Indian Asia,* 2 vol. (1955).

Architecture: JAMES FERGUSSON, *History of Indian and Eastern Architecture,* rev. ed., 2 vol. (1910, reprinted 1967); PERCY BROWN, *Indian Architecture,* 2 vol., 5th ed. (1965); and S.K. SARASWATI, "Architecture," in R.C. MASUMDAR (ed.), *History and Culture of the Indian People,* vol. 3 and 5 (1954–57), are standard works that survey the entire history of Indian architecture. JAMES FERGUSSON and JAMES BURGESS, *The Cave Temples of India* (1880, reprinted 1969), is a comprehensive account of rock-cut architecture. STELLA KRAMRISCH, *The Hindu Temple,* 2 vol. (1946), is concerned with principles and symbolism, and KRISHNA DEVA, *Temples of North India* (1969), presents a synoptic view of the various north Indian styles. G. JOUVEAU-DUBREUIL, *Archéologie du sud de l'Inde,* 2 vol. (1914), analyses the south Indian style. K.R. SRINVASAN, *Cave Temples of the Pallavas* (1964); and ARTHUR H. LONGHURST, *Pallava Architecture,* 3 vol. (1924–30), are important studies of early south Indian architecture. SIR JOHN MARSHALL, "The Monuments of Muslim India," in *The Cambridge History of India,* vol. 3, pp. 568–640 (1928, reprinted 1965); and PERCY BROWN, "Monuments of the Mughal Period," *ibid.,* vol. 4, pp. 523–576 (1937, reprinted 1963), are important essays on Islamic architecture in India. See also ELIZABETH S. MERKLINGER, *Indian Islamic Architecture: The Deccan 1347–1686* (1981).

Sculpture: S.K. SARASVATI, *A Survey of Indian Sculpture* (1957); and STELLA KRAMRISCH, *Indian Sculpture* (1933), discuss the broad historical and stylistic trends. LUDWIG BACHHOFER, *Early Indian Sculpture* (1929), is a fine stylistic analysis of sculptures from the third century B.C. to the third century A.D. The classic work on Gandhāra art is ALFRED FOUCHER, *L'Art gréco-bouddhique du Gandhâra,* 2 vol. (1905–41), though several of its conclusions are no longer tenable. SIR JOHN MARSHALL, *Taxila,* 3 vol. (1951), discusses works recovered from that site; and HAROLD INGHOLT, *Gandharan Art in Pakistan* (1957, reprinted 1971), provides excellent photographic documentation. R.D. BANERJI, *The Age of the Imperial Guptas* (1933), has a general discussion of the Gupta style; D.R. SAHNI, *Catalogue of the Museum of Archaeology at Sārnāth* (1914), contains information of interest on the school of Sārnāth. ELIKY ZANNAS, *Khajurāho* (1960); K.C. PANIGRAHI, *Archaeological Remains at Bhubaneswar* (1961); and R.D. BANERJI, *Eastern Indian School of Medieval Sculpture* (1933), are monographs on schools of north Indian medieval sculpture. A.H. LONGHURST, *op. cit.;* and K.A. NILAKANTA SASTRI, *The Cōlas,* 2nd ed. rev. (1955), contain information on the south Indian medieval styles. C. SIVARAMAMURTI, *South Indian Bronzes* (1963); P.R. SRINIVASAN, *Bronzes of South India* (1963); DOUGLAS E. BARRETT, *Early Cōla Bronzes* (1965), are important studies of south Indian bronzes; FREDERICK M. ASHER, *The Art of Eastern India: 300–800* (1980), a study of the pre-Pala period.

Painting: In considering the published literature it is important to remember that the study of Indian painting, confined to a limited number of scholars, is of comparatively recent growth, and is therefore full of controversies and uncertainties which keep shifting with the discovery of fresh materials. The standard work on Ajanta is GHULAM YAZDANI, *Ajanta,* 4 vol. (1930–55); and on the western Indian style, MOTI CHANDRA, *Jain Miniature Paintings from Western India* (1949). Much interesting information on the period of transition to the Rajasthani and Mughal styles has been brought together in KARL J. KHANDALAVALA and MOTI CHANDRA, *New Documents of Indian Painting* (1969). The Mughal school has been ably presented in PERCY BROWN, *Indian Painting under the Mughals, A.D. 1550 to A.D. 1750* (1924); and STUART C. WELCH, *The Art of Mughal India* (1964). DOUGLAS E. BARRETT, *Painting of the Deccan, XVI–XVII Century* (1958), is a brief introduction to the subject. The main publication on the Company style is MILDRED and WILLIAM G. ARCHER, *Indian Painting for the British, 1770–1880* (1955). The classic work on Pahari and Rajasthani painting is A.K. COOMARASWAMY's pioneering *Rajput Painting,* 2 vol. (1916). Fresh discoveries which have considerably changed the understanding of its history are summarized in KARL KHANDALAVALA, MOTI CHANDRA, and PRAMOD CHANDRA, *Miniature Painting: A Catalogue of the Exhibition of the Sri Motichand Khajanchi Collection* (1960). MOTI CHANDRA, *Mewar Painting in the Seventeenth Century* (1957); WILLIAM G. ARCHER, *Indian Painting in Bundi and Kotah* (1959); PRAMOD CHANDRA, *Bundi Painting* (1959); ERIC DICKINSON and KARL KHANDALAVALA, *Kishangarh Painting* (1959); WILLIAM G. ARCHER, *Central Indian Painting* (1958); and ANAND KRISHNA, *Malwa Painting* (1963), are informative summaries of the growing knowledge of the various schools of Rājasthān. The standard work, illustrated copiously, on the Pahari style is KARL KHANDALAVALA, *Pahāṛī Miniature Painting* (1958), and different in its account from WILLIAM G. ARCHER, *Indian Painting in the Punjab Hills* (1952). Books which cover most of the schools and are profusely illustrated include N.C. MEHTA, *Studies in Indian Painting* (1926); WILLIAM G. ARCHER, *Indian Miniatures* (1960); ROBERT SKELTON, *Indian Miniatures from the XVth to the XIXth Centuries* (1961); DOUGLAS E. BARRETT and BASIL GRAY, *Painting of India* (1963); and STUART C. WELCH and MILO C. BEACH, *Gods, Thrones and Peacocks: Northern Indian Painting from Two Traditions, Fifteenth to Nineteenth Centuries* (1965). (*Ceylon*): Standard works on Sinhalese art are SENERAT PARANAVITANA, *The Stūpa in Ceylon* (1946), *Art and Architecture of Ceylon: Polonnaruva Period* (1954), and *Ceylon: Paintings from Temple, Shrine and Rock* (1957). MOTI CHANDRA, *Studies in Early Indian Painting* (1975), covers the 5th through 16th centuries; CALAMBUR SIVARAMAMURTI, *The Art of India* (1977), covering all types of art with 1175 illustrations; STUART C. WELCH, *Room for Wonder: Indian Painting During the British Period, 1760–1880* (1978), an exhibition catolog with detailed comments on 125 illustrations.

Decorative arts: Scholarly literature on the decorative arts in India is scanty and mainly in learned journals. The *Journal of Indian Art and Industries* (1886–1916) is the most important and contains numerous pioneering studies. SIR GEORGE WATT, *Indian Art at Delhi* (1903); and SIR GEORGE BIRDWOOD, *The Industrial Arts of India,* 2 vol. (1880), are for the most part descriptive texts emphasizing the technical aspects of the decorative arts as they had survived up to the closing years of the 19th century. JOHN IRWIN, "Textiles and the Minor Arts," in LEIGH ASHTON (ed.), *The Art of India and Pakistan,* pp. 201–237 (1950), is a brief historical survey of the subject. MOTI CHANDRA, "Ancient Indian Ivories," *Bulletin of the Prince of Wales Museum of Western India,* no. 6, pp. 4–63 (1957–59), is a monograph on the history of Indian ivory carving. THOMAS H. HENDLEY, *Asian Carpets* (1905), treats Indian examples in the important collections of the Maharaja of Jaipur. GEORGE P. BAKER, *Calico Painting and Printing in the East Indies in the XVIIth and XVIIIth Centuries* (1921), is the most important work on the subject; WILBRAHAM EGERTON, *An Illustrated Handbook of Indian Arms* (1880), catalogs and describes the wide range and achievement of the armourer's craft.

(C.S./J.A.B.v.B./E.C.D./C.M.N./A.K.R./N.A.J./B.Ga./P.Ch.)

Mainland Southeast Asia

The mainland of Southeast Asia comprises the Indochinese Peninsula, which is divided into the following countries: Kampuchea (Cambodia), Laos, Malaysia, Myanmar (Burma), Singapore, Thailand, and Vietnam. The peninsula consists principally of a series of north–south-trending ranges and valleys associated with some of Asia's largest rivers, including the Mekong and the Irrawaddy. The climate is predominantly tropical, and the natural vegetation is forest.

The area is highly diversified culturally. Rice and corn (maize) are the principal dietary staples. Southeast Asia is an important source of rubber, rice, coconut oil, sugar, teak, hemp, tea, coffee, tobacco, and pepper and of tin, petroleum, bauxite, and other minerals.

Before the 16th century, Southeast Asia had contacts with India and China, and in several places, such as Kampuchea and Thailand, high civilizations had developed. The region was colonized by Western powers from the 16th to the 20th century. In the latter half of the 20th century, all the colonial states achieved independence. Only Thailand had avoided colonization, but it has been heavily influenced by the West. After World War II, Southeast Asia became enmeshed in power struggles between East and West and between the Soviet Union and China. (Ed.)

For a discussion of the artistic expressions of the peoples of Southeast Asia, see SOUTHEAST ASIAN ARTS.

For coverage of related topics in the *Macropædia* and *Micropædia,* see the *Propædia,* sections 937, 969, and 976, and the *Index*.

This article is divided into the following sections:

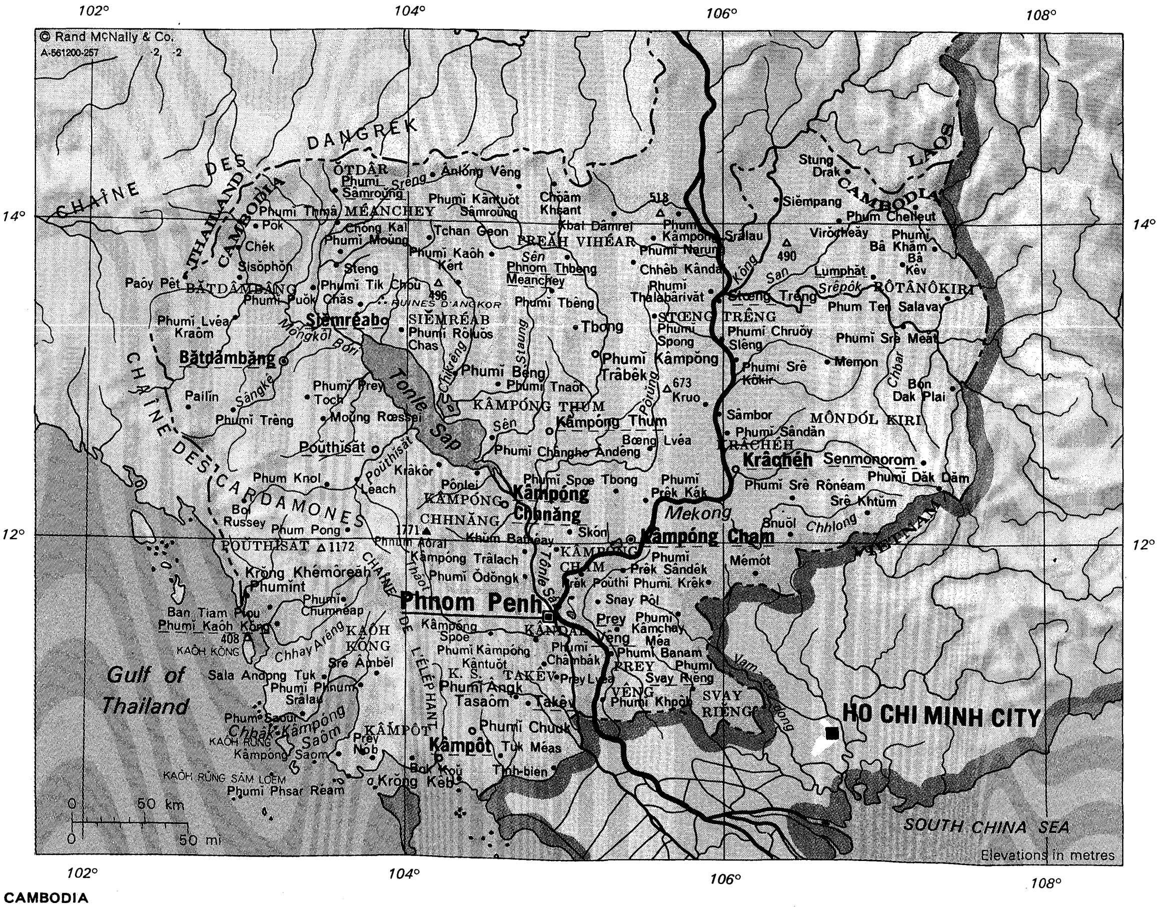
KAMPUCHEA

Kampuchea is a country located in Southeast Asia in the southwest part of the Indochinese Peninsula. Together with Laos and Vietnam, Kampuchea is part of what was formerly known as Indochina. Covering a land area of 69,898 square miles (181,035 square kilometres), it is bordered on the west and northwest by Thailand, on the northeast by Laos, on the east and southeast by Vietnam, and on the southwest by the Gulf of Thailand (Gulf of Siam).

The names Cambodia and Kampuchea are transliterations of the country's traditional name in the Khmer language. Cambodia was adopted as the Western conventional spelling for the French transliteration Cambodge. Under the Khmer Republic, established in 1970, official use of the transliteration Cambodia, with its colonial connotation, was abolished. The transliteration Kampuchea, however, was allowed. In this article Cambodia is used as a general historical term and Kampuchea is used for the country after 1975.

Jpheavals
f the
970s

Throughout the 1970s, Kampuchea was devastated by the enlarged Second Indochinese War, civil war, and social upheavals that damaged, destroyed, or redirected much of its economy and killed hundreds of thousands of its people. In April 1975, the government of the Khmer Republic collapsed under attacks by Communist guerrilla forces that had been fighting the regime, at first with Vietnamese help, since 1970. From 1975 to 1978, Kampuchea was known as Democratic Kampuchea and was governed by a pro-Chinese faction of the Kampuchean Communist Party. This regime's years in power were characterized by a dramatic attempt to increase Kampuchea's agricultural production by collectivizing its villages and regimenting its people. Under the regime, millions of people were moved from urban centres into rural areas, currency was abolished, and the postal system did not operate. Violent clashes along the border with Vietnam began in 1976 and developed into warfare in 1978. In January 1979 a pro-Vietnamese Communist movement, aided by Vietnamese armed forces, captured the national capital of Phnom Penh and began to administer the surrounding area as the People's Republic of Kampuchea. Remnants of the army of Democratic Kampuchea and thousands of Cambodians loyal to that government continued resistance in the northwestern region of the country.



CAMBODIA

Physical and human geography

THE LAND

Relief. Kampuchea's maximum extent is about 280 miles (450 kilometres) from north to south and 360 miles (580 kilometres) from east to west. The central region is a low-lying alluvial plain surrounding the Tonle Sap (Great Lake) and the beginnings of the Mekong River Delta. Extending outward from this region are transitional plains, thinly forested and with prevailing elevations no higher than several hundred feet above sea level. On the north, along the border with Thailand, the Cambodian plain abuts a sandstone escarpment that marks the southern limit of a mountain range, the Chuŏr Phnum Dângrêk. A southward-facing cliff, stretching for more than 200 miles in a west to east direction, rises abruptly from the plain to heights ranging from 600 to 1,800 feet (180 to 550 metres), forming a natural frontier boundary. East of the Mekong the transitional plains gradually merge with the eastern highlands, a region of densely forested mountains and high plateaus that extend northward and eastward into Laos and Vietnam. In southwestern Kampuchea two distinct upland blocks, comprising the Chuŏr Phnum Krâvanh (Cardamom Mountains) and the Chuŏr Phnum Dâmrei (Elephant Mountains), form another highland region that covers much of the land area between the Tonle Sap and the Gulf of Thailand. In this remote and largely uninhabited area is found Phnum Aôral (5,949 feet; 1,813 metres), Kampuchea's highest peak. The southern coastal region adjoining the Gulf of Thailand is a narrow lowland strip, heavily wooded and moderately populated. It is effectively isolated from the central plain by the southwestern highlands.

Drainage. The two most dominant topographical features of Kampuchea are the Mekong River and Tonle Sap. Rising in Tibet and emptying into the South China Sea, the Mekong enters Kampuchea from Laos and flows broadly southward to the border with Vietnam, covering a distance inside Kampuchea of approximately 315 miles. Its annual flooding during the rainy season deposits a rich alluvial sediment that accounts for the fertility of the central plain.

Tonle Sap (the Great Lake)

The Tonle Sap, joined to the Mekong by a river called the Tônlé Sab, serves as a reservoir for the Mekong. During the rainy season, from mid-May to early October, the Mekong's enormous volume of water backs up the connecting river for a distance of 65 miles and flows into the Tonle Sap lake, expanding its surface area from the dry season minimum of 1,200 square miles to a rainy season maximum of more than 3,000 square miles and increasing its maximum depth from seven feet to 35 feet. As the water level of the Mekong falls during the dry season, the process is reversed. Water drains from the Tonle Sap lake back into the Mekong, reversing its directional flow. As a result of this annual phenomenon, the Tonle Sap lake is one of the richest sources in the world of freshwater fish.

Climate. Kampuchea has a climate governed by rain-bearing winds (monsoons) and characterized by two major seasons. From mid-May to early October, strong prevailing winds blow out of the southwest, bringing heavy rains and high humidity. From early November to mid-March, winds are from the northeast and are light in velocity.

MAP INDEX

Political subdivisions

Bătdâmbâng	13·00n	103·00e
Kâmpóng Cham	12·00n	105·30e
Kâmpóng Chhnăng	12·00n	104·30e
Kâmpóng Spoe	11·30n	104·30e
Kâmpóng Thum	13·00n	105·00e
Kâmpôt	10·45n	104·15e
Kândal	11·30n	104·50e
Kaôh Kŏng	11·30n	103·30e
Krâchéh	12·30n	106·00e
Môndól Kiri	12·30n	107·00e
Ŏtdâr Méanchey	14·05n	103·35e
Poŭthĭsăt	12·30n	104·00e
Preăh Vihéar	14·00n	104·50e
Prey Vêng	11·30n	105·30e
Rôtânôkiri	13·00n	107·00e
Siĕmréab	14·00n	104·00e
Stœng Trêng	14·00n	106·15e
Svay Riĕng	11·15n	105·50e
Takêv	11·00n	104·50e

Cities and towns

Ânlóng Vêng	14·14n	104·05e
Bâ kêv	13·42n	107·12e
Ban Tiam Piou	11·34n	103·08e
Battambang (Bătdâmbâng)	13·06n	103·12e
Bœng Lvéa	12·36n	105·34e
Boi Russey	12·14n	103·01e
Bok Koŭ	10·37n	104·03e
Bon Dak Plai	12·58n	107·28e
Chék	13·48n	102·58e
Chhêb Kândal	13·45n	105·24e
Chŏâm Khsant	14·13n	104·56e
Chŏng Kal	13·57n	103·35e
Kâmpóng Cham, see Kompong Cham		
Kâmpóng Chhnăng	12·15n	104·40e
Kâmpóng Saôm, see Kompong Som		
Kâmpóng Spoe	11·27n	104·32e
Kâmpóng Thum, see Kompong Thom		
Kâmpóng Trâlach	11·54n	104·47e
Kâmpôt	10·37n	104·11e
Kbal Dâmrei	14·03n	105·20e
Khŭm Bathéay	11·59n	104·57e
Kompong Cham (Kâmpóng Cham)	12·00n	105·27e
Kompong Som (Kâmpóng Saôm, Sihanoukville)	10·38n	103·30e
Kompong Thom (Kâmpóng Thum)	12·42n	104·52e
Krâkôr	12·32n	104·12e
Kratie (Krâchéh)	12·29n	106·01e
Krŏng Kêb	10·30n	104·19e
Krŏng Khémôreăh Phumĭnt	11·37n	102·59e
Kruo	12·52n	105·55e
Léach	12·21n	103·46e
Lumphăt	13·30n	106·59e
Memon	13·08n	106·40e
Mémót	11·49n	106·11e
Moŭng Rœssei	12·46n	103·27e
Pailĭn	12·51n	102·36e
Paôy Pêt, see Poipet		
Phnom Penh	11·33n	104·55e
Phnom Thbeng Meanchey	13·49n	104·58e
Phum Chelleut	14·06n	107·14e
Phumĭ Ângk Tasaôm	11·01n	104·41e
Phumĭ Bâ Khăm	13·51n	107·21e
Phumĭ Banam	11·19n	105·18e
Phumĭ Béng	13·05n	104·18e
Phumĭ Châmbák	11·14n	104·49e
Phumĭ Chângho Ândêng	12·39n	104·35e
Phumĭ Chruŏy Slêng	13·14n	105·57e
Phumĭ Chumnéap	11·36n	103·35e
Phumĭ Chuuk	10·50n	104·28e
Phumĭ Dăk Dăm	12·20n	107·21e
Phumĭ Kâmchay Méa	11·35n	105·40e
Phumĭ Kâmpóng Kântuŏt	11·26n	104·49e
Phumĭ Kâmpóng Srâlau	14·05n	105·46e
Phumĭ Kâmpŏng Trâbêk	13·06n	105·14e
Phumĭ Kântuŏt Sâmroŭng	14·12n	104·37e
Phumĭ Kaôh Kért	13·47n	104·32e
Phumĭ Kaôh Kŏng	11·28n	103·13e
Phumĭ Khpôb	11·02n	105·12e
Phumĭ Krêk	11·46n	105·56e
Phumĭ Lvéa Kraôm	13·21n	102·54e
Phumĭ Moŭng	13·46n	103·33e
Phumĭ Narung	13·53n	105·34e
Phumĭ Ŏdŏngk	11·48n	104·45e
Phumĭ Phnum Srâlau	11·03n	103·42e
Phumĭ Phsar Réam	10·30n	103·37e
Phumĭ Prêk Kák	12·15n	105·32e
Phumĭ Prêk Sândêk	11·51n	105·22e
Phumĭ Prey Toch	12·54n	103·23e
Phumĭ Puŏk Chăs	13·26n	103·44e
Phumĭ Rôluŏs Chas	13·19n	104·00e
Phumĭ Sâmroûng	14·11n	103·31e
Phumĭ Sândăn	12·42n	106·01e
Phumĭ Spoe Tbong	12·20n	105·19e
Phumĭ Spong	13·27n	105·34e
Phumĭ Srê Kôkir	13·08n	106·04e
Phumĭ Srê Meăt	13·18n	107·10e
Phumĭ Srê Rônéam	12·16n	106·25e
Phumĭ Svay Riĕng	11·05n	105·48e
Phumĭ Tbêng	13·35n	104·55e
Phumĭ Thalabârivăt	13·33n	105·57e
Phumĭ Thmâ Pôk	13·57n	103·04e
Phumĭ Tik Choŭ	13·36n	103·24e
Phumĭ Tnaôt	12·56n	104·34e
Phumĭ Trêng	12·50n	102·54e
Phum Knol	12·21n	103·33e
Phum Pong	12·06n	103·40e
Phum Saour	10·55n	103·22e
Phum Ten Salavay	13·32n	107·26e
Poipet (Paôy Pêt)	13·39n	102·33e
Pônlei	12·26n	104·27e
Poŭthĭsăt	12·35n	103·55e
Prĕk Poŭthĭ	11·51n	105·07e
Prey Lvéa	11·10n	104·57e
Prey Nôb	10·38n	103·47e
Prey Vêng	11·29n	105·19e
Sala Andong Tuk	11·08n	103·31e
Sâmbor	12·46n	105·58e
Senmonorom	12·27n	107·12e
Siĕmpang	14·07n	106·23e
Siem Reap (Siĕmréab)	13·22n	103·51e
Sihanoukville, see Kompong Som		
Sisŏphŏn	13·35n	102·59e
Skón	12·04n	105·04e
Snay Pôl	11·40n	105·13e
Snuŏl	12·04n	106·26e
Srê Âmbĕl	11·07n	103·46e
Srê Khtŭm	12·11n	106·54e
Steng	13·43n	103·34e
Stœng Tréng, see Stung Treng		
Stung Drak	14·19n	106·47e
Stung Treng (Stœng Tréng)	13·31n	105·58e
Takêv	10·59n	104·47e
Tbong	13·21n	105·07e
Tchan Geon	13·54n	104·09e
Tinh-bien	10·36n	104·57e
Tuk Méas	10·41n	104·33e
Virŏcheăy	14·00n	106·45e

Physical features and points of interest

Angkor, Ruines d', *ruins*	13·25n	103·52e
Cardamones, Chaîne des, *mountains*	12·00n	103·15e
Chbar, *river*	13·22n	107·04e
Chhâk Kâmpóng Saôm, *bay*	10·50n	103·32e
Chhay Arêng, *river*	11·31n	103·25e
Chhlong, *river*	12·15n	106·00e
Chikrêng, *river*	12·51n	104·14e
Dangrek, Chaîne des, *mountains*	14·25n	103·30e
Éléphant, Chaîne de l', *mountains*	11·00n	104·05e
Kaôh Rŭng Sâm Loem, *island*	10·35n	103·28e
Kŏng, *river*	13·35n	106·00e
Kŏng, Kaôh, *island*	11·20n	103·00e
Mekong (Mékôngk), *river*	10·33n	105·24e
Môngkôl Bori, *river*	13·19n	103·24e
Phnom Aural (Phnum Aôral), *mountain*	12·02n	104·10e
Pôrŭng, *river*	12·55n	105·35e
Poŭthĭsăt, *river*	12·41n	104·09e
Rung, Kaoh, *island*	10·35n	103·17e
San, *river*	13·32n	105·57e
Sângkê, *river*	13·13n	103·41e
Sên, *river*	12·32n	104·28e
Srêng, *river*	13·21n	103·27e
Srêpôk, *river*	13·33n	106·16e
Staung, *river*	12·50n	104·19e
Thailand, Gulf of	10·45n	102·00e
Thâot, *river*	11·29n	104·57e
Tônlé Sab, *river*	12·10n	104·40e
Tonle Sap, *lake*	12·50n	104·00e

Cloudiness is variable, precipitation is infrequent, and humidity is low. Between these seasons the weather is transitional. Maximum temperatures are high throughout the year, ranging from the mid-90s (° F, about 35° C) in April, the warmest month, to the low 80s in January, the coldest month. Daily minimums are usually 15° to 20° F lower than maximums. Annual rainfall varies considerably throughout the country, from more than 200 inches (5,000 millimetres) on the seaward slopes of the southwestern highlands to 55 inches in the central lowland region. Between 70 and 80 percent of the annual rainfall occurs during the months of the southwest monsoon.

Plant and animal life. About three-fourths of Kampuchea's land area is forested. The central lowland region is covered with rice paddies, fields of dry crops (such as corn [maize] or tobacco), tracts of tall grass and reeds, and thinly wooded areas. Savanna (grassy parkland) is the prevailing vegetation of the transitional plains, with grass growing to a height of five feet. In the eastern highlands the high plateaus are covered with deciduous forest and grassland. Broadleaf evergreen forests grow in the mountainous areas to the north, with trees 100 feet high emerging from thick undergrowths of vines, rattans, palms, bamboos, and assorted woody and herbaceous ground plants. In the southwestern highlands open forests of pines are found at the higher elevations, while the rain-drenched seaward slopes are blanketed with virgin rain forest growing to heights of 150 feet or more. Vegetation along the coastal strip ranges from evergreen forests to nearly impenetrable mangrove forests.

Fruit, growing wild or cultivated, includes breadfruit, jackfruit, durian, mango, mangosteen, papaya, rambutan, and bananas.

Animal life includes elephant, wild oxen, tiger, panther and other leopard, bear, and innumerable small game. Among the more common birds are heron, crane, grouse, pheasant, peacock, pelican, cormorant, egret, and wild duck. Four varieties of snakes are especially dangerous: the cobra, the king cobra, the banded krait, and Russell's viper.

Settlement patterns. Until the upheavals of the 1970s, Kampuchea was a country of villages. Only a small fraction of the total population lived in urban areas of 10,000 or more inhabitants; the major part of the urban population was concentrated in Phnom Penh, which is situated at the confluence of the Mekong, Basăk (Bassac), and Tônlé Sab rivers.

illage life

Until the mid-1970s the vast majority of Cambodia's people inhabited the central lowland region, where the rural village was second only to the family as the basic social unit. The typical Kampuchean village, in those days, was to a large extent autonomous and self-sufficient, made up of ethnically homogeneous people and having a population of fewer than 300 persons. The village (*phum*) was part of a hamlet or community (*khum*) with which it shared one or more Buddhist temples (*wat*), an elementary school, and several small shops. Cambodian villages usually developed in a linear pattern along waterways and roads; more isolated villages were rectangular or, more rarely, circular in the arrangement of their houses. Houses were also dispersed through the countryside on largely self-contained paddy farms. Houses in Kampuchea were generally built on high wooden pilings and had thatched roofs, walls of palm matting, and floors of woven bamboo strips resting on bamboo joists. More prosperous houses, while still on pilings, were built of wood, and had tile or metal roofs.

Until they were forced off their holdings in 1975 and made to work as ordinary peasants, there were a few large landowners in Kampuchea. Before collectivization the typical villager owned and worked enough land to provide for his family, with a small surplus to be converted into cash for additional goods or the payment of taxes. Landholdings tended to be small in the crowded south-central regions of the country, where the least fertile soils occur. During the 1960s, the government of Prince Norodom Sihanouk was successful in colonizing frontier regions, especially in the northwest, with army veterans or needy farmers from more crowded parts of the country. These programs did not significantly alter Kampuchean settlement patterns, however.

Throughout rural Kampuchea, life-style was closely geared to the agricultural cycle, which was based, in large part, on family-oriented subsistence farming. Family members were awake before dawn, and the major portion of the day's work was accomplished before noon, although minor tasks were performed in the cool of the early evening. Electricity has always been rare in village areas, and country people were generally asleep soon after sunset. During the rice-growing season, all family members worked together in the fields, because the work of planting, transplanting, and harvesting must be done quickly. Without mechanical assistance, the work of more than one or two people is needed to grow enough rice to feed a family for a year. Because of the intensive labour requirements of paddy farming, reciprocal extra-family obligations would build up within a village during the agricultural season. Festivals and marriages, celebrated by a whole village, were usually held in the off-season, after the rice had been harvested.

Urban areas

The urban areas of Kampuchea, most of which were abandoned as an official policy in 1975, had developed, for the most part, during the first half of the 20th century as commercial and administrative centres serving their surrounding rural regions. Most of them were located at the intersections of land or river routes and were relatively accessible to the areas they served. Phnom Penh (*phnom* means "hill"; Penh is a woman's name) was Kampuchea's single metropolis before it was abandoned in 1975. Its population fluctuations reflect Kampuchea's recent history. Before the outbreak of war in 1970, it held some 500,000 persons; its population by 1975, then swollen with refugees, numbered some 2,000,000. Several thousand people moved back into the city in 1979, after its official depopulation. In 1970, Bătdâmbâng in the northwest was Kampuchea's second largest city, with about 40,000 people. No other city at that time held more than 20,000.

THE PEOPLE

Kampuchea's first national census as an independent nation was taken in 1962. The results showed that the population totalled about 5,700,000. Overall population density is difficult to determine, because of the enormous losses and movements of people in the years after 1970. Because so much of the country is poorly watered and without inhabitants, the actual density in populated areas is quite high.

Population distribution

The war and social revolution in the 1970s seriously affected the distribution of Kampuchea's population. Between 1975 and 1978, thousands of urban people were forcibly moved into the northwest to cultivate rice and to dig and maintain extensive networks of irrigation canals. During the war with Vietnam in 1978–79, the populated districts adjacent to Vietnam were devastated and deserted. Cities throughout the country were emptied and thousands of smaller population centres were destroyed by U.S. bombing in 1973, fought over in the civil war of 1970–75, or abandoned in 1975–78.

Ethnic distribution. Khmer (Cambodian) stock accounts for the vast majority of the total population. This has produced a homogeneity that is unique in Southeast Asia and has encouraged a strong sense of national identity. Before 1975, other major ethnic groups included the Chinese (350,000), Vietnamese (200,000), Cham-Malays (90,000), various tribal peoples (90,000), and Europeans, mainly French (5,000). All of these minorities, except for the ethnic Chinese and tribal peoples, were expelled from Democratic Kampuchea; many of the Chinese, in turn, sought refuge in Vietnam.

The Khmer are concentrated principally in the lowland regions surrounding the Mekong and Tonle Sap, on the transitional plains, and in the coastal area. They belong to the Mon-Khmer ethnolinguistic group. An end product of centuries of intricate cultural and racial blending, the Khmer descended before 200 BC into the fertile Mekong Delta from the Khorat Plateau of what is now Thailand. They were Indianized by successive waves of Indian influence and in the 8th century AD were exposed to Indo-

Malayan influences and perhaps immigration from Java. This was followed by Tai migrations from the 10th to 15th centuries, by a Vietnamese migration beginning in the 17th century, and by Chinese migrations in the 18th and 19th centuries. As a result of this racial admixture, the Khmer are generally classified as Austroasiatics linked to the Veddoid, Indo-Australoid, and Mongoloid peoples. Their physical characteristics reflect their mixed background. Despite wide variations, they tend to be of short stature (the average height of a male is about five feet, three inches), robust, and muscular. Skin colour is light to brown, and hair is black. About half show the epicanthic fold (a fold of skin extending from the eyelid over the inner corner of the eye; it is sometimes called the Mongolian fold). The typical Khmer family before 1975 consisted of a married couple and their unmarried children. Both sons and daughters usually left the parental home after marriage to establish their own households.

The ethnic minorities

Among the ethnic minorities in Kampuchea before 1975, the Chinese were clearly the most important, for they controlled the country's economic life and retained a high degree of cultural distinctiveness, even though many assimilated into Khmer society through intermarriage. The Vietnamese minority occupied a somewhat lower status, and most of them fled or were repatriated to Vietnam after 1970. Centuries of mutual dislike and distrust have clouded Vietnamese–Khmer relations, and intermarriage is infrequent. The next most important minority, the Cham-Malay group known in Cambodia as Khmer Islām, also maintained a high degree of ethnic homogeneity and was discriminated against in 1975–78. By contrast, scattered evidence suggests that the tribal people of Cambodia, living originally in the forested northeast of the country, received favoured treatment during that period perhaps because the Communists of Democratic Kampuchea had used tribal areas as guerrilla bases during the 1960s. This favouritism may also have reflected that regime's revolutionary concern for the downtrodden or forgotten inhabitants of Cambodia, for in the years before 1970, and indeed for centuries, the tribal minorities (known collectively as *phnong,* or "savages") had been mistreated by the ethnic Khmer.

Religion. The Khmer were almost universally Theravāda (Hīnāyana) Buddhists (*i.e.,* belonging to the earlier of the two great schools of Buddhism, the latter school being represented by the Mahāyāna), and Buddhism was officially recognized as the state religion. Between 1975 and 1978, Buddhist monks were made to work in the fields, like everyone else; images of Buddha from monasteries were destroyed; and temples were used as granaries and barracks. Although the social and psychological characteristics often ascribed to the Khmer—individualism, conservatism, patience, gentleness, and unconcern for material wealth—were often in the eyes of the beholder, they did represent Buddhist ideals toward which many Kampucheans aspired, and Buddhist precepts permeated Kampuchean education and ideology. These precepts were systematically attacked during the years of Democratic Kampuchea, which placed "superstition" in opposition to self-reliance.

Minority populations were not Theravāda Buddhists. Tribal people were animists and the ethnic Vietnamese and Chinese were eclectic, following Mahāyāna Buddhism, Taoism, and such syncretic Vietnamese religious movements as the Cao Dai. The Cham were strict Muslims, and a sizable number of Vietnamese were members of the Roman Catholic Church.

THE ECONOMY

Even before 1975, Kampuchea's economy was one of the least developed of the Southeast Asia region. It was heavily dependent upon two major products—rice and rubber—and consequently was vulnerable to profound annual fluctuations caused by vagaries in rainfall and world market prices. Agriculture dominated the economy. Most rural families were engaged in rice cultivation. Although the tradition of landownership was strong, family landholdings were relatively small. But even with small family farms, the rural population was largely self-sufficient. One hectare of rice paddy provides for the needs of a family of five persons, and supplementary requirements were traditionally satisfied by fishing, cultivating fruit and vegetables, and raising livestock. Famine was rare in Kampuchea, but the self-sufficiency of the rural family produced a conservatism that proved resistant to government efforts before 1975 to modernize the country's primitive agricultural methods.

Resources. Kampuchea has few known mineral resources. Some limestone and phosphate deposits are found in Kâmpôt province, and precious stones in limited quantities are mined in Bătdâmbâng. Iron and coal traces have not justified commercial exploitation. Electrical power sources are mainly dependent upon imported oil.

Rice production

Agriculture and fisheries. Rice is Kampuchea's principal food, its major crop, and, in times of peace, its most important export commodity. Rice is grown on most of the country's total cultivated land area. The principal rice regions surround the Mekong and the Tonle Sap, with cultivation particularly intensive in Bătdâmbâng, Kâmpóng Cham, Takêv, and Prey Vêng provinces. Lacking sufficient irrigation systems, Kampuchea has traditionally produced only one rice crop per year.

Under the government of Democratic Kampuchea, great strides were taken to build irrigation systems throughout the country, using mass labour and abolishing private ownership of land. According to scattered information, the results were often impressive, and in parts of the country peasants were able to grow two, and more rarely three, crops of rice per year. These irrigation works broke down or were abandoned in the war with Vietnam in 1978 and in the civil war that followed. In 1979–80 famine was widespread, and hundreds of thousands of Kampucheans fled into Thailand or became dependent, inside Cambodia, on food provided by international aid.

Under the traditional patterns of agriculture, planting normally begins in July or August, and the harvest period extends from November to January. The amount of rainfall, when there is little irrigation, determines the size and quality of the crop. Other food products include corn (maize), beans, soybeans, and sweet potatoes. The principal fruit crops, all of which are consumed locally, include oranges, bananas, and pineapple.

Fisheries and livestock are important components of the domestic economy. Fish in its various forms—fresh, dried, smoked, and salted—constitutes the single most important source of protein in the Kampuchean diet, and subsistence fishing is part of every farmer's activity. The annual freshwater catch includes perch, carp, lungfish, and smelts. Cattle, particularly water buffalo, are used principally as draft animals in the rice paddies and fields. Hog production has also played a large role in agriculture. The efforts of the government of Democratic Kampuchea to increase the number of livestock—seriously depleted by years of war—broke down in 1978, when war and counterinsurgency again occupied the government's attention.

Industry. Although industrial development remains at a low level, successive governments made strong efforts to build a modest industrial base suitable to the needs of the country.

Finance and trade. Foreign trade has been at a standstill in the years of warfare, and imports have taken the form of foreign aid. Traditionally, the bulk of Kampuchea's exports, consisting almost entirely of rice, rubber, corn (maize), and other agricultural products, went to other Asian nations, while imports came mainly from Japan, the United States, and western Europe.

The civil war in Kampuchea of 1970–75 devastated the countryside, sharply reduced foreign trade, and destroyed the nation's fragile economic infrastructure. After the Communist victory in 1975, large-scale economic policies involving the depopulation of urban areas, the construction of giant irrigation works, and the pursuit of industrial self-sufficiency met with mixed results and certainly left the people with few resources with which to fight a full-scale war against Vietnam. The port of Kâmpóng Saôm (formerly Sihanoukville) was used by both Democratic Kampuchea and the People's Republic of Kampuchea, largely to receive foreign aid. During the war of 1978–79, most of the earlier progress made in the countryside was erased.

Changing development policies

Administration of the economy. In seeking rapid economic development, the government in 1963 adopted a socialistic policy characterized by nationalization of the private banking system, establishment of a government monopoly over imports and exports, and extensive state participation in building and managing industrial enterprises. This policy proved to be unsuccessful, mainly because of defective economic planning, insufficient labour and managerial skills, and rigidly fixed agricultural prices that discouraged production incentive.

The government of the Khmer Republic, in turn, supported the idea of industrialization, but was unable to accomplish much because Cambodia's cities were under siege for most of 1972–75. The regime of Democratic Kampuchea pursued a policy of self-reliance modelled on that of the People's Republic of China. This meant that existing factories were expanded and new ones were built to meet local needs for cloth, cigarettes, sheet rubber, farm machinery, fertilizer, and cement. Smaller factories were installed inside the agricultural communes.

Transportation. Kampuchea's inland waterways and road systems constitute the main transportation routes. Each is invariably affected by the floods of the rainy season, which result in heavy silting and washouts (flash floods). Railroads rank third in significance. Domestic shipping and civil air facilities are limited. Maritime commerce is carried out almost exclusively by foreign vessels.

The road system

The road system eventually surpassed the country's inland waterways as the principal means for moving cargo and passengers. The network was originally designed by the French during the protectorate period to link the agricultural hinterland with the port of Saigon, now Ho Chi Minh City, Vietnam. Consequently, the system did not serve Kampuchea as a whole. Extensive land tracts in the northern, northeastern, and southwestern parts of the country were roadless. Of the total road network, only a fraction has been paved; other roads have been surfaced with crushed stone, gravel, or laterite or have been simply graded without being paved. The country's longest bridge—a ten-span structure more than 2,300 feet in length, traversing the Tônlé Sab at Phnom Penh—was destroyed in 1975.

The inland waterways have a collective extent of 1,200 miles, of which more than 90 percent are part of the Mekong and Tonle Sap systems. Phnom Penh, located about 200 miles from the mouth of the Mekong, can be reached by oceangoing vessels of less than 13-foot draft. North of Phnom Penh, the Mekong is navigable to Krâchéh for rivercraft, but rapids and winding channels in the 117-mile section between Krâchéh and the Laos border generally preclude commercial navigation.

Port at Kâmpóng Saôm

Kampuchea's single maritime port is located at Kâmpóng Saôm on the Gulf of Thailand. Completed in 1960, Kâmpóng Saôm can provide unlimited anchorage for oceangoing ships. The port is of strategic importance to Kampuchea, and considerable industrial development has taken place in the area. A modern four-lane highway links Kâmpóng Saôm with Phnom Penh.

The railroad system is owned and operated by the Kampuchean government. One line, completed prior to World War II, connects Phnom Penh with Paôy Pêt on the Thai frontier—a distance of 239 miles—and facilitates the movement of milled rice from the western provinces of Bătdâmbâng, Poŭthĭsăt, and Kâmpóng Chhnăng. Another line, completed in 1969, connects Phnom Penh with Kâmpóng Saôm, covering a distance of 168 miles through the provinces of Kândal, Takêv, and Kâmpôt.

ADMINISTRATIVE AND SOCIAL CONDITIONS

Government. The structure of government under the People's Republic of Kampuchea was not made clear to the international community. Local government, as it was known under more Western-style regimes, had probably disappeared under the pressures of civil war and contending Communist ideologies. Nothing was known of the judicial system then in effect; it appeared that no judicial system functioned under Democratic Kampuchea. One peculiarity of all the regimes that followed the Kingdom of Cambodia was that each condemned to death the leaders of the preceding regime.

Armed forces. The armed forces of Democratic Kampuchea include a number of ethnic Kampuchean forces loyal to the People's Republic of Kampuchea. There are considerable numbers of Vietnamese troops garrisoned throughout most of the country.

Education. Kampuchea's educational system, as it had developed in the first 70 years of the 20th century, was another casualty of warfare and ideology. Under Democratic Kampuchea, only primary schools were open; older students attended irregularly scheduled political and technical courses, often held in the communes. The People's Republic stated its eagerness to reopen the schools, but it was hampered by a shortage of funds, teachers, books, and students, as well as the continuing civil war.

Literacy rates

Literacy levels in Kampuchea have always been relatively low. Although Democratic Kampuchea claimed to have eliminated illiteracy in 1975–76, such an achievement seemed unlikely, given the conditions in the countryside, the absence of schools, and the government's discrimination against educated members of the "old society." Indeed, according to reports of refugees, illiterate peasants often rose to positions of power under that government.

Health and welfare. Throughout Kampuchea's history, an acute shortage of medical personnel has been a major obstacle to the implementation of an effective public health program. Even before the civil war, Kampuchea had few doctors, hospitals, or medical facilities. The civil war of 1970–75 strained and eroded this fragile structure. Democratic Kampuchea moved medical personnel onto collective farms and, as part of its policy of self-reliance, encouraged non-Western medical practices based on the use of local herbs. Because of unsettled conditions, a lack of sanitation, and a shortage of medicine, epidemics of cholera and malaria were reported and instances of other diseases, especially those related to malnutrition, were frequent.

CULTURAL LIFE

Khmer architecture

Before 1970, Cambodian culture and artistic expression were overshadowed by the greatness of the past. Although owing much to Indian influence, the achievements of the Khmer Empire represented original contributions to Asian civilization. The magnificent architecture and sculpture of the Angkorean period (802–1432), as seen in the temple complexes at Angkor Wat and Angkor Thom, marked the apex of Khmer creativity. Following the capture of Angkor by the Thai (15th century) and the crumbling of the empire, the region underwent four centuries of foreign invasions, civil war, and widespread depopulation. It was not until the establishment of the French protectorate in 1863 that internal security was restored, the country's borders stabilized, and efforts undertaken to revive traditional Khmer art forms. After gaining independence from France in 1953, the government placed particular emphasis upon accelerating that revival by establishing a national school of music, a national school of ballet and theatre, and a fine arts university. This coincided with the rapid expansion of elementary and secondary school facilities, and the emergence of education as the most important factor of social mobility.

Music occupied a dominant place in Kampuchean culture. It was sung and played everywhere—by children at play, adults at work, by young men and women while courting—and invariably was part of the many celebrations and festivals that took place throughout the year at Buddhist temples in the rural countryside. Instruments used in full orchestras included xylophones with wooden or metal bars, one- and two-stringed violins, wooden flutes, oboes, and drums of different sizes. The Kampuchean musical scale had five tones, compared with seven in the Western scale. Orchestral music had no harmony, in the technical musical sense. The players followed the lead of one instrument, usually the xylophone, and improvised as they wished.

Dancing and drama

Dancing and drama were also popular forms of artistic expression. The Royal Ballet in Phnom Penh exemplified the classic, highly stylized dance form adapted by the Kampucheans and Thai from the ancient dances of Angkor. Accompanied by an orchestra and choral narration, the dancers acted out stories and legends taken principally from Hindu epics such as the *Rāmāyaṇa.* In the countryside, folk drama and folk dances were performed at festivals and weddings by wandering troupes. The actors invariably depicted stereotyped characterizations familiar to all: the country yokel, the clumsy lover, the beautiful princess, the cruel father, and the greedy merchant. The visual arts revealed the essential conservatism of the Kampucheans. Ancient themes were preferred and rarely was there an effort to improve or adapt. The principal crafts were weaving, working silver and gold, making jewelry, and the sculpture of wood and stone. (L.C.O./D.P.Ch.)

For statistical data on the land and people of Kampuchea, see the *Britannica World Data* section in the BRITANNICA WORLD DATA ANNUAL.

History

The importance of Cambodia's cultural and historical contributions to the development of mainland Southeast Asia is out of proportion to its present reduced territory and limited political power. At the height of its power (11th–13th centuries), the Khmer (Cambodian) state stretched across a vast area of the Indochinese Peninsula and incorporated much of southern Vietnam, Laos, and Thailand, and the cultural influence of Cambodia on the other peoples of Southeast Asia has been enormous.

CAMBODIA BEFORE THE 9TH CENTURY AD

Knowledge of Cambodian prehistory is extremely limited. Excavations at three prehistoric sites—Samrong Sen and Longprao, in central Cambodia, and Melouprei, in northeastern Cambodia—have yielded remains from the Neolithic Age sufficient to permit only the barest reconstruction of life in the region more than 2,000 years ago. Nor is it possible to link the makers of the various artifacts that have been discovered with the succession of ethnic groups that moved through the Indochinese region as a whole in the prehistoric period, peopling both that region and the maritime areas of Southeast Asia as well.

Early inhabitants

Moving by both land and sea routes from areas to the north of Indochina, successive waves of migrants passed through the area. The original Australoid inhabitants were succeeded by Melanesians and then by Indonesians probably moving southward in two distinct phases, one by land and the other by sea. The languages spoken by these various migratory groups are increasingly seen by linguists as having had a distant common origin, whatever the later differentiations established among them. In the area of modern Kampuchea, a language of the Mon-Khmer eventually became the dominant tongue after the Khmers, relative latecomers to the scene, had asserted themselves as political leaders, first in the state of Funan—centred in present southern Kampuchea and the southernmost part of Vietnam—and later in the north-central kingdom of Chenla.

The story of human activity in the area before the establishment of Funan, which Cambodian tradition ascribes to the 1st century AD, is not known in any detail. Settlements were close to lakes and rivers; the inhabitants engaged in agriculture and maintained cattle and pigs as domestic animals. They hunted with weapons that included the bow and arrow, for which arrowheads of polished stone, bone, and iron have been found. Metal cultures reached the Cambodian region well before the 1st century AD, and a site at Melouprei (now Phumĭ Mlu Prey) has revealed traces of iron and bronze casting. Close to the modern capital of Phnom Penh, the discovery of a bronze urn having close links with the Dong Son style of northern Vietnam (5th to 1st centuries BC) suggests that highly developed metalworking may have been taking place in pockets of Cambodian territory as early as the 4th century BC. Despite this evidence of growing technical capabilities, excavations at Oc Eo, Funan's major port on the southern coast of modern Vietnam along the Gulf of Thailand (Gulf of Siam), suggest that the Neolithic culture characterized by polished stone implements survived alongside the more advanced metal culture.

Funan. While future archaeological work may lead to a better understanding of Cambodian prehistory, true historical knowledge begins with the rise of Funan. From its claimed mythical beginnings in the 1st century AD until its incorporation into the successor state of Chenla in the 6th century, Funan was vitally important as a recipient of Indian culture, which shaped the political institutions, art, and general culture of later Khmer states. Always regarded by later Khmer dynasties as the state from which they sprang, Funan owed its prosperity to its position on the great east–west trade route between India and China. The name of the state is generally believed to come from a Chinese transliteration of a Khmer word *bnam* (*phnom* in modern Cambodian), meaning "hill." If correct, such a derivation is important as evidence of substantial Khmer influence over a state in which there was also a strong Indonesian ethnic and cultural presence.

Legendary origins

The national birth legend of the Funanese state tells of the arrival of a foreign Brahmin, Kaundinya, who overcame and subsequently married a local woman named Willow Leaf and set Funan upon its path to greatness. The legend is suggestive of the early influence of Indian thought and culture in the Indochinese area. Rejecting an earlier belief that large colonies of Indians settled in Southeast Asia, modern scholarship increasingly holds to a view that a relatively limited number of Indians, sometimes Buddhist traders, at other times Brahmins (high-caste Hindus), established themselves in positions of influence and propagated their more advanced knowledge of statecraft, religion, and art. A part of this view is the assumption that the existing populations of such areas as Funan were far from primitive savages and may have mastered the intricacies of wet rice cultivation.

The extent of implanted Indian influence in Funan before the middle of the 3rd century is uncertain, but a clearer picture of that period emerges from records left by two Chinese ambassadors who visited the country, *c.* AD 245. By that time Funan had extended its political power over the lower reaches of the Mekong and Tônlé Sab rivers and as far west as a section of the Malay Peninsula where trade passed from the Indian Ocean into the Gulf of Siam. This advance had taken place under the direction of King Fan Shih-man (*c.* 205–*c.* 225) and his nephew, Fan Chan (*c.* 225–*c.* 240), who took the throne as a usurper. Their successor, Fan Hsun (*c.* 240–*c.* 287), another usurper, was ruling at the time of the Chinese ambassadors' visit. The account of one of these ambassadors, K'ang T'ai, speaks of Funan as a land of walled cities whose inhabitants went unclothed through the streets but whose rulers gathered taxes in the form of precious metals, jewels, and perfumes. An Indian-based script was in use, and there were libraries and archives. The reference to the use of an Indian script is a clear indication of Indian influence, but it was not until the end of the 4th century that the full impact of Indian culture was felt in Funan.

This second "Indianization" of Funan came after a period about which there is little if any knowledge of developments within the state. Fan Chan's and Fan Hsun's reigns had seen the growth of official relationships between Funan and China and India. Relatively little is known, however, of the extent to which these official contacts were matched by a significant spread of Indian culture among the population at large. The appearance of a foreigner, Chandan, on the throne of Funan in 357 has been seen by some historians as a reinforcement of Indian influence and by others as reflecting a link with the Iranian world. Whatever may be the case, under Chandan's successors in the latter part of the 4th century and particularly in the 5th, there is a sense of Indian influence at a much wider level. Sanskrit inscriptions provide a more precise chronology of the reign of Funanese rulers and reveal that Hinduism and Mahāyāna Buddhism coexisted within the state.

Chandan's successor, according to the Chinese *History of the Liang,* was another Brahmin called Kaundinya, whose arrival in Funan was received with popular rejoicing and who instituted many reforms based on Indian models of law and administration. However mythical this Kaundinya may have been, the account of his arrival appears to capture correctly the importance of a new infusion of Indian culture. This cultural influence is abundantly apparent in the records that have survived describing the reigns of Funanese rulers during the 6th century. Funan was at the height of its power under the rule of King Kaundinya Jayavarman (d. 514).

'unan's reatest eriod

Excavations at the port city of Oc Eo have indicated the extent to which Funan prospered as a trading kingdom with links, indirectly, even with countries of the Mediterranean. Aerial photography has revealed a complex series of irrigation and communication canals that were established in the hinterland away from the coast. While Oc Eo may have been the most important port city, the royal capital was inland from the sea, at one time near the modern Cambodian town of Banam in Prêy Veng province and bearing the name Vyādhapura (Sanskrit: "City of the Hunters").

From passages preserved in Chinese dynastic histories there is considerable information on the material culture of Funan in its heyday as well as insights into the state's religious life. It was a slave-owning society in which the rich decked themselves with gold and silver jewelry and wore rich brocades. The king lived in a richly constructed palace and travelled on the back of an elephant. The common people amused themselves with cock fights and contests between fighting pigs. Justice was administered through trial by ordeal. Since most buildings were constructed of wood, little is known of the style of Funanese building. A few architectural remnants at Oc Eo reinforce the evidence of Indian influence. The same initial source of inspiration was present in Funanese statuary. Sculpture from the 4th century has very obvious links with Gupta art in India. In the 6th century a distinctively Funanese art emerged in which Indian iconography was given a Cambodian form.

Hinduism was the religion of the rulers of Funan in the 5th and 6th centuries, but Buddhism was widely diffused in the kingdom also. There is evidence that Buddhism also enjoyed a measure of royal patronage. The importance of Buddhism in Funan in the 6th century was so great that a Chinese embassy visited the country between 535 and 545 to ask the ruler, Rudravarman (514–539), for Buddhist texts and teachers who could give Buddhist instruction in China.

Rudravarman was the last king of Funan to reign over an independent state. The subjugation of Funan by the rulers of the emerging state of Chenla marks a major division in Cambodian history. Although there are gaps in the history of Funan, the importance of the state is clearly apparent. It was through Funan that the initial process of Indianization took place, and it was this state which was always regarded by later Cambodians as the origin of their dynasties. Though this view was propounded in part to enhance legitimacy, it also reflected an awareness of the vital links that Funan had with its Angkorian successor state (9th-15th centuries). While the details are unclear, Funan before its absorption into Chenla already had links with the Khmers, whose base area was well to the north in the middle Mekong or Champassak area of modern Laos. There is clear evidence that Khmer princes were already associated with the Funanese ruling house and that some areas under the control of the Funanese state in the north of the kingdom were settled by Khmers.

henla– ınan lation- ıip

Chenla. The first ruler claiming sovereignty over both Chenla and Funan was Bhavavarman (reigning 598), the grandson of Rudravarman of Funan, a fact which, along with his claim to power over both kingdoms, has led some authorities to suggest that the Chenla and Funan kings originally belonged to two branches of the same dynasty. Since Funan continued to send embassies to China until the beginning of the 7th century, the southern state apparently retained a degree of autonomy under Bhavavarman, whose capital was probably on the northern shore of the Cambodian Great Lake (Tonle Sap). Bhavavarman's brother, who succeeded him as Mahendravarman (*c.* 600–*c.* 611), left inscriptions that reveal considerable activity in the northeastern region of modern Cambodia around Krâchéh and Stœng Trêng. With Mahendravarman's son, Iśanavarman (*c.* 611–635), the power of the Chenla rulers over the declining state of Funan became absolute. Under Iśanavarman the power of Chenla was extended to the west as well as to the south, and good relations with the state of Champa (in modern Vietnam) were sealed by Iśanavarman's marriage to a Cham princess. Iśanavarman's capital, Iśanapura, at the modern site of Sambor Prei Kuk, about 12 miles north of Kâmpóng Thum, is the most impressive group of pre-9th-century remains so far discovered in Cambodia. A series of brick towers prefigure a number of later developments in Khmer art, and the site remained important as a centre of scholarship during the Angkorian period.

Under Iśanavarman's successors, Bhavavarman II (whose reign dates are unknown) and Jayavarman I (who reigned through most of the latter half of the 7th century), Khmer power extended throughout the Indochinese region. Yet, despite the description of Jayavarman as "glorious lion of kings" on one of his many inscriptions, the later division of the Chenla dominions into two separate regions probably began during his reign. Having expanded their power rapidly, the Chenla rulers were unable to maintain strict control over the various territories that they claimed the right to rule. Nevertheless, whatever political difficulties later sprang from the period, the 7th century was a time of continuing artistic development. Chinese accounts of court life under Iśanavarman give a picture of a rich material culture probably little different from that which had existed in Funan. Indian influence continued to be important in the art of the 7th century, and the two great Indian religions of Hinduism and Buddhism continued their coexistence. Of great significance for the future was the fact that the Khmers, when they established their settlements in lowland Cambodia, used hydraulic techniques for agriculture that they had developed in lands further to the north. Their "captive water" technique, which dammed water to be distributed by canals to less elevated regions, was later to be exploited to its full in the complex systems of Angkor.

Chenla's decline

Following Jayavarman I's death, the state of Chenla split into rival petty principalities. The history of the 8th century is obscure, but certain broad trends may be recorded. Land Chenla, the name given by the Chinese to the northern of the two divisions, apparently enjoyed a fairly stable existence with the centre of the state probably to the north of Angkor. By contrast, Water Chenla, the southern division, was an unsettled state beset by dynastic rivalries. As heir to Funan's position as an entrepot on the great east–west trade route, Water Chenla was grievously affected in the 6th and 7th centuries when traders began to bypass the Malay Peninsula, travelling to China by way of the Indonesian state of Śrivijaya in southern Sumatra. Along with the interruption of the trading link with the West, there was a sharp decline in direct cultural contact with India during the 7th century. Despite the diminution, and even elimination, of direct Indian influence, Khmer artistic achievements remained impressive until the latter half of the 8th century. And even this period of relative decline was rapidly followed by a dramatic efflorescence of Khmer artistic output, inspired by local genius rather than foreign models. This resurgence was the cultural accompaniment to the emergence of the Angkorian dynasty under the great Khmer ruler Jayavarman II (reigned *c.* 802–850), a man only distantly related to the ruling dynasties of pre-9th-century Cambodia.

THE ANGKOR PERIOD

Jayavarman II and his successors. In the latter part of the 8th century the Śailendra rulers of Java appear to have created a sphere of influence in Cambodia, and it was from Java that Jayavarman II returned to establish his kingdom about the year 800. Not the least remarkable aspect of his reign is the fact that since he left no inscriptions his achievements must be reconstructed from 11th-century inscriptions that give a clear picture of the vital importance of Jayavarman's reign. The concept of a unified Cambodian state was established, never to be forgotten by his successors, and a new state religion that established the Khmer ruler as a *devarāja* (god-king) was introduced during his reign.

On his return to Cambodia from Java, Jayavarman first established a capital at Indrapura, in modern Kâmpóng Cham province. But his subsequent three capitals were all in the region farther north, close to but not actually in the area of the later Angkor complex. Strategically, the area on the east of the Great Lake was distant from the threat of external attack. The lake was a source of fish, and the plains about it were, with the application of hydraulic engineering, a great source for rice. Once Jayavarman had chosen this region as the centre for Khmer power, it remained so for most of the Angkorian era.

Khmer kingship

At his last capital of Mahendraparvata, in the Kulén hills to the east of Angkor, he took part in a Śaivite Hindu ceremony consecrating him as a *devarāja* and establishing both him and his kingdom as independent of any other power, especially that which the Javanese had claimed to exert over Cambodia. In this ceremony—perpetuated by Jayavarman's Angkorian successors—the king's sacred personality, the very essence of the kingdom, was enshrined in a sacred *liṅga* (a ritual phallic symbol) that was housed in a pyramidal structure called a temple mountain. This temple mountain was both the centre of the earthly kingdom over which the king ruled and the symbolic centre of the universe within which the kingdom existed. The details of this highly complex concept are possibly less important than the recognition that through participation in it Jayavarman and his successors saw themselves as universal rulers, subservient to no other state. The same symbolic purpose led to the establishment of the great Angkor temples, which were prefigured by those of Jayavarman's reign. In sculpture and the plastic arts, generally, the 9th century witnessed a revival and development of styles and techniques that blended Khmer traditions with those of Java and Champa.

After Jayavarman's death in 850 he was succeeded by a son, Jayavarman III (reigned 850–877), who left no significant mark upon Khmer history. The next ruler, Indravarman I, a cousin of Jayavarman III and possibly a usurper, played an important role in the development of the Angkorian state. Although his capital was not yet located on the site of the later Angkorian kings, it was at Roluos, not far distant, where Jayavarman II had founded one of his four capitals. At Roluos, Indravarman built two great stone temple mountains, Bakong and Preah Ko, which mark the beginning of Khmer classical art, the form from which the later temples developed. Even more important, however, was Indravarman's construction of a vast irrigation system that, by means of an artificial reservoir, permitted the assured exploitation of rice lands that otherwise would have remained unproductive. The later elaboration of artificial irrigation systems allowed the Khmers at Angkor to maintain a densely populated and highly centralized state in a relatively limited area. Only with such an irrigation system was it possible to feed the immense labour force necessary for the construction of the Angkor temples.

The 8th-century empire

Indravarman's successor, Yaśovarman I (reigned 889–900), inherited a kingdom from his father that was at peace, and he proceeded to build solidly upon Indravarman's achievements. His new capital of Yaśodharapura, the first Khmer city in the Angkor area proper, centred on the hill Phnom Bakheng, which Yaśovarman crowned with a temple. The city complex occupied some 16 square miles; the irrigation of this area depended upon a new system, fed from a great reservoir east of the city, the so-called *baray oriental,* or eastern *baray* of modern times, which was fed from the Siĕmréab River. The power which Yaśovarman accumulated at Angkor enabled him to extend Cambodian domination over a territory as large as Funan's at that empire's height. It is uncertain whether this extension of power was the result of armed conquest or the readiness of other states to acknowledge the suzerainty of the leading power in the region. Some authorities believe that achievements attributed to Yaśovarman rightly belong to later reigns, but he was certainly a figure of great importance. He built monasteries throughout his kingdom for both Śaivites and Buddhists, and the remains of his temples, particularly that of Phnom Bakheng, show a continuing development of the Khmer genius of blending architectural construction with religious symbolism.

The century following Yaśovarman's presumed death in 900 was one for which political records are now at a minimum. External commentary from Chinese sources vanished with the internal division of China after the fall of the T'ang dynasty. Within Cambodia itself the progress of the state must be charted through inscriptions, which provide a highly selective view of developments, and through the material evidence of temples, both royal and nonroyal in foundation. Nevertheless, close study of the extant inscriptions does permit some insight into the life of Cambodian rulers, the narrow oligarchy of priests and advisers who supervised the complex bureaucracy that presided over the daily business of the state, and even, occasionally, the activities of less elevated members of the population. Six Khmer kings ruled during the 10th century. Probably the most significant political developments of their reigns were the temporary abandonment of the Angkor capital under the usurper Jayavarman IV (reigned 928–942) and its subsequent re-establishment by Rajendravarman II (reigned 944–968). Rajendravarman's successor, Jayavarman V (968–1001), is known chiefly for his continuation of Rajendravarman's efforts to control Champa and for his promotion of scholarship and learning. In religious matters the 10th century continues to provide a record of Hindu cults coexisting with Buddhism. Although Saivism (worship of the Hindu god Śiva) was the principal royal cult, Buddhism also received its share of patronage.

Angkorian government and society. Though the Khmer monarchs were absolute rulers in theory, their power was often limited by the complex bureaucracy of the Angkorian state. The king was the final arbiter in legal matters and sometimes led his troops in battle, but much of his time was taken up with ceremonial and religious functions deemed necessary to insure the well-being of the state. If he was not a forceful personality, he might be reduced to a figurehead.

The bureaucracy

The bureaucracy was composed of learned Brahmins (members of hereditary priestly families) and military leaders who were sometimes members of the royal family. These men formed a class separate from the population at large; there is considerable evidence that their separate identity was reinforced through intermarriage. They were particularly influential in regions more distant from the capital. Although the construction of a temple mountain was the prerogative of the king alone, the great officials founded their own shrines, each supported by surrounding villages, which met the high cost of maintaining the temple and its priests. None of these shrines rivals the great royal foundations in size, but the temple of Banteay Srei, a little to the northeast of the Angkor complex, is considered by many to be the most perfect of all Khmer temples. It was built in 967 by a Brahmin, Yajñavarāha, a man of royal descent who had supervised the education of both Rajendravarman II and Jayavarman V.

Between the great officials, on the one hand, and the peasants and slaves (about whom little is known), on the other hand, there was another group, whose existence is recorded in inscriptions but whose role in the Khmer state has yet to be studied in any detail. This was a group of land-owning men who held no official position and who, unlike the priestly officials of the great foundations in provincial regions, appear to have had little contact with the capital. Their existence, and the existence of "castes" of teachers and priests who were recruited from provincial areas but who were clearly not of the elevated status of the great officials, provided some qualification to the narrow classifications into which Khmer society has frequently been divided in modern historical accounts. While the size of these intermediate groups is unknown, their existence suggests a more complex society than that which envisages a single, sharp division between the king and his officials, and the peasantry and slaves.

The 11th century. At the beginning of the 11th century one of the most energetic of the Khmer monarchs, Suryavarman I (reigned *c.* 1002–50), challenged the position of Jayavarman's legitimate successor, Udayadityavarman I, to hold the throne. A Cambodian prince, but with scant direct claim to the throne, Suryavarman had defeated Udayadityavarman by 1002, but a long civil war dragged on until he finally gained control over the whole kingdom about 1011. Although the evidence comes from chronicles of a considerably later period, Suryavarman appears to be correctly credited with the extension of Khmer power over much of modern southern Thailand and over Mekong valley regions as far north as Luang Prabang. In contrast to this activity in the west, his relations with Champa and Vietnam, to the east, seem to have been calm and stable. His long reign enabled the restoration of many of his predecessors' foundations both at Angkor itself and in other parts of the country, and the completion, notably, of the Phimeanakas and Ta Keo temples begun by Jayavarman V.

new
ynasty

Suryavarman's son, Udayadityavarman II (reigned 1050–66), carried on his father's expansionist policies. While some historians judge that this led to the Khmer state reaching its furthest limits, these policies also provide part of the explanation for the record of revolts that marked Udayadityavarman II's reign. Cham encouragement appears to have been behind a revolt in the southeast of the kingdom. And in the west the Khmers found themselves in confrontation with the newly victorious Burmese, who had conquered the Mon capital of Thaton. Other revolts within Cambodia itself are believed by some historians to have reflected resentment at the favour which the ruler showed toward Buddhism, a religion that his father had done much to promote. Despite the record of warfare and revolt during his reign, Udayadityavarman II found time to build another great *baray,* or reservoir, the western *baray,* which greatly increased the cultivable area of the region around Angkor. Udayadityavarman's successor was his brother, Harshavarman III (reigned 1066–80), who was unable to prevent a major incursion into Cambodia by a Cham army that marched to Sambhupura, a city on the Mekong near the modern town of Krâchéh, and sacked it. Harshavarman was overthrown by a usurper, who reigned as Jayavarman VI (1080–1107), but possibly never occupied the capital at Angkor. He and his successor, Dharanindravarman I (1107–13), had to contend with the opposition of members of Harshavarman's family.

From Suryavarman II to Jayavarman VII. *Suryavarman II.* This opposition was finally and forcefully crushed by Suryavarman II, one of the greatest figures in Cambodian history, who came to the throne in 1113. Reducing the kingdom of Champa to vassal status, he annexed it completely when the Cham king would not cooperate in his campaigns against the Vietnamese state farther north. Suryavarman's expeditions against the Vietnamese took his armies as far north as Thanh Hoa in modern North Vietnam. The extent of his success in the Mae Nam Chao Phraya Valley is open to some doubt, but the substantial Khmer influence found in the architecture of the Thai state of Lopburi suggests an important impact on that region (present Thailand). Khmer influence also reached well down into the Malay Peninsula.

ıryavar-
an's
npire

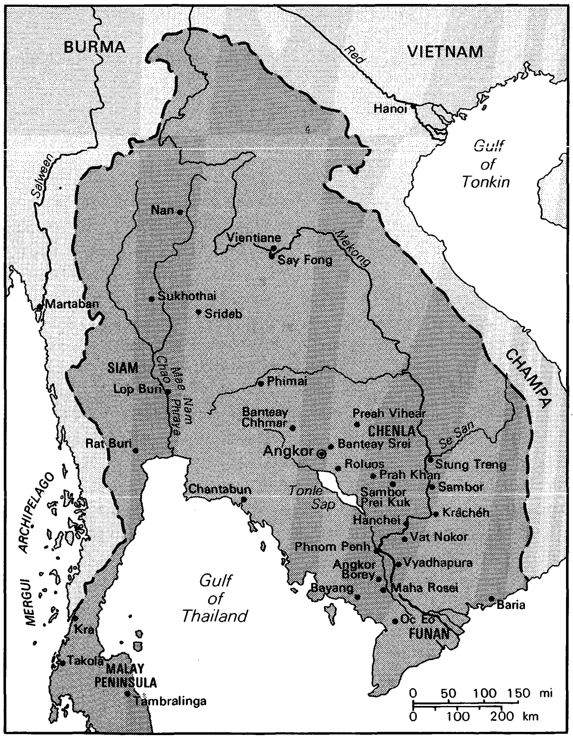
Suryavarman's military and political conquests are matched by the magnificence of his achievements in building the great temple complex of Angkor Wat. Completed at about the time of his death in 1150, Angkor Wat was only one of the outstanding architectural achievements of his reign. Beng Mealea, some 25 miles to the east of Angkor, and Banteay Samre on its outskirts are other magnificent examples of Khmer art and architecture.

Suryavarman blended Śaivism and Vaiṣṇavism (the cult of Vishnu, or Viṣṇu) in his state religion. Within the inner sanctuary of Angkor Wat was a gold statue of Vishnu mounted on the fabled winged Garuḍa, representing the deified Suryavarman as Vishnu. Yet once again, there was no religious exclusiveness within the Khmer state. Buddhism, even if overshadowed by the Hindu cults, continued nonetheless to have its adherents.

The first period of decline. Suryavarman's successes rested on uncertain foundations, and by the end of his reign the combined effects of an overambitious external policy and the demands that his building program placed upon the resources of the state entailed heavy costs. His repeated expeditions against the Vietnamese failed to achieve his goal of subjugating that energetic neighbouring state. A year before his death (*c.* 1150), the Chams freed themselves from Khmer domination and their strength in the immediately succeeding decades was directed toward wiping out the stigma of Khmer overlordship by attacking the Cambodian heartland. Indeed, between 1050 and the accession of Jayavarman VII in 1181, the fragility of the Khmer state after a period of greatness was demonstrated time and again. Internal rebellion was added to the challenge of the Chams to further weaken the power of the ruler. Suryavarman's immediate successor, Dharanindravarman II (1150–60), had an apparently uneventful reign, notable largely for the fact that the new king was a devotee of Buddhism rather than of Hinduism. But at his death the rightful successor, Jayavarman, was passed over and the throne was occupied by his younger brother Yaśovarman II (reigned 1160–*c.* 1166). In the space of six years Yaśovarman faced a peasant or slave revolt, which was successfully subdued, and the rebellion of an ambitious official, Tribhuvanadityavarman, which caused the king's dethronement and death. Although Jayavarman came out of exile in Champa to defend the king, he arrived after the usurpation had taken place and again retired from the scene. During Tribhuvanadityavarman's reign the resurgent Chams began their long campaign of attrition against the Khmers. Border skirmishes begun in 1167 finally culminated in the humiliating and highly successful Cham attack against the capital at Angkor in 1177. The capture of the capital and the death of the ruler left Cambodia in chaos.

Cham
attacks

Adapted from J. Buttinger, *The Smaller Dragon: A Political History of Vietnam* (1970); Praeger Publishers



Khmer Empire, *c.* 1200.

Jayavarman VII. The accession of Jayavarman VII brought a change. This ruler ranks with Suryavarman II among the greatest of all Angkorian rulers. His energy was phenomenal: in both the success of his campaigns against the Chams and his devotion to temple construction he revivified the Khmer empire. As was the case with Suryavarman II, however, this massive outpouring of energy exacted a heavy cost. Jayavarman's policies so exhausted the state that, once his leadership had gone, no other ruler was able to maintain it in the same manner. The most active builder of all the Angkorian monarchs, he left a large number of inscriptions. To his reign belongs the great complex of Angkor Thom, which rivals Angkor Wat in magnificence.

Jayavarman was nearly 60 when, following a period of almost constant campaigning, he achieved sufficient control to assume the kingship in 1181. He repaid in full the injuries his country had suffered at the hands of the Chams. The Khmer armies laid waste to their kingdom and sacked the Cham capital of Vijaya. Through skillful use of Cham allies opposed to their own ruling family, Jayavarman prepared a campaign that finally led to the annexation of Champa, which remained under Khmer control from 1203 to 1220. Within Cambodia itself Jayavarman was unchallenged after the first year of his reign. To the west and south of Cambodia, the Chao Phraya Valley and parts of the Malay Peninsula recognized Cambodian suzerainty. An inscription near modern Vientiane, Laos, records his domination of that area.

Though Jayavarman's Buddhist adherence was to the Mahāyāna school, which had long been dominant in Cambodia, during his reign Theravāda Buddhism was introduced into Burma by the Mon monk Chapata. Among Chapata's companions was a Khmer prince who has been identified as one of Jayavarman's sons. This link with the Theravāda school foreshadowed its later patronage by the Khmer kings, which led to its supplanting the Mahāyāna school and so to a major change in the state's religion. Although modern scholarship rejects the view that the conversion of the Cambodian state to Theravāda Buddhism provides a single convincing explanation for the decline of Angkorian power, the change when it came was a profound one, not least because the Theravāda cult did not call for the construction of great temples as did both Hinduism and Mahāyāna Buddhism.

Jayavarman VII's greatest energies during his reign, as far as external policies were concerned, had been directed toward the conquest and subjugation of the Chams. Before the end of his life, however, the most formidable threat to Khmer power was developing to the west.

The Thai threat

The Thais had been filtering into the Chao Phraya Valley from well before the beginning of the 12th century. In a famous section of the Angkor Wat bas-relief, Thai mercenaries are depicted as ill-disciplined and strangely dressed men of little civilization. Thai chieftains living under Khmer rule in what was the northwestern extension of the Cambodian empire were, however, growing in power and absorbing much of the civilization of their overlords. Only 20 years after Jayavarman's death the Thais of the state of Sukhothai were powerful enough to break the shackles of foreign rule. That this should have been so emphasizes the rapid reversals that followed the end of Jayavarman's rule (*c.* 1215, possibly as late as 1219). He was the last of the great kings. Yet even if the 13th and 14th centuries were centuries of decline, they were by no means a period of eclipse, despite the repeated incursions of the Thais, which forced a temporary abandonment of Angkor before the end of the 14th century.

The final decline. After Jayavarman's death, no significant building took place at Angkor, and the Khmer empire began to contract. Control over Champa ended in 1220. In 1238 a Thai chieftain—son-in-law of Jayavarman VII—mounted a successful coup d'etat against the Khmer governor of the upper Chao Phraya region and founded the Thai state of Sukhothai. Khmer overlordship in the Malay Peninsula also disappeared after Jayavarman's death. For the greater part of the 13th century, however, Angkor remained a glittering and wealthy city, one which inspired wonder in the eyes of the Chinese traveller Chou Ta-kuan when he visited the city in 1296. The account that Chou Ta-kuan provides of Angkor at this time is both the longest and most detailed description of the Khmer capital, and despite its inaccuracies and reliance on legend it gives a remarkably vivid picture of the rich court society that continued to exist, even though the Chinese writer records the fact that Cambodia had been under attack from the Thais shortly before. Chou Ta-kuan left a picture of a bustling city in which the king still went forth in great pomp and ceremony. Whatever weakness now characterized the Khmers, Chou Ta-kuan still judged Cambodia to be the strongest of the states to the south of China. Notable among his observations was the presence in the Khmer capital of Buddhist monks of the Theravāda school. Thus, by the end of the 13th century the Cambodian state had abandoned its adherence to Hinduism and, at the court and in the capital at least, adopted Theravāda Buddhism.

Theravāda Buddhism

This change took place during the long reign of Jayavarman VIII (reigned 1243–95), probably about the middle of the century. Its causes remain a matter for some debate. The view of some of the earliest scholars of Cambodian history, that Theravāda Buddhism was a "democratic" religion that appealed to the Cambodian masses bowed down before the overwhelming exactions of their kings, has been severely qualified. The adoption and propagation of a particular religion within Angkorian society is much more likely to have been a matter for royal patronage. One intriguing suggestion is that the rulers and their advisers at Angkor were aware of the political success of the increasingly powerful Thai states to the west and attributed this to the form of Buddhism, the Theravāda school, which the Thai rulers followed. Adherence to Theravāda Buddhism was never incompatible with the pursuit of power for the Thai rulers and would not have been viewed so by those in power at Angkor. Moreover, placing importance on the population's acceptance of Buddhism seems illogical in view of the fact that their most fundamental religious adherence was to animistic cults. Viewed from a slightly different perspective, however, there may be value in considering the permeation of Theravāda Buddhism into the ranks of the general population as reflecting some readiness on the part of the people to accept a new religion that made few demands upon them—in contrast to Hinduism and Mahāyāna Buddhism, which had sapped their material and physical resources. Not only did the introduction of the new religion mean the end of temple building on the old scale, it also led to the disappearance of the use of Sanskrit; the latest Cambodian Sanskrit inscription dates from 1327. Henceforth Pāli was the sacred language.

For a short period at the beginning of the 14th century the Khmers enjoyed a respite from external threat. The great Thai ruler of Sukhothai, Ramkhamhaeng, died *c.* 1317 and his death removed the motive force that had brought Sukhothai to a position of considerable power, posing a threat to Angkor. The respite, however, was a brief one. Whereas the selection of the Angkorian region as a site for the Khmer capital in the 9th century had ensured its isolation from external threats, the ever-growing presence of the Thais in the Chao Phraya Valley meant that Angkor was in an increasingly exposed position. The king at Angkor when Chou Ta-kuan visited the city in 1296 was Indravarman III (1296–1308). His successors, the last to leave inscriptions, were Indrajayavarman (1308–27) and Jayavarman Paramesvara, who acceded to the throne in 1327 but for whom no date of death is recorded. With an absence of inscriptions and the unreliability of other records, there has been a measure of historical disagreement over developments during the 14th and 15th centuries. Recent research has challenged the previous assumption that the Thais captured Angkor in 1431, from which date the Khmers abandoned their capital. According to this new view, based on both Cambodian and Chinese sources, Angkor fell to Thai attacks in 1369 and 1389 and was on each occasion occupied by Thai forces. It was not, however, until 1444 that the final (third) great attack against Angkor took place that did, indeed, lead to its ultimate abandonment.

The fall of Angkor

The proposed new chronology, which is accompanied by evidence of considerable Thai involvement in the family rivalries of the Khmer royal house, is helpful for an assessment of why it was that Angkor was finally abandoned in such a dramatic fashion. Emphasis on the political significance of the change to Theravāda Buddhism remains important, with its overtones of a sense of crisis in Khmer national confidence. But other, more obvious, material factors were clearly involved. The record of Thai attacks against Angkor highlights the extreme fragility of the Angkorian economic system. Under pressure from the Thais, and on occasion from the Chams also, the Khmer rulers at Angkor had the greatest difficulty in maintaining the intricate hydraulic system upon which Angkor's survival depended. The reservoirs and canals needed constant attention, which was impossible under conditions of war. Damage to the system during actual attacks on the city, and the shipping off of thousands of prisoners of war, further weakened the economic base of the state. Recent analysis now discounts the view that malaria, spreading from the shattered hydraulic system, played a part in Angkor's decline. When, in a manner that became so characteristic in the years of decline after the fall of Angkor, squabbles between members of the royal family further weakened the Khmers, the stage was set for the collapse of the Angkorian empire. After the abandonment of the former imperial city in 1444, the Khmer court moved south to a site near modern Phnom Penh, and thence to Phnom Penh itself. For the remainder of Cambodian history to the present, this region was to be the centre of a contracting Khmer world.

The artistic traditions of Angkor did not immediately disappear. The expression of these traditions was no longer the mighty temples in stone but rather the more intimate art of the religious statue, and even this art form declined as political power waned more rapidly during the 17th and 18th centuries before the ever greater power of the Thais and the Vietnamese. When a French colonial presence was established in Cambodia shortly after the middle of the 19th century, it was common for the European newcomers to speak of Cambodia as having "fallen from its antique splendour." This comment accurately assessed the manner in which the once singularly powerful Khmer empire had, in the years after the fall of Angkor, gradually ceased to play any major role in the Indochina region.

THAI AND VIETNAMESE HEGEMONY

In the century and a half following the abandonment of Angkor, much of Khmer history is a confusing mixture of uncertain dates and complex dynastic rivalries within the Khmer royal family. Dominating the years up to the Thai conquest of Lovek (a later Khmer capital a little to the north of Phnom Penh) in 1594 is the record of almost incessant warfare between the Khmers and the Thais, now based in their capital of Ayutthaya. The contest was not entirely uneven. During the reign of Ang Chan (1516–66) the Khmers carried their campaign deep into Thai territory on one occasion, in 1564, but reached the city of Ayutthaya only to find it occupied by Burmese invaders. Yet despite the temporary success of his son Barom Reachea I (1566–76), which enabled the Cambodian court to reoccupy the city at Angkor temporarily, the tide was flowing against the Khmers. Once the Thais had successfully resisted the challenge posed by repeated Burmese invasions, they were able under the leadership of King Naresuan to plan and deliver a crushing blow against the Cambodian state. This was the capture of the capital at Lovek in 1594.

The "Spanish interlude." Before Lovek fell to the Thais the Cambodian king Satha (1576–*c.* 1594) had sought to gain assistance from the Portuguese and Spaniards, who were now established in Southeast Asia. Satha's efforts to enlist their support and the ambitions of both missionaries and soldiers of fortune led to one of the more curious episodes in Cambodian history. For a brief period, after the fall of Lovek, Spanish soldiers became arbiters of power within the Khmer state. A Spanish expedition arrived in Cambodia in 1596 with the intention of giving aid to King Satha. By then, however, Satha had been deposed and a usurper, Chung Prei, occupied the throne. After a series of disagreements and the sacking of the Chinese quarter of Phnom Penh by the Spanish forces, the Spaniards attacked the king in his palace at Srei Sânthôr, killing him and his son. Deciding to return to Manila after this affair, the Spaniards later changed their minds while sailing along the Vietnamese coast and marched overland to Laos, where they found one of Satha's sons and successfully installed him upon the Cambodian throne in 1597 as Barom Reachea II (1597–99). The great officials within the Cambodian court were by this time deeply resentful of the power wielded by the foreigners and were anxious to depose Barom Reachea in favour of Satha's brother, Soryopor. An incident involving one of the Spanish leaders in Phnom Penh led to a massacre of the Spanish garrison in mid-1599 and, after a period in which three princes occupied the throne, Soryopor returned to gain the throne as Barom Reachea IV (1603–18), with the assistance of a Thai army. The Spanish interlude was the last significant contact the Cambodian court had with Europeans until the onset of French colonialism in the mid-19th century. It was an exotic interregnum that had no lasting political effect but did, however, provide later historians with one of the few accounts of the Angkorian irrigation system, written by an observer who visited Angkor at a time when it had been reoccupied toward the end of the 16th century.

irst
uropean
ontact

Cambodia as a vassal state. Cambodian history from the accession of Barom Reachea IV until the establishment of the French Protectorate in 1863 is a sorry record of weak kings, almost always under challenge from dissident members of the royal family and forced to seek the protection of either of their stronger neighbours, Siam (Thailand) and Vietnam. Between 1603 and 1848, the date of the last Cambodian ruler to assume the throne free from French political control, no fewer than 22 monarchs uneasily occupied the throne. Several of these rulers had more than one reign, as they gave up their position either through choice or under duress, only to return when their successors either proved incapable or were themselves deposed. The details of this unstable period are less important than the record that is available of the manner in which Cambodia slowly slipped under the dual suzerainty of Siam and Vietnam. Early in the 17th century, Chey Chetta II (1618–28) successfully, if rather temporarily, declared his country's independence from Siam. In order to strengthen his position, the king sought assistance from the Nguyen lords of southern Vietnam. This fatal decision, which had involved Chey Chetta's marrying a Vietnamese princess, brought a Vietnamese demand for the right to settle Vietnamese colonists in areas of modern southern Vietnam, near Saigon, which until this time had been Cambodian territory. From this point onward, rival claimants for the Khmer throne sought to advance their interests through the respective support of either the Thais or the Vietnamese, at the ever-increasing cost of loss of territory or loss of power. For the Vietnamese, assistance given to Cambodian princes was followed by the extension of Vietnamese settlement into former Cambodian territories. For the Thais, the provision of aid was joined to an insistence upon Cambodia's assuming a vassal status, but it did not involve the colonization of Cambodian lands.

Vietnamese
expansion

That Cambodia survived at all during this bitter period was attributable in large part to the other preoccupations that weighed upon the Thais and the Vietnamese. In the middle years of the 18th century, Thai energies were largely taken up with wars against the Burmese, which led, in 1767, to the Burmese capture of Ayutthaya. Later in the same century, the Nguyen rulers of southern Vietnam were engaged in a prolonged campaign to regain power from the usurping Tay Son rebels. Yet if these facts had slowed the pace of Cambodia's decline, the situation by the end of the 18th century was nonetheless grim. Over the course of the century, all of what was to become the French colony of Cochinchina had fallen under Vietnamese control. In the west, Thai suzerainty had been firmly established over the great provinces of Bătdâmbâng and Siĕmréab. As the century drew to a close, and as the Vietnamese leader Nguyen Anh, later to reign as Gia Long, still fought to unify his country, the Thai court at Bangkok was firmly in control of the Cambodian royal family. King Ang Eng (reigned 1779–96) was actually crowned in Bangkok and was placed on the throne at the Khmer capital, Oudong, north of Phnom Penh, through the support of a Thai army. Following Ang Eng's death, Ang Chan, a minor, succeeded to the throne. He was crowned in 1806 and ruled uncertainly until 1835.

Ang
Chan's
reign

The reign of Ang Chan II confirmed Cambodia's dual vassalage to Siam and Vietnam. Faced with his rebellious brothers, Ang Chan gained assistance from Vietnam, while the Thais supported his brothers. The uneasy calm that was finally established, as Cambodia acknowledged both Thai and Vietnamese suzerainty, ended with Ang Chan's death. Vietnamese pressure was strong enough to ensure that a weak and powerless princess was then placed on the throne while almost total Vietnamese control was exercised over the country. Not until 1841, when Ang Duong (reigned 1848–60) returned from exile in Bangkok supported by Thai troops, were the Cambodians able to exercise even a minimal degree of independence. Because Thai support had helped him to the throne, Siam exercised a preponderant influence over Ang Duong, but he remained in tributary relations with Vietnam. As the last king to rule free from French control, Ang Duong has often been idealized in modern Cambodia. He was devoted to revitalizing the state, but his resources were desperately limited. Cambodia survived essentially because of the restraint of its neighbours, and before the end of his reign, Ang Duong's rule was troubled by internal rebellions.

FRENCH RULE

The protectorate. French control over Cambodia developed as an adjunct to French colonial involvement in Vietnam. Before his death, Ang Duong, on the urging of a French missionary, had sought to gain French assistance to regain former Cambodian lands held by the Vietnamese. He did not seek a French protectorate over his country, as many French accounts have suggested. There was no response to Ang Duong's call for help, and the French decision to advance into Cambodia came only when, with a colony implanted in Cochinchina, the French began to fear that British and Thai expansion might lead to hostile control of the Mekong River. In an almost classic use of gunboat diplomacy, the French in 1863 intimidated the Cambodian king Norodom (acceded 1860) into signing a protectorate treaty that gave France control of Cambodia's foreign affairs. When Norodom was crowned at Oudong in 1864, the French stood by as new suzerains of the weak Cambodian state.

Reign of Norodom

In the early years of the French protectorate, the European newcomers interfered little in the affairs of the Cambodian state. Indeed, French protection had distinct similarities to the part played by Cambodia's Asian suzerains in the past. At odds with his half brothers and faced with rebellion, Norodom gained rather than lost as the result of the French presence. Although he bitterly resented the French recognition of Thai control over Bătdâmbâng and Siĕmréab provinces, he might well have lost the throne if it had not been for French support. By the mid-1870s, however, French officials in Cambodia were pressing for greater control over internal affairs. Shocked by what they regarded as the profligacy and barbarity of Norodom's court and methods of government, they sought to introduce reforms. In pursuing their goals they acted with the knowledge that Norodom's half brother Prince Sisowath was ready to cooperate with them and was most ambitious to replace Norodom on the throne. Yet, despite the power that the French could exercise, their repeated efforts to change the methods of Cambodian administration foundered on the rock of Norodom's stubborn opposition to change. Exasperated by the king's intransigence, Gov. Charles Thomson of Cochinchina, who also held responsibility for the French position in Cambodia, forced Norodom in June 1884 to sign a convention virtually transforming Cambodia into a colony. The result of this action was a Cambodian rising that broke out in January 1885 and lasted for two years.

Anti-French rebellion

This rebellion was the only major anti-French movement to occur in Cambodia until after World War II. It gravely threatened the French position in the country and was finally put down only when French officials were able to persuade the King to call upon his countrymen to lay down their arms in return for some French concessions to the King's position. This was, basically, a hollow victory for Norodom. What the French had been unable to achieve in a single sweep by the convention of 1884 they proceeded to gain through piecemeal action. As Norodom's health declined and as senior Cambodian officials began to see their interests linked as much with French power as with their monarch, the way was opened for greater and greater French control over the actual business of government. In 1897 the French representative in Phnom Penh assumed the chairmanship of the Cambodian Council of Ministers, and the King's influence in government was reduced to a bare minimum. Norodom died a bitter man in 1904.

Yet Norodom's reign was not entirely one of constant loss for the Cambodian monarchy. Although the French gained political ascendancy, they ensured the survival of the Cambodian state, and, by maintaining the king in a splendour that had probably not been equalled since Angkorian times, they greatly enhanced his symbolic position within the kingdom. This fact was to be of considerable importance in the post-independence period following World War II.

King Norodom's ambitious half brother Sisowath finally did succeed to the Cambodian throne in 1904. Already in his 60s, he nonetheless continued to reign until 1927. While Sisowath held the throne, no difficulties arose between the Cambodian king and the French, nor was there any growth of nationalist activity comparable to that in Vietnam. By preserving the Cambodian monarchy, the French administration successfully prevented the development of any alternative focus for national identification. The one significant event that occurred during Sisowath's reign involved a mass peasant protest against tax and corvée (forced labour) requirements in 1916. Yet even this peasant protest was essentially peaceful and the protesters disbanded once the King had given his assurance that royal consideration would be given to ameliorating conditions. During Sisowath's reign, and that of his son and successor, Monivong (1927–41), the first notable economic developments derived from French rule took place. Red-earth lands in the east were brought under rubber cultivation, and, for the first time, there was a realization that Cambodia's economic opportunities went beyond the export of rice.

Rubber cultivatio

World War II and the First Indochina War. When Monivong died in 1941, the Japanese had occupied Indochina while leaving the French administration in nominal control. In these difficult circumstances, the French governor general of Indochina, Adm. Jean Decoux, placed Prince Norodom Sihanouk on the Cambodian throne. Although some uncertainties remain about the reasons for Sihanouk's elevation, Decoux seems to have been guided by the expectation that, as a young man of 18, Sihanouk, unlike the late King Monivong's son, Prince Monireth, would be readily controlled by the French. The error of the French estimation of Sihanouk's pliability became apparent in the years after the end of the war.

The effect of the Japanese occupation on Cambodia was less profound than it was in many other Southeast Asian countries. The overthrow of the French administration throughout Indochina in March 1945 brought the opportunity for political development to Cambodia. Although Sihanouk declared his country's independence from France, he and his advisers looked forward to cooperation with the former protecting power. Others, led by Son Ngoc Thanh, who had urged opposition to the French in the 1930s and been forced into exile in 1942, urged a sharp break with France.

The end of World War II brought the re-establishment of French control over Cambodia. Despite abolition of Cambodia's protectorate relationship, as it became an "autonomous state within the French Union," real power remained in French hands. Yet between 1945 and Cambodia's final achievement of independence in 1953, important political developments occurred. There were more demands for full independence, and a contest developed between the King and his supporters, who wished to uphold the supremacy of the crown, and those who sought to make Parliament the predominant power.

Re-establishment of French control

Cambodia was ill prepared for parliamentary democracy, and the years between 1945 and 1953 were marked by factional strife. The dominant Democratic Party was frequently at odds with the King and suffered from internal dissension. The death in 1947 of its able leader, Prince Youtevong, was a blow to the party and to those who hoped for a constitutional monarchy. Outside Parliament, Son Ngoc Thanh, now a bitter opponent of Sihanouk's policies, led the dissident Khmer Issarak (Independent Cambodians), who opposed both the King and the French. With this internal disunity and with the

French fighting the Vietnamese revolutionaries, the Viet Minh, throughout Indochina, Sihanouk acted. In January 1953 he dissolved Parliament, declared martial law, and left Cambodia to seek the support of world opinion for his country's independence.

INDEPENDENCE

Sihanouk finally succeeded in his personal campaign, known as the Royal Crusade, when the French granted independence on November 9, 1953. At the Geneva Conference in May of the following year, Sihanouk's representatives achieved a further triumph when his government was recognized as the sole legitimate authority within the country. This decision prevented the Viet Minh forces and the limited number of left-wing Khmers who were their allies from gaining any regional authority within the country, as happened in Laos. The King's achievements left him in a momentarily unchallengeable position; even Son Ngoc Thanh appeared ready to cooperate with Sihanouk.

Independence and the end of the First Indochina War did not, however, bring an end to political factionalism. Finding himself once again in conflict with his domestic opponents, Sihanouk took a dramatic new step. In March 1955 he abdicated in favour of his father, Norodom Suramarit, and formed a new mass political movement, the Sangkum Reastr Niyum (People's Socialist Community). Loyalty to Sihanouk and the country were the guiding principles of the Sangkum. In the elections of September 1955 it won every seat in the National Assembly. From then until his fall in March 1970, Sihanouk was the central figure in Cambodian politics, sometimes as prime minister and, after his father's death in 1960, when no new monarch was named, as chief of state. Overt political life was carefully controlled by the Prince and his advisers.

ounding

ingkum

Before the mid-1960s, external crises posed the greatest threat to Cambodia's stability. Conscious of historical precedent, Sihanouk saw South Vietnam and Thailand as the greatest threats to Cambodia's survival, particularly as these two neighbouring states were allied with the United States, which Sihanouk distrusted. At the same time Sihanouk feared the eventual success of the Vietnamese Communists and the danger he saw in a unified Vietnam. In these circumstances he proclaimed a policy of neutrality in international affairs. Convinced, however, of U.S. involvement in plots against his state and his family in 1959 and 1960, Sihanouk finally broke relations with Washington in May 1965 and aligned Cambodia more closely with China, North Vietnam, and the National Liberation Front in South Vietnam. From 1965, in an attempt to anticipate the future, he agreed clandestinely to the use of Cambodian territory by the Vietnamese Communists.

This decision further alienated right-wing elements in Cambodian society, particularly the army officer corps. The conservative urban elite already resented the economic policies introduced in 1963 and 1964, and they saw the closer alignment with the Vietnamese Communists as potentially ruinous to their interests.

From 1966 Cambodia's internal politics developed in a complex fashion. Although Prince Sihanouk remained a revered, even semi-divine figure for most of the peasantry, he was increasingly seen as a threat by both the right and the left. Young left-wing Cambodians, many of them educated abroad, resented Sihanouk's internal policies, which did not tolerate radical dissent. The outbreak of a rural rebellion in 1967 convinced Sihanouk that the greatest danger to his regime came from the left. He increasingly bowed to pressure from the army and other conservatives to follow a policy of harsh repression against left-wing elements.

By 1969 Sihanouk's personal position had deteriorated, and conflict between the army and left-wing opponents of the Prince's regime had increased. Right-wing politicians, most notably Gen. Lon Nol and Prince Sisowath Sirik Matak, plotted to depose Sihanouk, whom they saw as the root cause of their country's problems. They enlisted the support of Sihanouk's long-time enemy Son Ngoc Thanh, who had remained a dissident throughout the 1960s at the head of the Khmer Serei (Free Cambodians), a group receiving clandestine assistance from the United States. The extent of U.S. involvement in this plotting against Sihanouk is a matter of controversy. There seems little doubt, however, that U.S. intelligence services were aware of developments.

Matters came to a head in March 1970 when, with Sihanouk absent in France, Lon Nol and Sirik Matak called for the evacuation of all Vietnamese Communist forces from Cambodian soil and instigated demonstrations against the North Vietnamese and National Liberation Front embassies in Phnom Penh. At the same time, Sihanouk's conservative opponents called on him to renounce his international and domestic policies. Within a week it was clear that neither the Vietnamese Communists nor Sihanouk would negotiate under threat, and on March 18, Lon Nol's faction deposed the Prince. Sihanouk took up residence in Peking. From there he became the head of both a government in exile and a National United Front of Cambodia (NUFC), which was dedicated to overthrowing the Lon Nol regime.

Deposition of Sihanouk

Sihanouk's overthrow brought Cambodia's full involvement in the Second Indochina War, particularly after the U.S. and South Vietnamese invasion of the country in May 1970. At first the Lon Nol regime confronted a largely Vietnamese military threat, but a Khmer revolutionary force slowly emerged as the principal opponent. The leaders of this newly important group included men such as Khieu Samphan, Hou Yuon, and Hu Nim, who had opposed Sihanouk's policies in the past and who had fled into the countryside before his overthrow to escape the summary justice of the Prince's security forces.

As these and other left-wing leaders emerged into prominence, Sihanouk's importance declined. Although he remained a symbol of the legitimacy of those who opposed the Lon Nol regime, Sihanouk was clearly not in charge of either the daily tactics or the long-term strategy of those fighting against the Phnom Penh forces. By 1973 he spoke publicly of his difficulties with the Cambodian revolutionaries.

Despite proclaimed policies of reform and United States military and economic aid, the Lon Nol regime never succeeded in gaining the initiative against its opponents. Gestures such as declaring Cambodia the Khmer Republic in 1970 were of little account, as the revolutionary forces pursued a strategy of attrition, choking off provincial capitals from the countryside. Following the end of the U.S. bombing of Cambodia in August 1973, the balance tilted toward the eventual success of the NUFC forces despite continuing strong U.S. aid to the Lon Nol regime. This success was finally achieved when Phnom Penh fell to the revolutionary forces on April 17, 1975.

Great secrecy surrounded developments after April 1975. Sihanouk returned to Phnom Penh in September 1975 as titular chief of state but resigned three months after a new constitution of Democratic Kampuchea was instituted on January 5, 1976. Amid reports by refugees of mass executions and death from disease, major political and social changes took place. Phnom Penh was forcibly evacuated and programs were instituted to reorient the country toward agricultural self-sufficiency and to remove the economic and social patterns of the prerevolutionary period. Elections for the People's Representative Assembly were held in March 1976, and in April Khieu Samphan was named head of state and Pol Pot was named premier. In 1977 the Communist Party of Kampuchea was first officially recognized as the country's governing body.

Pol Pot as premier

(M.E.O.)

The regime of Pol Pot and Khieu Samphan gained a reputation of being one of the cruelest in modern history, with reports of arrests and mass executions common. In the meantime, the border clashes between Kampuchea and Vietnam escalated throughout 1977, culminating in Cambodia breaking relations with Vietnam on December 31. The differences between Cambodia and Vietnam were ideological as well as territorial. Vietnam, always suspicious of its northern neighbour, China, was antagonized by Cambodia's pro-Chinese policies. Pol Pot's strengthening of ties with China led to an escalation of successful Vietnamese attacks. Not only was the Cambodian government weakened by warfare with its neighbour but the internal disruption of the economy and agricultural system led to massive famines. It was estimated that at least 1,000,000 people died in 1975–78 from starvation or disease or in government purges.

In January 1979 guerrillas backed by Vietnam seized Phnom Penh and set up a government headed by Heng Samrin. Vietnam's occupation of Cambodia did nothing to promote stability or stop the bloodshed. Both Khieu Samphan and Pol Pot managed to escape to China, where they organized resistance to the new government. Sihanouk, who had been under house arrest during the Pol Pot regime, also went to Peking as well as to North Korea to ask for aid against the Vietnamese.

There ensued a protracted civil war resulting in many casualties and a large number of refugees, many of whom fled to an already beleaguered Thailand. Resistance to the Vietnamese-backed government consisted of three groups: the Khmer Rouge, headed by Pol Pot and Khieu Samphan; the staunchly anti-Communist Khmer People's National Liberation Front (KPNLF) headed by Son Sann, a former Kampuchean premier; and the supporters of Sihanouk, later called the Armée Nationale Sihanoukist.

Although the guerrillas under Pol Pot met with only limited military success, they fared better on the diplomatic front, with the UN refusing to recognize any Kampuchean government but Democratic Kampuchea. The Khmer Rouge won strong support from China, but most nations, even those ideologically sympathetic, were repelled by the Pol Pot regime's reputation for butchery during its years in power, and many chose to support the KPNLF or the more moderate resistance group headed by Sihanouk. Despite some success at unified action, these groups continued to fight each other as well as the Vietnamese occupiers. (Ed.)

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

LAOS

Laos (in full Lao People's Democratic Republic; Lao: Sathalanalat Paxathipatai Paxaxôn Lao) is a small, landlocked, underdeveloped country on the Indochinese Peninsula. It is bounded on the north by China, on the northeast and east by Vietnam, on the south by Kampuchea, and on the west by Thailand and Myanmar. Following the contour of the Indochinese Peninsula, Laos extends more than 600 miles (966 kilometres) from northwest to southeast, and it has a total area of approximately 91,400 square miles (236,800 square kilometres). The capital is Vientiane.

In the 1950s Laos became a focus of struggle among the major powers of the world. Despite international agreements to neutralize and isolate the country from external threat and to support its unity and development, Laos continued until 1975 to be the scene of others' battles, and then became the object of political competition between Vietnam and the Soviet Union, on one hand, and the People's Republic of China, on the other, while its own people remained divided and in conflict.

Located in an area of strategic importance among contending powerful neighbours, Laos possesses neither the natural barriers nor a population large enough to withstand foreign encroachment. The majority of its population live in the lowlands and in the valley of the Mekong River in the western and central part of the country. The remainder live in small groups in the eastern hill areas. The Laotian people are divided by differences of religion, language, and custom. Another source of disunity is the fact that, until the 19th century, Thailand, Myanmar, and Vietnam either held territory that at present is a part of Laos or exercised suzerainty and indirect rule over parts of the land and its people. Although the peoples of Laos can look back to a golden age when the country was strong and united under its own kings, since the 18th century it has been divided and weak.

Rise of Pathet Lao

Laos gained legal unity and the formal status of a nation on October 22, 1953, when France recognized the Kingdom of Laos as fully independent and sovereign; in December 1955 it was admitted to the United Nations. In reality, however, it was at best a divided state. In 1953 the Communist Pathet Lao, a dissident group supported by the Viet Minh military forces of Vietnam, established itself in the northeast and claimed to be the legal government. The Geneva Conference of 1954, while recognizing the unity of Laos and the legality of the kingdom, permitted the Pathet Lao forces to occupy the provinces of Phôngsali and Houaphan (Sam Neua), "pending a political settlement." No real settlement occurred, despite efforts by the leaders of Laos to integrate the dissidents, and internal warfare continued until 1975 as foreign forces supported rival Laotian factions and used Laos as a battlefield in the Vietnam war. Intra-Laotian strife became intertwined with the larger conflict. In 1975 the Pathet Lao gained control of the country, and the monarchy was ended and the Lao People's Democratic Republic inaugurated in December.

Physical and human geography

THE LAND

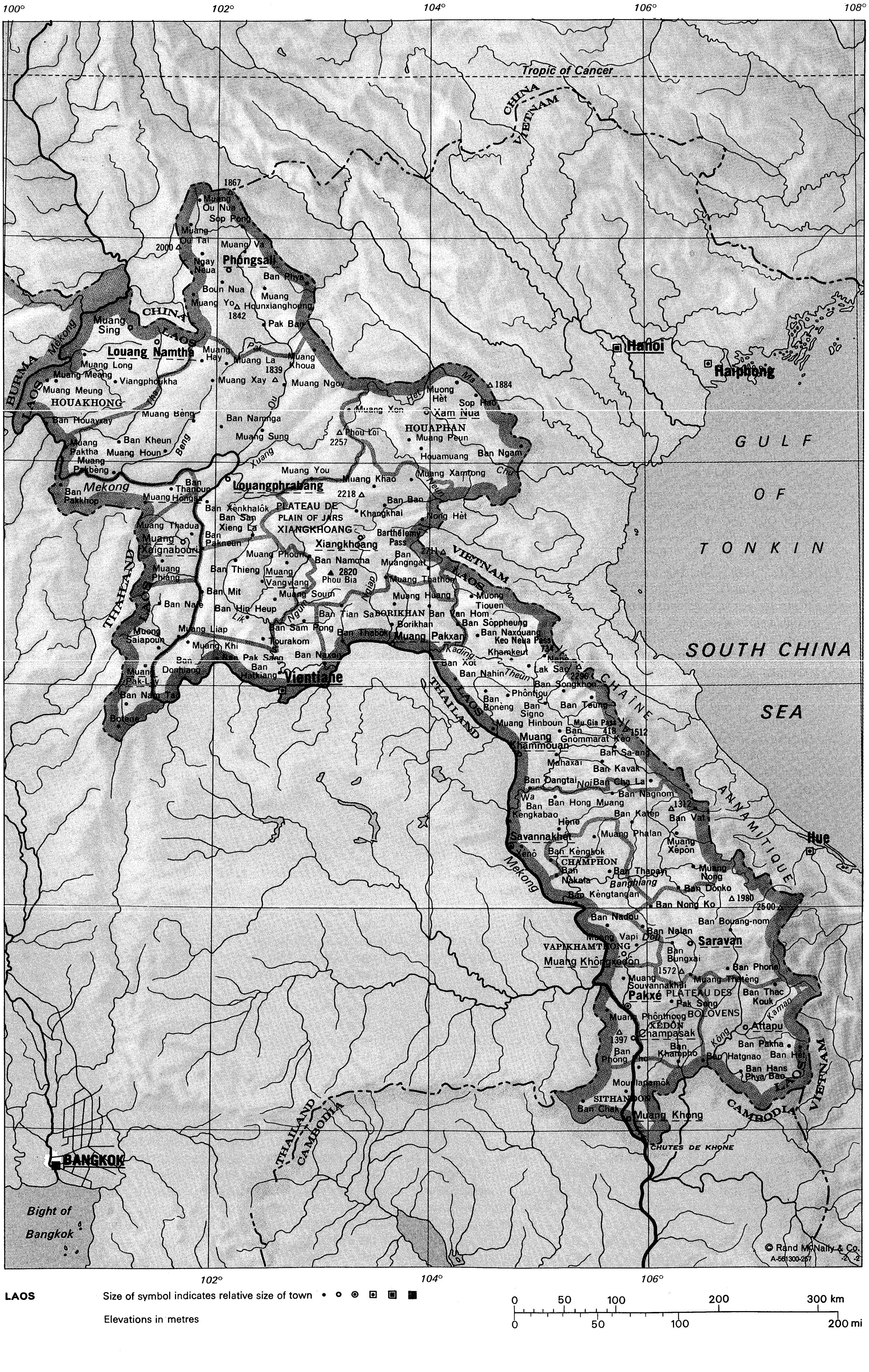
Relief. Dominating the landscape of Laos are its inhospitable, forest-covered mountains, which in the north rise to a maximum of 9,248 feet (2,819 metres) above sea level. The principal range lies along a northwest–southeast axis, but in the north there are lower, secondary ranges along a northeast–southwest axis. The ranges are broken by narrow, deep valleys through which rivers flow. A vital area strategically and politically is the Plain of Jars (a name derived from large prehistoric stone jars discovered there; Laotian Thông Haihin), a part of the Plateau de Xiangkhoang that often changed hands during the Vietnam war. In the central and northern parts of the country, the mountains form part of the Chaîne Annamitique. The only lowlands lie alongside the eastern bank of the Mekong. The general slope of the land is upward to the east.

Drainage. The rivers on the eastern mountain slopes flow toward Vietnam and the South China Sea; the rivers on the western side drain into the Mekong, which flows in a southerly direction from the province of Yunnan in China to the South China Sea, forming Laos' boundaries with Myanmar and (for the most part) with Thailand. The floodplains of the lowlands are formed from the alluvial soils carried by the river and its tributaries. The only other important fertile area is the Plateau des Bolovens, at an elevation of about 3,500 feet. The hills and mountainsides can be cultivated temporarily by the slash-and-burn method, but they quickly lose their fertility and the cultivators must move to other areas.

Climate. The climate in Laos reflects its location between the latitudes of 14° and 22° N. During the rainy season, May to October, the winds blow from the southwest and deposit an average rainfall of between 50 and 90 inches (1,300 and 2,300 millimetres). In the Plateau des Bolovens precipitation reaches 160 inches (4,100 millimetres) per year. The dry season, from November to April, is dominated by northeast winds. Temperatures average between 60° and 70° F (16° and 21° C) in the cool months of December through February, increasing to more than 90° F (32° C) in March and April, just before the rains. In the wet season, the average temperature is 80° F (27° C).

Plant life. Laos has tropical rain forests of broadleaf evergreens in the north and monsoon forests of mixed evergreens and deciduous trees in the south. In the monsoon forest areas the ground is covered with tall, coarse grass called *tranh;* the trees are mostly second growth, with an abundance of bamboo, scrub, and wild banana.

Settlement patterns. More than half the population is concentrated in the lowlands, where most are engaged in wet-rice farming. The isolated valley communities preserve different traditions and different dialects. Village populations range between 50 and 2,000, usually located close to rivers and roads that give the people access to itinerant


100°
102°
104°
106°
108°
22°
20°
18°
16°
14°
Tropic of Cancer
CHINA
VIETNAM
BURMA
LAOS
CHINA
LAOS
THAILAND
LAOS
VIETNAM
LAOS
THAILAND
LAOS
THAILAND
CAMBODIA
CAMBODIA
LAOS
VIETNAM
Mekong
Ma
Nam
Hanoi
Haiphong
GULF OF TONKIN
SOUTH CHINA SEA
CHAÎNE ANNAMITIQUE
Hue
BANGKOK
Bight of Bangkok
CHUTES DE KHONE
1867
Muang Ou Nua
Sop Pong
Muang Ou Tai
2000
Muang Va
Ngay Neua
Phongsali
Ban Phya
Boun Nua
Muang Yo
1842
Muang Hounxianghoung
Pak Bay
Muang Sing
Louang Namtha
Muang Hay
Muang La
1839
Muang Khoua
Muang Long
Muang Meang
Viangphoukha
Muang Meung
Muang Xay
Muang Ngoy
HOUAKHONG
Muang Beng
Ban Houayxay
Ban Namnga
Muang Sung
Muang Pakbeng
Muang Paktha
Ban Kheun
Muang Houn
Beng
Ou
Xuang
Ban Pakkhop
Ban Thanoun
Muang Hongsa
Louangphrabang
Muang You
Muang Khao
2218
Ban Ban
Ban Xenkhalok
Ban San Xieng La
PLATEAU DE PLAIN OF JARS
XIANGKHOANG
Khangkhai
Nong Het
Muang Thadua
Muang Xaignabouri
Ban Pakneun
Xiangkhoang
Barthélemy Pass
2711
Muang Phoun
Ban Namoha
Ban Muangngat
Muang Phiang
Ban Thieng
Muang Vangviang
2820
Phou Bia
Muang Thathom
Ban Mit
Muang Soum
Muang Huang
Muang Tiouen
Ban Nale
Ban Hin Heup
Lik
Ngum
Ban Tian Sa
BORIKHAN
Ban Van Hom
Ban Soppheung
Muang Saiapoun
Muang Liap
Ban Sam Pong
Ban Thabok
Borikhan
Muang Pakxan
Ban Naxouang
Keo Neua Pass
734
Muang Khi
Tourakom
Ban Naxai
Kading
Khamkeut
Nape
Ban Pak Sang
Ban Hatkiang
Vientiane
Ban Xot
Ban Nahin
Lak Sao
2286
Muang Pak-Lay
Ban Donthuang
Ban Nam Tai
Botene
Theun
Ban Songkhon
Ban Boneng
Phôntiou
Ban Signo
Ban Teung
Muang Hinboun
Mu Gia Pass
418
1512
Muang Khammouan
Ban Gnommarat
Keo
Ban Sa-ang
Mahaxai
Ban Kavak
Ban Dangtai
Noi
Ban Cha La
Wa
Ban Nagnom
Ban Kengkabao
Ban Hong Muang
1312
Hene
Ban Katep
Ban Vat
Savannakhet
Muang Phalan
Muang Xepon
Xenô
Ban Kengkok
CHAMPHON
Ban Nakala
Ban Thapay
Banghiang
Muang Nong
Ban Donko
Ban Kengtangan
Ban Nong Ko
1980
2500
Ban Nadou
Ban Bouang-nom
Muang Vapi
Don
Ban Nalan
Saravan
VAPIKHAMTHONG
Muang Khongxedon
Ban Bungxai
1572
Ban Phone
Muang Souvannakhili
Muang Thateng
Pakxé
PLATEAU DES BOLOVENS
Ban Thac Kouk
Pak Song
Muang Phônthong
XEDON
Kaman
Champasak
Attapu
1397
Kong
Ban Phong
Ban Khampho
Ban Pakha
Ban Hatgnao
Ban Hê
Ban Hans Phya Bao
Moulapamok
SITHANDON
Ban Chak
Muang Khong
Mekong
1884
Muong Hèt
Sop Hao
Muang Xon
Hêt
Xam Nua
HOUAPHAN
Muang Peun
Phou Loi
2257
Houamuang
Ban Ngam
Chu
Muang Xamtong
© Rand McNally & Co.
A-561300-257
LAOS
Size of symbol indicates relative size of town • ○ ◎ ▣ ▣ ■
Elevations in metres
0 50 100 200 300 km
0 50 100 200 mi

MAP INDEX

Political subdivisions

Attapu	14·45n	107·00e
Borikhan	18·30n	103·40e
Champasak	14·45n	105·45e
Champhon	16·25n	105·25e
Hôngsa	19·40n	101·10e
Houakhong	20·45n	101·00e
Houaphan	20·20n	104·00e
Khammouan	18·00n	105·00e
Louangphrabang	20·20n	102·30e
Pak-Lay	18·20n	101·15e
Phôngsali	22·00n	102·00e
Saravan	15·40n	106·45e
Savannakhét	16·40n	105·00e
Sithandon	14·15n	105·50e
Vangviang	19·00n	102·35e
Vapikhamthong	15·40n	105·50e
Vientiane	18·10n	102·35e
Xaignabouri	19·20n	101·30e
Xédôn	15·00n	106·10e
Xiangkhoang	19·15n	103·30e

The name of a political subdivision if not shown on the map is the same as that of its capital city.

Cities and towns

Attapu	14·48n	106·50e

B. = Ban

B. Ban	19·38n	103·34e
B. Bonèng	17·58n	104·35e
B. Bouangnom	15·47n	106·47e
B. Bungxai	15·42n	106·14e
B. Chak	14·17n	105·25e
B. Cha La	17·11n	106·05e
B. Dang Tai	17·06n	104·57e
B. Dong Ko	16·12n	106·17e
B. Donhiang	18·05n	101·48e
B. Gnômmarat Kèo	17·36n	105·10e
B. Hans Phya Bao	14·31n	106·51e
B. Hatgnao	14·40n	106·35e
B. Hatkiang	18·11n	102·40e
B. Hèt	14·44n	107·29e
B. Hin Heup	18·38n	102·20e
B. Hong Muang	17·04n	105·12e
B. Houayxay	20·18n	100·26e
B. Katèp	16·48n	105·52e
B. Kavak	17·18n	105·37e
B. Kèngkabao	16·48n	104·45e
B. Kèngkok	16·26n	105·12e
B. Kèngtangan	16·05n	105·22e
B. Khampho	14·37n	106·17e
B. Kheun	20·13n	101·07e
B. Mit	18·51n	101·55e
B. Muangngat	19·05n	104·04e
B. Nadou	15·51n	105·38e
B. Nagnom	17·02n	105·44e
B. Nahin	18·14n	104·13e
B. Nakala	16·17n	105·11e
B. Nalan	15·50n	106·04e
B. Nalè	18·42n	101·34e
B. Namcha	19·09n	102·53e
B. Namnga	20·22n	102·19e
B. Nam Tao	17·50n	101·15e
B. Naxon	18·12n	103·05e
B. Ngam	20·11n	104·53e
B. Nong Ko	16·01n	106·04e
B. Pakha	14·38n	107·24e
B. Pakkhop	19·49n	100·36e
B. Pakneun	19·14n	101·50e
B. Pak Sang	18·13n	102·08e
B. Phone	15·25n	106·42e
B. Phông Pho	14·36n	105·52e
B. Phya	21·35n	102·55e
B. Sam Pong	18·32n	102·47e
B. San Xieng La	19·27n	102·26e
B. Signo	17·51n	105·04e
B. Songkhon	17·58n	105·10e
B. Sôppheung	18·33n	104·17e
B. Teung	17·54n	105·29e
B. Thaboc	18·22n	103·12e
B. Thanoun	19·50n	101·29e
B. Thapayi	16·19n	105·41e
B. Tian Sa	18·43n	103·14e
B. Van Hom	18·44n	104·01e
B. Vat	16·54n	106·25e
B. Xènkhalôk	19·42n	101·54e
B. Xot	18·11n	104·05e
Borikhan	18·33n	103·43e
Botene	17·35n	101·12e
Boun Nua	21·38n	101·54e
Champasak	14·53n	105·52e
Hène	16·42n	105·17e
Houamuang	20·09n	103·38e
Khamkeut	18·15n	104·43e
Khangkhai	19·28n	103·15e
Lak Sao	18·11n	104·58e
Louang Namtha	20·57n	101·25e
Louangphrabang	19·52n	102·08e
Mahaxai	17·25n	105·12e
Mounlapamôk	14·20n	105·52e

M. = Muang

M. Bèng	20·22n	101·44e
M. Hay	21·03n	101·49e
M. Hinboun	17·35n	104·36e
M. Hôngsa	19·43n	101·20e
M. Houn	20·09n	101·27e
M. Hounxianghoung	21·37n	102·18e
M. Huang	18·54n	103·42e
M. Khammouan	17·24n	104·48e
M. Khao	19·47n	103·29e
M. Khi	18·27n	101·46e
M. Không	14·07n	105·51e
M. Khôngxédôn	15·34n	105·49e
M. Khoua	21·05n	102·31e
M. La	20·52n	102·07e
M. Liap	18·29n	101·40e
M. Long	20·57n	100·48e
M. Meung	20·43n	100·28e
M. Ngoy	20·43n	102·41e
M. Nong	16·22n	106·30e
M. Ou Nua	22·18n	101·48e
M. Ou Tai	22·07n	101·48e
M. Pakbèng	19·54n	101·08e
M. Pak-Lay	18·12n	101·25e
M. Paktha	20·06n	100·36e
M. Pakxan	18·22n	103·39e
M. Peun	20·13n	103·52e
M. Phalan	16·39n	105·34e
M. Phiang	19·06n	101·32e
M. Phônthong	15·05n	105·39e
M. Phoun	19·07n	102·43e
M. Sing	21·11n	101·09e
M. Soum	18·46n	102·36e
M. Souvannakhili	15·23n	105·49e
M. Sung	20·19n	102·27e
M. Thadua	19·26n	101·50e
M. Thatèng	15·26n	106·23e
M. Thathôm	19·00n	103·36e
M. Va	21·53n	102·19e
M. Vangviang	18·56n	102·27e
M. Vapi	15·40n	105·55e
M. Xaignabouri	19·15n	101·45e
M. Xamtong	19·51n	103·51e
M. Xay	20·42n	101·59e
M. Xépôn	16·41n	106·14e
M. Xon	20·27n	103·19e
M. Yo	21·31n	101·51e
M. You	19·49n	102·50e
Muong Hèt	20·48n	104·03e
Muong Tiouen	18·51n	104·23e
Napè	18·18n	105·06e
Ngay Neua	21·50n	101·54e
Nong Hèt	19·29n	103·59e
Pak Ban	21·14n	102·28e
Pak Song	15·11n	106·13e
Pakxé	15·07n	105·47e
Phôngsali	21·41n	102·06e
Phôntiou	17·53n	104·37e
Saravan	15·43n	106·25e
Savannakhét	16·33n	104·45e
Sop Hao	20·33n	104·27e
Tourakom	18·26n	102·32e
Viangphoukha	20·41n	101·04e
Vientiane	17·58n	102·36e
Xam Nua	20·25n	104·02e
Xénô	16·35n	104·50e
Xiangkhoang	19·20n	103·22e

Physical features and points of interest

Annamitique, Chaine (Annamese Cordillera), *mountains*	17·00n	106·00e
Banghiang, *river*	16·03n	105·15e
Barthélemy Pass	19·26n	104·06e
Beng, *river*	19·53n	101·08e
Bia, Phou, *mountain*	18·59n	103·09e
Bolovens, Plateau des	15·20n	106·20e
Chu, *river*	19·53n	105·45e
Don, *river*	15·07n	105·48e
Het, *river*	20·49n	104·01e
Jars, Plain of	19·27n	103·10e
Kading, *river*	18·19n	104·00e
Kaman, *river*	14·48n	106·51e
Keo Neua Pass	18·23n	105·09e
Khone, Chutes de (Khone Falls), *rapids*	13·56n	105·56e
Kông, river	14·28n	106·20e
Lik, *river*	18·31n	102·31e
Loi, Phou, *mountain*	20·16n	103·12e
Mekong, *river*	13·55n	105·58e
Mu Gia Pass	17·40n	105·47e
Neun, *river*	19·42n	104·03e
Ngiap, *river*	18·24n	103·36e
Ngum, *river*	18·09n	103·06e
Noi, *river*	17·05n	105·02e
Ou, *river*	20·04n	102·13e
Pak, *river*	21·05n	102·31e
Tha, *river*	20·07n	100·36e
Theun, *river*	18·14n	104·38e
Wa, *river*	16·55n	104·48e
Xiangkhoang, Plateau de	19·30n	103·10e
Xuang, *river*	19·58n	102·15e

traders as well as to each other. Most villages are laid out around a main street or open area, farmlands being adjacent to the residential areas. Every village, if it can, has a pagoda and supports at least one Buddhist monk. The pagoda compound usually includes a public building that serves as a school and a meeting hall. Village leadership is usually divided, the headman having authority in secular matters and the monk in religious.

The hill peoples

The hill peoples are usually organized on tribal lines and live in smaller groupings. They are hunters and gatherers of forest products as well as farmers; their techniques of shifting cultivation prevent them from establishing permanent villages. Their political and social structures are varied. The tribal Tai have a stratified social structure and a political hierarchy, while the Mon-Khmer tend to have a simple political organization with a single headman. The Meo (Miao) tribesmen have a tradition of a king and chiefs; in theory their villages are tribal, but in fact they usually have a headman. Hill peoples living close to the Lao and Tai tend to acquire the languages and cultures of their neighbours and to engage in limited trade with them; those living at higher elevations remain unacculturated.

Urban life in Laos is limited to the capital, Vientiane, the former royal capital, Louangphrabang, and four or five large towns. With the exception of Louangphrabang, all are located near the Mekong in the floodplain area. Their populations are predominantly Lao, with many Chinese, Vietnamese, and Europeans. The Lao elite tends to be Westernized in its life-style and dress. By comparison with the cities of Thailand, Malaysia, or Vietnam, those of Laos are small and provincial.

THE PEOPLE

The peoples of Laos are divided by language, culture, and location. Lao officials distinguish four basic linguistic-political groups: Lao-Lu, or valley Lao; Lao-Tai, or tribal Tai; Lao-Theng, or Mon-Khmer; and Lao-Soung, or Meo and Man. Mountain people sometimes are called Kha ("Slave"), a pejorative term.

The Lao-Lu live in the lowlands, on the banks of the Mekong and its tributaries, and in the cities. They speak Laotian Tai, which is closer to the language spoken by Thais in Thailand than it is to the language of the local Tai-speaking tribes. It is the language that the minorities living near Lao areas are gradually adopting.

The Lao-Tai include such local groups as the Black Tai and Red Tai, both names referring to the dress of the women; the Tai Neua, or Tai of the north; the Tai Phuan; and the Phon Tai. The Lao-Tai live in all parts of the country, chiefly in upland areas. The various dialects of the Lao-Tai are mutually intelligible.

Mon-Khmer

The Lao-Theng, better known as Mon and Khmer, or Mon-Khmer, include many groups of people scattered throughout Laos, northeastern Myanmar, northern Thailand, and southern China and are thought to be descendants of the earliest populations in the region. These people do not form a single coherent group but rather include between 25 and 30 distinct groups, some closely related, others only tenuously identified as part of this linguistic group.

The Lao-Soung, which include the Meo (Miao) and the Man (Yao), are believed to have come from southern China since the late 18th century. They are divided into subgroups, and neither constitutes a large proportion of the population of Laos.

The other distinct linguistic groups are few. Tibeto-Burman-speakers, who came from southern China, live in the north and northwest. Chinese and Vietnamese live primarily in the urban areas. Initially, French was the language of the Lao elite and of the cities, but by the 1970s English had begun to displace it. Under the leadership of the Lao People's Revolutionary Party, Vietnamese has become the third language of the elite.

Prior to the emergence of the Lao People's Democratic Republic in 1975, it was accurate to say that the Lao peoples had a distinct pattern of culture and dress. They also had a well-defined social structure, differentiating between royalty and commoners. The members of the elite were said to number about 2,000, which included only a few outsiders who were not descendants of nobility. Most of the elite lived in the cities, drawing their incomes from rural land rents or from urban occupations. After 1975 a new elite emerged representing the victorious leftist forces. Many of this group were of aristocratic origin.

Traditionally the Lao-Tai, or tribal Tai, were organized in groups larger than villages, called *muong.* Each was ruled by a hereditary ruler called the *chao muong.* Within this broad grouping, however, there were ethnic variations. Among the Black Tai, the nobility consisted of two descent groups, the Lo and the Cam, who provided the rulers of the *muong.* The religious leaders came from two other descent groups, the Luong and the Ka. The Black Tai tribal organization had three levels: the village; the commune, which was composed of a number of villages; and the overall *muong.* The latter two were ruled by nobles, while the village headman was selected from among the commoners by the heads of households. The Red Tai had a similar social structure, with the addition of a council of five to aid the *chao muong.* The nobility owned the land and had the right of service from the commoners. Not all of them were Buddhists; those living in the higher and more isolated areas retained their traditional culture.

The Lao-Theng, or Mon-Khmer, have no political or social structure beyond the village. They are led by a village headman, who is their link to the central government; but his role in the village is not clear. The people traditionally have been spirit worshippers; one subgroup, the Lamets, practices ancestor worship. Some have adopted Buddhism.

Among the Lao-Soung, the Meo (also known as the Hmong) have a tradition of a king and subchiefs and a large-scale organization, although in practice this is usually limited to the village. The village consists of several extended families. In some villages, all the heads of households are members of a single clan, and the head of the clan is the headman of the village. Where several clans reside together in a large village there are several headmen, one being the nominal head and the link to the government. The headman has real authority in the village and is aided by a council. The Meo have extended their organization beyond the village for military purposes. They are spirit and ancestor worshippers.

The religion of most of the people of Laos is Theravāda Buddhism, professed by most Lao and by a small number of other ethnic groups. The rest of the people are animists or spirit worshippers. Many see no contradiction in being both, since Buddhism shows the way to Nirvāṇa while spirit worship helps a person to cope with daily and local problems. Among the hill peoples, especially those who have migrated from southern China, there are groups that mix Confucian ideas with Buddhism and animism. There were Catholic and Protestant missionaries in the country before 1975, but it is estimated that no more than 2 or 3 percent of the population is Christian. The Vietnamese, who live both in the cities and in the northeastern rural areas, practice a mixture of Mahāyāna Buddhism and Confucianism.

Laos remains an underpopulated country, with considerable migration taking place in and out of China, Vietnam, and Thailand.

THE ECONOMY

Resources. The natural resources of Laos include coal, iron, copper, lead, gold, tin, and precious stones. Only tin has been extracted on a commercial scale, though the others have been mined in primitive and unsystematic ways. Most of the tin ore is produced from the mines at Phôntiou and Nong Sun.

ce

Agriculture. The chief occupation of the people is agriculture, with an estimated 90 percent engaged in rice farming. Since the creation of the Lao People's Democratic Republic in 1975, natural calamities, including both drought and excessive rainfall, have prevented the people from achieving self-sufficiency in rice production, causing the nation to look to imports and gifts to make up for inadequate production. Modest amounts of corn (maize), sugarcane, tobacco, and cotton are produced. The forests provide teak and other woods, gum benzoin, cardamom, and stick lac. A major agricultural commodity that does not find its way into the statistics is opium, grown mainly by the Meo and traded illegally.

Nearly every household raises livestock—including cattle and buffalo—and poultry. Leather and hides are traded.

Under the revolutionary government, a serious effort is being made to relocate hill peoples onto the plains and to get them to adopt wet rice farming practices in place of their traditional slash and burn. It has been reported that some of the Meo have been forced to resettle in agricultural cooperatives in the region of the Plain of Jars. The majority of hill dwellers are resisting change, however, despite pressure and persuasion to alter their agricultural practices, social customs, and traditions. If the government persists in this policy it may provoke outright opposition to, and revolt against, the most serious challenge to Meo traditional life.

Industry. The tiny industrial sector manufactures bricks and ceramics, matches, cigarettes, soft drinks and beer, rubber sandals, cloth, pottery, and plastic bags. A major hydroelectric plant produces surplus electricity that is sold to Thailand.

Finance and trade. The chief exports are timber, tin, coffee, leather and hides, cardamom, gum benzoin, and stick lac. The major buyers formerly were Singapore, Malaysia, Thailand, the United States, and Hong Kong. The major suppliers of imports were Thailand, Japan, the United States, France, and the United Kingdom. New trading relationships are beginning to develop with Thailand and Vietnam and with Communist European states. Imports always vastly exceed exports in value.

River
traffic

Transportation. A major obstacle to the economic and social development of Laos is its transportation system. Rivers and roads are the major avenues of communication, supplemented by air transport. The Mekong River is the major north–south commercial artery; all but two sections of it—Chutes de Khone (Khone Falls) and the rapids of Khemmarat (Khemarat)—are navigable either all or part of the year. Large barges operated by Chinese ply the deeper sections of the rivers between towns, but most of the water traffic is carried in Lao-operated sampans and pirogues. The latter average 25 to 35 feet in length and can carry up to seven passengers and half a ton of cargo. Some of the mountain people build bamboo rafts and float their goods to market, selling the raft along with the goods.

During French rule, a primitive network of roads was created. The main artery joined Saigon with Louangphrabang, and several lesser roads led eastward through the four mountain passes to Vietnam and to the main towns and population centres. During the 1960s, with U.S. assistance, an all-weather road was added between Vientiane and Pakse; another was to be built between Vientiane and Louangphrabang. In addition, the North Vietnamese developed a complex of roads and trails across eastern Laos for their own use during the war. Laos itself has no railways, but a railroad from Bangkok, Thailand, to the Lao border serves as a major artery.

The newest means of transport is the airplane. The Wattay international airport at Vientiane connects with Bangkok, Phnom Penh and Hong Kong.

ADMINISTRATIVE AND SOCIAL CONDITIONS

Government. Before December 2, 1975, Laos was a constitutional monarchy. Real power rested with the prime minister and the legislature. There had been nominal political parties since 1947, when the constitution was promulgated, but they were loose coalitions of leaders with little popular following. In 1953 there were two main parties, the National Progressive and the Independent, and two splinter parties, the National Union and the Democratic. In 1958 the Lao Patriotic Front, backed by the Communist Pathet Lao, participated in the elections, as did the Peace Party; together they won several seats. In the late 1950s two new parties emerged, the Rally of the Lao People and the Committee for the Defense of the National Interest.

Beginning in 1973, when a cease-fire agreement was reached, a series of coalition governments sought to exercise power. The traditional elite continued to rule in urban areas, while the bulk of the people in the countryside remained outside the national political process. The coalition government gave way in December 1975 to a new body, the National Congress of People's Representatives, which, under the guidance of the Pathet Lao, met and passed several resolutions that ended the monarchy and replaced it with the Lao People's Democratic Republic. The King abdicated.

Supreme People's Council

The government is headed by a president, who is also chairman of the Supreme People's Council, an interim (pending elections) legislative body. The secretary general of the Lao People's Revolutionary Party (the only political party) serves as prime minister. The party, which replaced the Pathet Lao as leader of the Socialist revolution, was formerly called the People's Party of Laos and is the Communist core of the Lao Patriotic Front. It is organized much like other Communist parties, with a Central Committee headed by a Politburo.

The country is divided into 20 provinces, each subdivided into districts, which, in turn, are subdivided into towns and villages. At each level of local government there are "people's revolutionary committees," which receive directions from the Central Committee of the Lao People's Revolutionary Party.

Armed forces. Between independence, in 1953, and the cease-fire, in 1973, the military forces grew to become a major institution. In addition to the Royal Laotian Army, Royal Lao Air Force, Laotian River Flotilla, and Lao National Police Corps, there were the separate armies of the Pathet Lao and the neutralists. The government forces received aid from the United States. The Pathet Lao army was supported by North Vietnam and China.

After the victory of the Pathet Lao (officially Lao People's Liberation Army from 1965), the Royal Laotian Army was in part integrated with it and the rest was disbanded. There is a small navy and an air force.

Education. Education has been reorganized. The government has set up a number of agricultural schools, sent teachers to give literacy classes in provincial villages, and opened new primary, secondary, and teacher-training schools.

Health and welfare. The country has around 40 hospitals, supplemented by infirmaries and rural dispensaries. With the departure of the majority of medical doctors after 1975, the government began building village infirmaries in most of the provinces and training medical workers. Using medicinal herbs, these village medical workers provide most of the primary health care. Malaria and gastroenteritis are the major health problems.

CULTURAL LIFE

The basis of Laotian culture is religion and tradition. Art, literature, music, and drama draw mainly from these sources, and there appears to be little Western influence outside Vientiane.

Religious basis of cultural life

Theravāda Buddhism entered the country in the 14th century and along with Hinduism has been a major influence on cultural and intellectual life. The story of the Buddha and Hindu myths are the subjects of the carvings and sculptures found in all religious places. In the south, Khmer influences on the peoples of Laos are strong; in the north, Myanmar and Thai influences are readily apparent. As elsewhere in Southeast Asia, religious symbols, stories, and themes have been modified and localized. The snake, for example, representations of which adorn religious and royal buildings, represents the benevolent spirit of the water and the protector of the king.

The Laotians have a variety of folk arts, including weaving, basketmaking, wood and ivory carving, and silverwork and goldwork. There are a number of Laotian musical instruments, of which the *khene,* a bamboo wind instrument, is most widely known. Music is not written down but is played from memory.

Dancing is a profession rather than a form of recreation; the professional dance troupes travel throughout the country performing for religious celebrations or on important holidays. Their main themes are drawn from the Indian epics. All professional dancers are male, the female roles being performed by young men and boys.

Laotian literature is predominantly religious and linked to the Buddhist tradition. There is also a secular literary stream based on themes of the Hindu epic poems, which have been transmuted into popular language, as in the Laotian epic the *Sin Xay.* The popular poems and songs are often satirical. (J.Si.)

For statistical data on the land and people of Laos, see the *Britannica World Data* section in the BRITANNICA WORLD DATA ANNUAL.

History

The Lao people, the predominant ethnic group in present-day Laos, are a branch of the Tai peoples who by the 8th century AD established a powerful kingdom, Nanchao, in southwestern China. From Nanchao the Tai gradually penetrated southward into the Indochinese Peninsula; their migration was accelerated in the 13th century by the Mongol invasions of southern China by Kublai Khan. The Lao, together with other Tai peoples, gradually supplanted various primitive tribes (collectively known as Kha, or "Slaves") that from the 5th century on had lived in what is now Laos under the suzerainty of the Indianized Khmer Empire of Cambodia. During the 12th and 13th centuries they established the principality of Muong Swa (later Luang Prabang, now Louangphrabang), which was ruled by various Tai leaders and the history of which survives in Laotian legend and myth.

LAN XANG

The first Laotian state

Recorded Laotian history begins with Fa Ngum, who founded the first Laotian state, Lan Xang (Kingdom of the Million Elephants), with the help of the Khmer sovereign at Angkor. Fa Ngum was a great warrior, and between 1353 and 1371 he conquered territories that included all of present-day Laos and much of what is today northern and eastern Thailand. He extended the Indo-Khmer civilization to the upper Mekong River and introduced Theravāda Buddhism, preached by Khmer missionaries from Angkor.

In 1373 Fa Ngum was succeeded by his son Phya Sam Sene Thai, who did much to organize the pattern of administration and defense for the kingdom. After his death in 1416, a long period of calm, broken only by a Vietnamese invasion in 1478, allowed his successors to complete the work of organizing Lan Xang. This period of peace and tranquillity ended with Photisarath (reigned 1520–47), who involved Lan Xang in a struggle against Myanmar and Siam that lasted two centuries. Photisarath waged three wars against Siam and succeeded in placing his son Sethathirath on the throne of the Tai state of Chiang Mai (Chiengmai), marking Lan Xang's maximum territorial expansion. On Photisarath's death, Sethathirath returned to occupy his father's throne. His reign was marked by the loss of Chiang Mai to the Myanmar, by the transfer of the capital from Luang Prabang to Vientiane, and by the repulsion of two Myanmar invasions that took place *c.* 1565 and 1570.

When he died (1571) the Myanmar seized Vientiane (1574) and ravaged the country, which lapsed into anarchy until Souligna-Vongsa ascended the throne in 1637 and restored order. He fixed the frontiers with Vietnam and Siam by means of treaties and led two victorious expeditions against the principality of Chieng Khouang in the south. A defender of Buddhism and a patron of the arts, he embellished Vientiane and made it a centre of intellectual brilliance. His reign is considered by Laotians to be a Golden Age.

Vietnames rule

When Souligna-Vongsa died in 1694, one of his nephews seized the throne with the help of a Vietnamese army, thus placing Lan Xang under Vietnamese rule and initiating a period of chaos that ended in the partition of the kingdom of Lan Xang. Other members of the royal family refused to accept Vietnamese vassalage. With the northern provinces under their control, they declared themselves independent (1707) and established the separate kingdoms of Luang Prabang and Vientiane. The south seceded in turn and set itself up as the kingdom of Champassak (1713). Split into three rival kingdoms, Lan Xang ceased to exist.

UNDER FOREIGN RULE

During the 18th century the three Laotian states, continually at loggerheads, tried to maintain their independence from the Myanmar and the Siamese, who were contending for the control of western Indochina. Their weakness, resulting from their disunity, inevitably caused them to fall prey to the Siamese.

Vientiane, which had sided with the Myanmar, was invaded (1778), annexed, and made a state subject to Siam (1782). Luang Prabang, which had supported the Siamese, was invaded by the Myanmar (1752), who imposed their rule upon it until the Siamese supplanted them (1778). In the south, Champassak, which had supported Myanmar revolt against the Siamese, was also invaded (1778) and transformed into a dependency of Siam. Each of these kingdoms was placed under the control of a Siamese commissioner. The kings of Champassak, Vientiane, and Luang Prabang were allowed to rule in their respective kingdoms but had to pay tribute to Bangkok. Their appointments to the throne were made in Bangkok.

Chao Anou (king of Vientiane 1805–28) attempted to shake off this yoke. First, he strengthened the bonds of allegiance uniting Vientiane to Vietnam (1806), whose influence in Indochina had grown to rival that of Siam. Next, he persuaded Bangkok to give his son the governorship of Champassak, thus extending his frontiers as far as the old southern boundaries of Lan Xang. Thinking that the British, who had just defeated Myanmar, were going to attack Siam, he led three armies against Bangkok (*c.* 1826). But the Siamese regrouped their forces, marched on Vientiane, and defeated Anou, who fled to Vietnam. Vientiane was pillaged and destroyed. In 1819 Anou attempted another attack but was again defeated. Vientiane was made a Siamese province.

For the Siamese the annexation of Vientiane was the first step toward the creation of a great empire. They next extended their colonization of the left bank of the Mekong to protect themselves from an eventual Vietnamese expansion westward. They therefore garrisoned Champassak (1846) and Luang Prabang (1885) and stationed troops as far as the Chaîne Annamitique. Siamese expansion toward the northeast—where the mountain states were placed under the cosuzerainty of Vietnam and Luang Prabang—provoked the protests of the French, who had established a protectorate over Vietnam. France entered into negotiations with Bangkok (1886) to define the Siamese-Vietnamese frontier and won the right to install a vice consul in Luang Prabang. The office was entrusted to Auguste Pavie, who, owing partly to his popularity with the Laotians, succeeded in winning Luang Prabang over to France. After a number of Franco-Siamese incidents in the Mekong Valley, French ships made a show of strength off Bangkok. On the advice of the British, Siam withdrew from the left bank of the Mekong and gave official recognition to the French protectorate in the evacuated territory (1893). French annexation was completed by treaties with Siam (Thailand) in 1904 and 1907.

ench
otec-
rate

The French organized this territory as the Protectorate of Laos and allowed it autonomy in local matters. The kingdom of Luang Prabang survived, but the other provinces were placed under the direct authority of a French official. France paid little attention to Laos until 1941, when, under Japanese pressure, the Vichy government restored to Thailand the territories acquired in 1904. In March 1945 the Japanese drove the French from Indochina and proclaimed the independence of Laos.

Two movements sprang up at this time. The first was anti-Japanese and was represented by the court of Luang Prabang and Prince Boun Oum of Champassak; the second was anti-French (the Free Laos movement, or Lao Issara), was located in Vientiane, the former French colonial administrative centre, and was led by Prince Petsarath. These two movements remained in conflict until the return of French troops, which compelled the supporters of the Lao Issara to flee to Thailand. In 1946 France, in a temporary agreement, recognized the internal autonomy of Laos under the king of Luang Prabang, Sisavang Vong. Finally, after the promulgation of a constitution and general elections, a Franco-Laotian convention was signed on July 19, 1949, by which Laos was granted limited self-government within the French Union. All important power, however, remained in French hands.

Although many of the Lao Issara leaders were prepared to work with the French under this new arrangement, their decision was opposed by a more radical group led by Prince Souphanouvong. Under Souphanouvong's presidency a new political movement, the Pathet Lao (Lao Country), was created (1950) that joined forces with the Viet Minh of Vietnam in opposing the French. The Pathet Lao remained unreconciled when the French took further steps toward granting independence to Laos in October 1953, while still retaining control of all military matters in the kingdom. Between 1950 and early 1954 the Pathet Lao gained strength in northeastern Laos and had a firm grip on two of the country's provinces when the Geneva Conference brought the First Indochina War to an end.

INDEPENDENT LAOS

Geneva
Conference

At the Geneva Conference the 14 participating nations (including France, Great Britain, the U.S., China, and the Soviet Union) agreed on the establishment of Laos as a unified, independent buffer state between Thailand and North Vietnam, which were allied, respectively, to the West and to the Communist bloc. But this compromise agreement concerned only the international aspects of Laos' neutral status; the Laotians were left with the problem of making it work on the local level. The deep political divisions within Laos made it impossible for the country to function effectively as a neutral buffer state.

Relative calm prevailed during 1955–58 as a government of national union, including Pathet Lao representatives, sought to implement the Geneva accords. When elections held in 1958 showed there was significant support for the Pathet Lao, the right reacted by forcing the neutralist prime minister, Prince Souvanna Phouma, out of office and installing their own candidate, Phoui Sananikone. Then in July 1959 the right-wing forces in control of Vientiane imprisoned Prince Souphanouvong (he escaped a year later). In further violation of the Geneva agreement, the rightists accepted military aid from the U.S. and the Philippines. The Pathet Lao retaliated by seizing control of Phôngsali and Houaphan (Sam Neua) provinces, and a period of inconclusive hostilities followed. When Phoui showed some readiness to move toward a neutralist position, he was deposed by rightist army officers and his place was taken by Gen. Phoumi Nosavan. A period of great confusion followed as Phoumi Nosavan and his supporters were themselves deposed in a coup led by a parachute battalion commander, Kong Le, in August 1960, only to succeed in recapturing Vientiane, with considerable U.S. and Thai support, in the following December. By this stage the situation in Laos was a cause for growing tension between the U.S. and the Soviet Union.

Souvanna
Phouma
coalition

A change in U.S. policy under the new administration of Pres. John F. Kennedy brought acceptance of the concept of a neutral Laos, and in May 1961 the Geneva Conference was reconvened to seek a formula that would allow Laos to occupy a neutral position as a buffer zone. In June 1962 Prince Souvanna Phouma formed a new coalition government that included Pathet Lao, neutralist, and rightist representatives. Despite international support for such a solution, Laos could not disentangle its affairs from the war being fought in neighbouring Vietnam. Nor were the major political groupings within the country ready to give up control over those areas they dominated by means of their armed forces.

Increasingly from 1963, developments in Laos were linked with those in Vietnam. Sections of eastern Laos were a vital part of the Vietnamese Communist supply lines to the south (the Ho Chi Minh Trail), and, as the Vietnamese Communists gave their aid to the Pathet Lao, the U.S. joined in military support of the forces of the government in Vientiane. Despite the external assistance received by both sides, neither the government in Vientiane nor the Pathet Lao was able to break the pattern by which the territory of Laos was divided between them, with little apparent possibility of either gaining a decisive victory. In these circumstances, and with the U.S. and

North Vietnam negotiating for a cease-fire in Paris, the Vientiane government and the Pathet Lao began negotiations in 1972 and signed a cease-fire in February 1973. (P.-B.L./M.E.O.)

Beginning in that year, yet another effort was made to find a way to rule Laos through a coalition of rightists, neutralists, and the Pathet Lao. After further protracted negotiations, a Provisional Government of National Unity was inaugurated in April 1974 that had an equal number of members drawn from those who had been associated with the Vientiane government and from the Pathet Lao. Once associated with this provisional government, the Pathet Lao demonstrated an organizational ability that contrasted with the disunity of their rightist opponents. When Saigon and Phnom Penh fell to Communist forces in April 1975, morale dropped sharply among rightists in Laos; and from May onward, as many of the most important rightist politicians and soldiers fled to Thailand, the Pathet Lao moved swiftly and firmly to establish control over the whole of the country.

Lao People's Revolutionary Party

After the Pathet Lao gained control over Laos, the existence of a Laotian Communist Party (the Lao People's Revolutionary Party, founded in 1955 as the People's Party of Laos) was revealed. This party controls the Lao People's Democratic Republic, which was inaugurated in December 1975. The Communist regime has developed close military and economic ties with Vietnam, and it has supported Vietnam's involvement in Kampuchea. Despite considerable economic aid from Vietnam, the Soviet Union, and other countries, Laos has remained one of the world's poorest nations. (P.-B.L./M.E.O./Ed.)

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

MALAYSIA

Malaysia, a country in Southeast Asia, is composed of two regions—West Malaysia, on the Malay Peninsula, and East Malaysia, consisting of the territories of Sarawak and Sabah on the island of Borneo. It has a total land area of 127,316 square miles (329,747 square kilometres).

West and East Malaysia

West Malaysia has an area of about 50,800 square miles and occupies most of the Malay Peninsula south of latitude 6°40′ N. To the north it is bordered by Thailand, with which it shares a land boundary of 300 miles (480 kilometres). To the south, its nearest neighbour is the island republic of Singapore, with which it is connected by a causeway; to the southwest, across the Strait of Malacca, is the Indonesian island of Sumatra. East Malaysia, consisting of Sarawak and Sabah (formerly North Borneo), is separated from West Malaysia by some 400 miles of the South China Sea. Sarawak, with an area of 48,050 square miles, and Sabah, 28,460 square miles, occupy most of the northwestern coastal part of the large island of Borneo. Both these territories share land boundaries with each other and with the Indonesian provinces of Borneo (Kalimantan). Within Sarawak is the small enclave of Brunei, a sultanate under British protection, politically separate from Malaysia.

Malaysia, a member of the Commonwealth, represents the political marriage of territories that were formerly under British rule. When it was established on September 16, 1963, Malaysia was composed of Malaya (now West Malaysia), Singapore, Sarawak, and Sabah. In August 1965 Singapore left the federation and became an independent republic.

While not large by Asian standards, Malaysia is nevertheless one of the richest countries in the region. It occupies a central position in Southeast Asia, commanding one of the major sea lanes of the world, the Strait of Malacca. Malaysia is unique in that it is the only country that has territory on both the mainland and insular regions of Southeast Asia. The physical separation between mainland Malaysia and insular Malaysia (Sarawak and Sabah), however, inevitably poses problems of communication. The division between West and East Malaysia is not only physical. There are wide disparities and differences as well in the degree of development and in the living standards of the population.

Malaysia inherited many problems from its colonial past. The country was faced with the tasks of instilling a polyglot society with a sense of common identity and national purpose and of resolving the tensions arising from inequalities of political and economic conditions. Economically, difficulties have included a high rate of population growth (which poses the related problems of finding productive employment for an expanding labour force and of at least maintaining, if not raising, living standards); an initial inordinate dependence on rubber and tin production, which has been eased somewhat by the export of palm oil and timber; disparities in income between rural and urban peoples and between ethnic groups; and a shortage of skilled manpower.

Physical and human geography

THE LAND (WEST MALAYSIA)

Relief. The Malay Peninsula is a long, narrow strip of mountainous land extending southward from Myanmar and Thailand. West Malaysia occupies the southern half of the peninsula and is about 500 miles long and—at its widest east–west axis—about 200 miles wide. About half of Malaysia is covered by granite and other igneous rocks, a third more is covered by stratified rocks older than the granite, and the remainder is covered by alluvium. At least half of the land area is more than 500 feet (150 metres) above sea level.

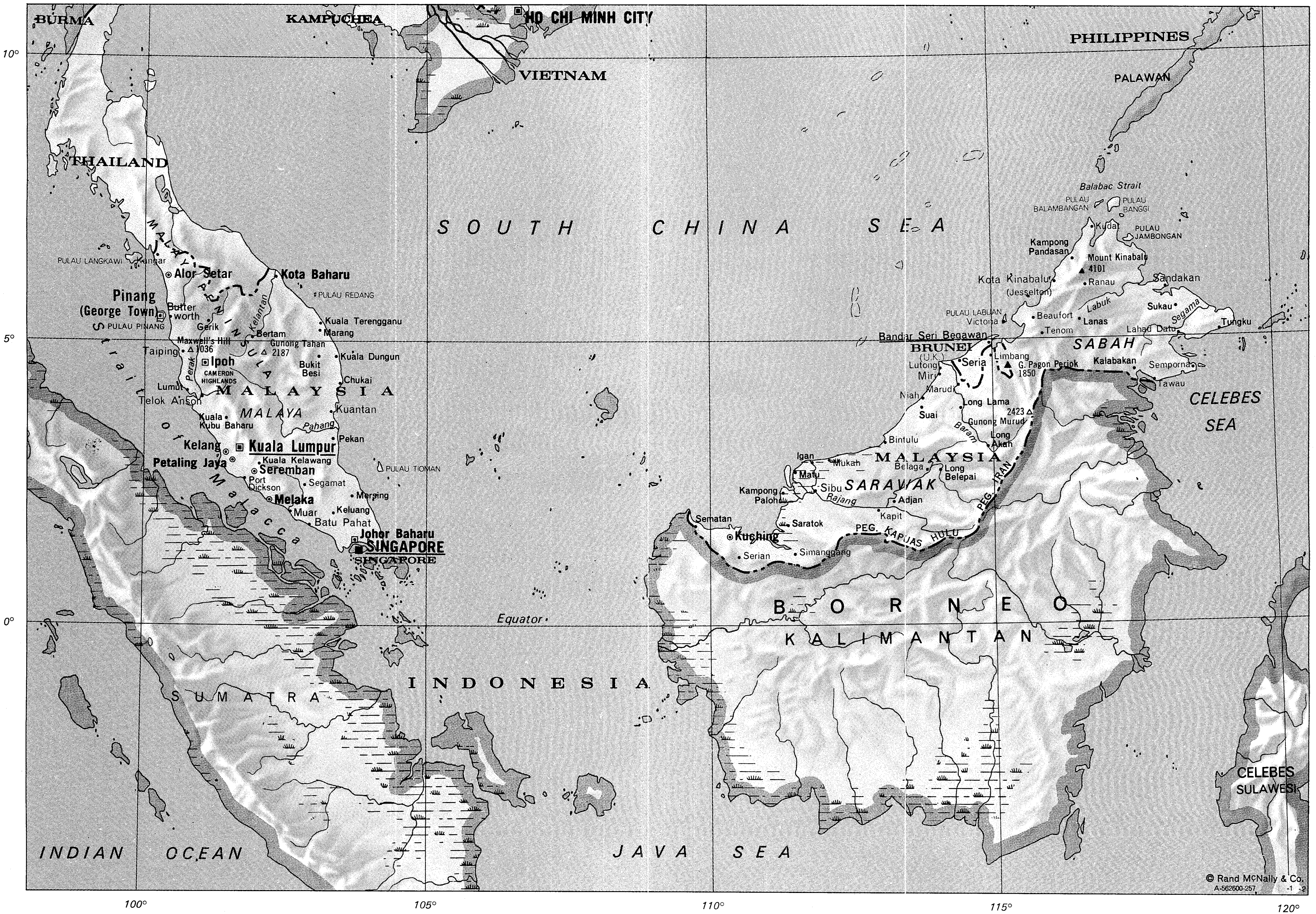
Mountain ranges

Topographically, the country is dominated by its mountainous core, which consists of a number of roughly parallel mountain ranges aligned north–south. The most prominent of these is the Main Range, which has peaks rising more than 7,000 feet above sea level and is about 300 miles long. Limestone hills, with characteristically steep, whitish-gray sides, stunted vegetation, caves created by the dissolving action of water, and subterranean passages, are distinctive landmarks in central and northern West Malaysia. Bordering the mountainous core are the coastal lowlands, 10 to 50 miles wide along the west coast of the peninsula but narrower and discontinuous along the east coast. Settlement and development have taken place primarily along the west coast.

Drainage. The peninsula is drained by an intricate system of rivers and streams. The longest river—the Sungai Pahang—is only 270 miles long. The year-round rainfall results in perennial streamflow, but the volume of water transported fluctuates because of the localized and torrential nature of the rainfall. In the western part of the peninsula such heavy falls may occur at any time of year, but in the eastern part they are more likely to occur during the northeast monsoon that occurs from about November to March. Prolonged rains often cause floods, especially in those areas where the natural regimes of the rivers have been disrupted by uncontrolled mining or agricultural activities.

Soils. The soils of the peninsula have been exposed for an extremely long period of time to intense tropical weathering, with the result that most of their plant nutrients have been leached out. The soils are typically strongly acid and coarse textured; they are deficient in nitrogen, phosphorus, potassium, and magnesium and have low amounts of organic matter. Only a small proportion of the soils of West Malaysia are fertile; regular applications of fertilizers are therefore necessary in order to sustain crop yields. On sloping ground, such additional measures as the building of contour embankments or the planting of protective cover crops are also needed in order to minimize soil erosion.

Climate. The Malay Peninsula has an equatorial climate, but its narrowness and topographical configuration—a central mountainous core with flat, flanking coastal plains—facilitate the inland penetration of mar-



MALAYSIA

Size of symbol indicates relative size of town • ◎ ▣ ◙ ■

Elevations in metres

0 100 200 400 600 km

0 100 200 400 mi

MAP INDEX

Cities and towns

Adjan	2·13n	113·09e
Alor Setar	6·07n	100·22e
Batu Pahat	1·50n	102·56e
Beaufort	5·22n	115·38e
Belaga	2·42n	113·47e
Bertam	5·08n	102·03e
Bintulu	3·13n	113·04e
Bukit Besi	4·46n	103·12e
Butterworth	5·24n	100·24e
Chukai	4·14n	103·24e
Georgetown, see Pinang		
Gerik	5·25n	101·08e
Igan	2·53n	111·42e
Ipoh	4·35n	101·04e
Jesselton, see Kota Kinabalu		
Johor Baharu	1·27n	103·45e
Kalabakan	4·23n	117·28e
Kampong Paloh	2·25n	111·13e
Kampong Pandasan	6·28n	116·32e
Kangar	6·26n	100·12e
Kapit	2·02n	112·54e
Kelang	3·02n	101·28e
Keluang	2·01n	103·19e
Kota Baharu	6·07n	102·15e
Kota Kinabalu (Jesselton)	5·58n	116·04e
Kuala Dungun	4·47n	103·26e
Kuala Kelawang	2·56n	102·05e
Kuala Kubu Baharu	3·34n	101·39e
Kuala Lumpur	3·09n	101·43e
Kuala Terengganu	5·19n	103·09e
Kuantan	3·50n	103·20e
Kuching	1·32n	110·19e
Kudat	6·53n	116·46e
Lahad Datu	5·03n	118·19e
Lanas	5·19n	116·35e
Limbang	4·45n	115·00e
Long Akah	3·17n	114·50e
Long Belepai	2·45n	114·04e
Long Lama	3·46n	114·24e
Lumut	4·13n	100·39e
Lutong	4·32n	114·00e
Marang	5·12n	103·11e
Marudi	4·11n	114·19e
Matu	2·42n	111·26e
Melaka	2·12n	102·15e
Mersing	2·26n	103·51e
Miri	4·23n	113·59e
Muar	2·02n	102·34e
Mukah	2·54n	112·02e
Niah	3·55n	113·41e
Pekan	3·28n	103·24e
Petaling Jaya	3·07n	101·39e
Pinang (George-town)	5·24n	100·19e
Port Dickson	2·32n	101·48e
Ranau	5·57n	116·40e
Sandakan	5·53n	118·05e
Saratok	1·44n	111·20e
Segamat	2·31n	102·49e
Sematan	1·49n	109·41e
Semporna	4·27n	118·36e
Seremban	2·44n	101·56e
Serian	1·12n	110·31e
Sibu	2·19n	111·51e
Simanggang	1·12n	111·32e
Suai	3·48n	113·38e
Sukau	5·32n	118·15e
Taiping	4·51n	100·44e
Tawau	4·17n	117·54e
Telok Anson	4·00n	101·02e
Tenom	5·08n	115·55e
Tungku	5·06n	119·01e
Victoria	5·18n	115·14e

Physical features and points of interest

Balabac Strait	7·35n	117·00e
Balambangan, Pulau, *island*	7·15n	116·57e
Banggi, Pulau, *island*	7·17n	117·12e
Baram, *river*	4·36n	113·59e
Cameron Highlands	4·29n	101·23e
Celebes Sea	3·00n	119·00e
Datu, Tandjung, *cape*	2·06n	109·39e
Iran, Pegunongan, *mountains*	2·05n	114·55e
Jambongan, Pulau, *island*	6·40n	117·27e
Kapuas Hulu, Pegunongan, *mountains*	1·25n	113·15e
Kelantan, *river*	6·11n	102·15e
Kinabalu, Mount	6·05n	116·33e
Labuan, Pulau, *island*	5·21n	115·13e
Labuk, *river*	5·54n	117·30e
Langkawi, Pulau, *island*	6·22n	99·50e
Malacca, Strait of	2·30n	101·20e
Malay Peninsula	6·00n	102·00e
Malaya, *historic region*	4·00n	102·00e
Maxwell's Hill	4·52n	100·48e
Murud, Gunong, *mountain*	3·52n	115·30e
Pagon Periok Gunong, *mountain*	4·18n	115·19e
Pahang, *river*	3·32n	103·28e
Perak, *river*	3·58n	100·53e
Pinang, Pulau, *island*	5·23n	100·15e
Rajang, *river*	2·04n	111·12e
Redang, Pulau, *island*	5·47n	103·00e
Sabah, *historic region*	5·20n	117·10e
Sarawak, *historic region*	2·30n	113·30e
Segama, *river*	5·27n	118·48e
South China Sea	10·00n	113·00e
Sulu Sea	8·00n	119·00e
Tahan, Gunong, *mountain*	4·38n	102·14e
Tioman, Pulau, *island*	2·48n	104·10e

Effects of monsoons

itime climatic influences. In addition, monsoonal effects further modify the climate. The country is influenced by eight or nine major airstreams, flowing from the northeast, the south, and the west; the advance and retreat of these airstreams are responsible for the division of the climatic year into four seasons. These are the northeast monsoon (from November or December until March), the first intermonsoon period (from March to April or May), the southwest monsoon (from June to October), and the second intermonsoon period (from October to November). The onset and termination of the monsoons are not sharply defined.

The mean annual rainfall is approximately 100 inches (2,540 millimetres). The driest location, in Kuala Kelawang (formerly Jelebu), about 30 miles southeast of Kuala Lumpur, receives about 65 inches of rain a year; while the wettest, Maxwell's Hill, about 25 miles northwest of Ipoh, receives 200 inches annually. Temperatures are uniformly high throughout the year, averaging 78° to 82° F (25° to 28° C) for the majority of lowland areas. Relative humidities are also persistently high, averaging from 82 to 86 percent. The climate, in consequence, tends to be hot and humid.

Plant and animal life. The characteristic vegetation of the peninsula is the dense, evergreen rain forest—the climax vegetation of the hot, wet tropics. Rain forest still covers about half of the total land area; another fraction is under swamp forest. The flora of the Malayan rain forest is among the richest in the world. There are about 8,000 species of flowering plants, of which at least 2,500 are trees. An acre (0.4 hectares) of forest may have as many as 100 different species of trees, as well as shrubs, herbs, lianas (creepers), and epiphytes (plants that grow nonparasitically upon others, deriving nourishment from the atmosphere). The forest canopy is so dense that little sunlight can penetrate it. As a result, the undergrowth is usually poorly developed and—contrary to popular belief—is not impenetrable. Much of the original rain forest has been destroyed by severe wind and lightning storms, by aborigines clearing it for temporary agricultural exploitation, or by clearances made for agricultural or commercial purposes. When such cleared land is subsequently abandoned, coarse grassland, scrub, and secondary forest develop. These types of vegetation constitute a smaller fraction of the total land area.

The forests and scrubland are inhabited by a large variety of animal life. Mammals include the elephant, the tiger, the magnificent seladang (or Malayan gaur, a massive wild ox), the Sumatran rhinoceros, the tapir (a hoofed and snouted quadruped), the wild pig, and many species of deer, including the *pelandok*, or mouse deer (a small, deerlike ruminant). Crocodiles, monitor lizards, cobras, and king cobras also are indigenous to the country, while the green sea turtle and the giant leathery turtle nest regularly on the beaches of the east coast.

Settlement patterns. West Malaysian settlements range from the rudimentary settlements of the Orang Asli (aborigines) to modern cities. This complexity and variety reflect in part the range of economic activities, in part the cultural diversity of the residents, and in part the long history of settlement in the peninsula. The first settlers were small groups of Orang Asli, who were later followed by peoples of Malay stock and by immigrants from China and India.

The Malay are predominantly a rural people. Their settlements are similar in appearance and pattern to those of their rural counterparts elsewhere in Southeast Asia. The basic unit is the kampong (village, or community of houses), composed of dwellings on stilts, which are commonly erected beside rivers, canals, beaches, roads, and footpaths. The houses are usually built of wood and a thatched roofing called atap, woven from the leaves of the nipa palm, a species also used for basketry. Each house is surrounded by trees bearing coconuts, bananas, papayas, and other fruits. The four main types of Malay settlements—fishing villages, paddy (rice-field) villages, cash-crop villages, and mixed-crop villages—despite their variations, all conform to the same basic pattern.

Types of rural settlement

Most other rural settlements are associated with peoples who have settled in the country since the early 19th century. The earliest of these were the mining camps, which sprang up in the tin fields on the western peninsula. Some have since grown into large towns, but others—especially in the Kinta Valley—still remain small. The British introduced the plantation system of agriculture, and the subsequent cultivation of rubber and of the oil palm has changed the face of rural West Malaysia, bringing to the landscape the plantation, or estate, settlement. This is usually a group of buildings composed of the processing factory and storehouse, the labourers' quarters, and the manager's house.

A modern type of settlement is the New Village, which originally was simply a group of buildings occupying a defensive site enclosed with barbed wire near a road. A total of about 550 New Villages were established between 1948 and 1960, during the Emergency, the formal name for the period when the administration was engaged in suppressing the Communist guerrilla uprising. They were part of a plan to resettle rural dwellers in easily defended positions, so as to prevent contact between them and the armed Communists. With the end of the Emergency in 1960, some of the New Villages were abandoned and have since disappeared. Most of them, however, took root and now have become permanent settlements. With their barbed-wire fences dismantled, they differ little from other small towns.

The cities and large towns were built up during the colonial and postcolonial periods and are distributed mainly along the tin and rubber belt in the western peninsula. The towns are associated with mining, purchasing, processing, distributing, exporting, and administrative functions; each town usually performs several of these functions. Some towns are located at coastal or riverine sites, emphasizing the early importance of water transport; more modern towns have been built in inland areas served by road, rail, and air transport. Slightly more than half the population of the urban centres is Chinese, and about one-third is Malayan. Indians and Pakistanis make up most of the remainder.

Jrban
opulation

Except for satellite towns such as Petaling Jaya (outside Kuala Lumpur), the towns of West Malaysia are unplanned, having grown up around pre-urban nuclei. Urban land use is generally mixed, and buildings are put to multiple uses. Streets, built for a more leisurely era, are narrow and often congested. In the larger centres, such as Kuala Lumpur, Ipoh, and George Town (Pinang), distinct central business districts, similar to those in Western cities, have emerged. These are characterized by heavy population and traffic densities, high land values, and a concentration of shopping, banking, insurance, entertainment, and other facilities.

THE LAND (EAST MALAYSIA)

Relief. East Malaysia is an elongated strip of land approximately 670 miles long, with a maximum width of about 160 miles. The coastline of 1,400 miles is paralleled inland by a 900-mile land boundary with Kalimantan (Indonesian Borneo). For most of its length the relief is composed of three topographic features. First, there is the flat coastal plain, which in Sarawak, where the coastline is regular, averages 20 to 40 miles in width. In Sabah, where the coastline is rugged and deeply indented, it is only 10 to 20 miles in width. Inland from the coastal plain is the second topographic feature—the hill and valley region. This is generally less than 1,000 feet in altitude, but isolated groups of hills reach heights of 2,500 feet or more. The terrain in this region is usually very irregular, with steep-sided hills and narrow valleys. The mountainous backbone that forms the divide between East Malaysia and Kalimantan is the third topographic feature. This backbone is both higher and nearer the coast in Sabah than it is in Sarawak. It is composed of an ill-defined complex of plateaus, ravines, gorges, and mountain ranges, all of which have been cut by erosion. The summits of the ranges are between 4,000 and 7,000 feet. Mt. Kinabalu (13,455 feet), the highest peak in the Malay Archipelago, towers above this mountain complex; no other peak in the country reaches 8,000 feet in height.

Drainage. As in West Malaysia, the drainage pattern is set by the interior highlands, which also form the watershed between East Malaysia and Kalimantan. The rivers are perennial because of the year-round rainfall. They form a dense network of drainage lines covering all parts of the country. The longest river in Sarawak—the Rajang—is 350 miles long and is navigable by shallow-draft boats for about 150 miles from its mouth; its counterpart in Sabah—the Kinabatangan—is of comparable length but is only navigable to about 120 miles from its mouth. The rivers are of importance as they provide a means of communication between the coast and the interior. Settlement also has taken place along the rivers, as it did on the peninsula in an earlier period.

ommuni-
ıtion by
ver

Soils. The soils of East Malaysia, like those of the peninsula, developed under intensive tropical weathering conditions over a long period. They are heavily leached and are generally poor in plant nutrients. When exposed, the organic matter in such soils is rapidly oxidized, and the soils consequently become even poorer. Soil erosion is always a danger on sloping land. In such an environment, agricultural development poses considerable problems. Generally, soil conditions in Sarawak and Sabah do not differ greatly from those in West Malaysia. Of these three territories, only Sabah has appreciable areas of fertile soils. These are found in particular on the Semporna peninsula, where the parent material from which the soil is formed is composed of basic volcanic rocks.

Climate. Both continental and insular Malaysia are in the same latitudes and are influenced by similar airstreams. They consequently have similarly high temperatures and humidities, heavy rainfall, and a climatic year patterned around the northeast and southwest monsoons. In coastal areas in East Malaysia minimum temperatures range from 72° to 76° F (22° to 24° C), and maximum temperatures from 88° to 92° F (31° to 33° C); temperatures are lower in the interior because of higher altitudes. As on the peninsula, rainfall variations are pronounced. Mean annual rainfall in Sabah varies from 60 to 120 inches, while most parts of Sarawak receive between 100 and 160 inches. The northeast monsoon brings heavy rain and rough seas to the exposed coasts of southwestern Sarawak and northern and northeastern Sabah. The southwest monsoon, however, affects mainly the southwest coastal belt of Sabah. Floods are common, especially along the west coast of Sabah. Neither mainland nor insular Malaysia is in the typhoon belt, but their coasts are occasionally subject to the heavy rainstorms associated with squalls.

Plant and animal life. Large expanses of East Malaysia—some three-fourths of Sarawak and Sabah—are still covered by dense evergreen rain forest. Soil type, location, and altitude produce distinctive vegetation zones, varying from tidal swamp forest on the coast, to freshwater and peat swamp forest on the ill-drained parts of the coastal plains, to lowland rain forest on the well-drained parts of the coastal plains and foothills up to an altitude of about 2,000 feet, to submontane and montane (lower mountain-type) forest above that altitude. The highly leached and sandy soils of parts of central Sarawak and the coast support an open, heathlike forest, known locally as *kerangas* forest. Large expanses of the forested land, especially in the upland areas, have been cleared for shifting cultivation, and the original forests have given way to scrub and savanna.

Animal life in East Malaysia is even more varied than it is on the peninsula. In addition to the peninsula species, East Malaysia is also the home of the fast-disappearing orangutan ape, the rhinoceros, the honey bear, and the unique proboscis monkey—a reddish tree-living species. There are also vast numbers of cave swifts, whose nests are regularly collected and sold as the main ingredient of bird's-nest soup.

Settlement patterns. About three-fourths of the population of East Malaysia is still rural, and it is in the rural areas that the greatest variety of settlement types is encountered. This variety is a direct reflection of the almost bewildering ethnic complexity of the population and of the fact that indigenous as well as immigrant groups are settled in the rural areas. The indigenous ethnic groups, such as the Iban (Sea Dayak), Land Dayak, Kenyah, Kayan, and Murut, are thinly scattered in the foothill country and, to some extent, in the coastal lowlands as well. They are primarily shifting cultivators and live in locations on or near the banks of rivers. Their characteristic dwelling is the longhouse, more commonly found in Sarawak than in Sabah. Each longhouse is raised on stilts and is composed of a number of rooms, known as *bileks;* each *bilek* houses a family. A longhouse can grow by accretions of related families, and an Iban longhouse may in time reach a length of 50 *bileks.* Some groups, such as the Melanau of Sarawak and the Kadazan (Dusun) of Sabah, have abandoned the longhouse settlement form, adopting instead the individual dwelling of the Malay.

The Malay and Melanau of East Malaysia share many common characteristics with their rural counterparts on the peninsula. They tend to be a riverine and coastal people, with an economy based on agriculture and fishing. Many live in kampongs set in the midst of coconut, mangrove, or other swamp trees. Their houses are generally built on stilts. The Melanau specialize in sago production; their kampongs are built in the large delta swamp region between Bintulu and Rajang. The rural Chinese in Sarawak have settled in the region between the coast and the uplands, usually in homesteads strung along both sides of the roads. They grow cash crops such as rubber, pepper, fruits, and vegetables in small holdings. Their houses are commonly built at ground level and are thus easily distinguishable from the stilt-raised dwellings of the Dayak and Malay.

Principal towns

Urbanization in East Malaysia has proceeded slowly. Only a small percentage of people live in towns. The largest towns in East Malaysia are Kuching, Sibu, and Miri in Sarawak, and Sandakan, Kota Kinabalu, and Tawau in Sabah. As on the peninsula, the urban population is predominantly Chinese. The large towns are invariably located on coastal or riverine sites. The layout and appearance of these towns are markedly similar: a wharf area; rows of Chinese shop-houses in the central business districts; more substantial buildings in the governmental administrative area; and one or more timber and atap kampongs built on the river banks.

THE PEOPLE

The population of Malaysia is unevenly divided between the peninsula and East Malaysia, with the vast majority living in West Malaysia. The population shows great ethnic, linguistic, cultural, and religious diversity. There are important differences between the indigenous and immigrant peoples, as well as among the indigenous peoples themselves, and between the Muslims, Buddhists, Christians, and tribes adhering to traditional religions.

West Malaysia. The Malay Peninsula, situated at one of the great maritime crossroads of the world, has been the meeting place of peoples from other parts of Asia. As a result, the population shows the ethnographic complexity typical of Southeast Asia as a whole. In essence there are four groups of people—the Orang Asli, the Malay, the Chinese, and the Indians and Pakistanis.

The Orang Asli constitute the smallest group and can be divided ethnically into the Jakun, who speak an archaic Malay, and the Semang and Senoi, who speak languages of the Mon-Khmer language family. They are primarily adherents of traditional religions, but a number have been converted to Islām.

The Malay originated from different parts of the peninsula and the Malay Archipelago. They comprise slightly more than half of the population and are politically the most important group. They share with each other a common culture, speak a common Austronesian language, Bahasa Malaysia, which is the national language, and are firm adherents of the Muslim religion. Adherence to Islām is regarded as one of the most important factors distinguishing a Malay from a non-Malay; the number of Malays who are not Muslim is negligible. Minor differences in dialect, culture, and physical characteristics are noticeable among the Malay living in the south in Johor state, on the east coast in the states of Kelantan and Terengganu, and on the west coast in the states of Negeri Sembilan, Perak, Kedah, and Perlis.

The Chinese, who make up about one-third of the peninsular population, were originally from the provinces of southeastern China and are ethnically homogeneous. They are, however, less homogeneous than the Malay in language and religion. These peoples from southeastern China comprise several different dialect groups, each with its own spoken language. Oral communication between two Chinese may thus sometimes depend on Mandarin Chinese, English, or Malay. A minority, the Baba Chinese, speak a Malay patois, although otherwise they remain Chinese in customs, manners, and habit. The most important cultural and linguistic Chinese groups are the Hokkien, from Fukien Province; the Cantonese, from the Canton region of Kwangtung Province; the Hakka, from the area between Canton and Swatow; the Tiechiu, from the Swatow region of Kwangtung Province; the Hainanese, from the island of Hainan; the Kwongsai, or Kwangsi, from Kwangsi Province; and the Hokchiu, from the hinterlands of Fukien Province. The Chinese do not have a dominant religion; most of them, while subscribing to the Confucian code of moral behaviour, are either Buddhists or Taoists. A small minority is Christian.

Origins of the Chinese populatioı

The peoples from the Indian subcontinent—the Indians, Pakistanis, and Tamil from Sri Lanka—constitute about 10 percent of the population of West Malaysia. Linguistically, they can be subdivided into speakers of Dravidian languages (Tamil, Telugu, Malayalam, and others) and speakers of Indo-European languages (Punjabi, Bengali, Pashto, and Sinhalese). Numerically, the Tamil speakers are the largest group. Most of the Indians and Sri Lankans are Hindus, while the Pakistanis are predominantly Muslim. Some Indians have been converted to Christianity. The Sikh, from the Punjab, speak their own language and also adhere to their own religion, Sikhism.

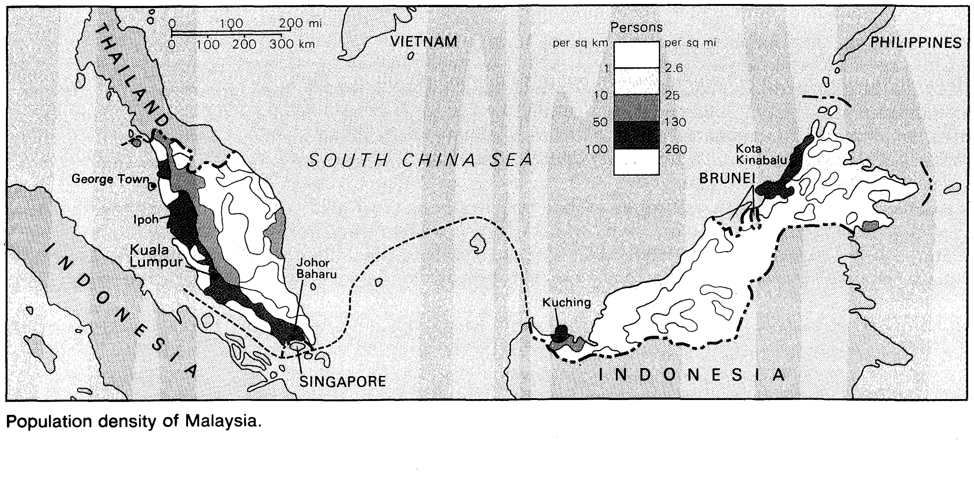
The population of West Malaysia also includes small numbers of Europeans, Americans, Eurasians, Arabs, and Thai.

Ethnic diversity

East Malaysia. The population of East Malaysia is ethnographically even more complex than that of West Malaysia. The government has tended to oversimplify the situation, listing seven main ethnic groupings in Sarawak and seven in Sabah, while there are, in fact, 25 ethnic groups in Sarawak and in Sabah, not counting minor tribal groups. The main ethnic groups in Sarawak are the Chinese, almost one-third of the total population; various speakers of mutually unintelligible Austronesian languages including the Iban (less than one-third); the Malay (less than one-fifth); the Land Dayak (about 9 percent); the Melanau (about 5 percent); and other minor groups, who make up the rest of the population.

The Chinese of Sarawak, drawn from the southeastern provinces of China, share common regional origins with the Chinese of West Malaysia, but the relative importance of each dialect group varies. Thus the Hakka and Foochow (Hokchiu) groups, who are less numerous than the Cantonese and Hokkien in West Malaysia, are the dominant linguistic groups, and together they comprise two-thirds of the Chinese population in Sarawak. The other important dialect groups are the Hokkien, the Cantonese, the Tiechiu, the Henghua, and the Hainanese. As in West Malaysia, the Chinese of Sarawak practice Buddhism, Taoism, and Confucianism. Only a few of the Chinese are Christians.

The Iban are the largest and most important indigenous



Population density of Malaysia.

group in Sarawak. Their origins are obscure, but they were originally headhunters. They are a homogeneous people speaking a language described as a type of pre-Islāmic Sumatran Malay. Most of them are longhouse dwellers, practicing shifting cultivation in the interior uplands. They have a distinctive culture, in which religion plays an important part; nearly every activity is influenced or governed by animist beliefs.

The Malay of Sarawak are a heterogeneous group of people, among whom only a few are of immigrant West Malaysian origin. Most are the descendants of indigenous peoples who were converted to Islām and adopted the Malay way of life some time since the mid-15th century. Although ethnically diverse, they are culturally homogeneous, speaking a common language and observing the Muslim faith.

he and ayak

The Land Dayak live in hill country; most of them are found in the First Division (one of Sarawak's seven administrative units that were created at different stages in its historical growth as a political entity). Although all are of the same ethnic group, they speak five different but related dialects that are to some extent mutually intelligible. The majority of the Land Dayak are adherents of traditional religions, but Christian missionaries have made some converts among them.

The Melanau differ ethnically from the Sarawak Malay, being shorter, broader shouldered, and of a lighter complexion. Their dialects, which are distinct from Malay, do not differ sufficiently to constitute a barrier to communication. Three-fourths of the Melanau population are Muslims; the rest, except for a small number of Christians, follow traditional religions. Other indigenous peoples—the Kenyah, Kayan, Kedayan, Murut, Kelabit, Bisaya, Punan, and others—contribute much to the ethnic and cultural diversity of East Malaysia.

Sabah also has a kaleidoscopic mixture of peoples. Of the total population, almost one-fourth are Kadazan; about one-fifth Chinese; about one-tenth Bajau; almost 5 percent Murut; while about one-fifth are composed of such indigenous peoples as the Brunei Malay, Kedayan, Orang Sungei, Bisaya, Sulu, Tidong, and Sino-native. Europeans, Eurasians, Malay, Indonesians, Filipinos, and Indians and Pakistanis comprise the remainder.

The Kadazan are composed of a number of tribes, each speaking a dialect that the others can understand. About a fourth of the Kadazan population are Christians, a small percentage are Muslims, and the remainder are animists. More than half of the Chinese are Hakka-speaking; the other important Chinese dialect groups are Cantonese, Hokkien, Tiechiu, and Hainanese. The Bajaus are not a cohesive community, being split into two main groups—the sedentary agriculturists, who have settled on the north coast, and those who live by the sea on the east coast. Most are Muslims, but not all of them can communicate with each other. The Murut of Sabah are ethnically different from the Murut of Sarawak, being descended from the same stock as the Kadazan. They are shifting cultivators. Although they are divided into subtribes, their languages are mutually intelligible. Although most follow traditional religions, approximately one-fourth are Christians.

THE ECONOMY

Malaysia's economy is one of the strongest in Southeast Asia. The monetary unit is the Malaysian dollar (*ringgit*). The country is the world's largest producer of rubber, tin, and palm oil. Its economy, compared with those of most developing countries, is relatively advanced in its infrastructure (*i.e.,* roads, water supplies, and electricity), diversification, institutional patterns, commercial and financial patterns, and management, professional, and administrative skills.

Resources. There are many mineral resources in Malaysia. West Malaysia's resources include tin, iron ore, bauxite (from which aluminum is made), coal, gold, china clay (kaolin), monazite (the principal ore of thorium, a malleable, radioactive metallic element), ilmenite (a titanium mineral), columbite (the principal ore of niobium, used in making steel alloys), manganese, tantalite (a mineral containing tantalum, a metallic element resistant to corrosion), wolframite (an important ore of tungsten, a metallic element with a high melting point), and zircon (the chief ore of zirconium, a metallic element, used in steel metallurgy). Antimony (a brittle, lustrous, white metallic element, used in alloys), mercury, gold, coal, petroleum, bauxite, phosphate, and glass sand are found in Sarawak; and gold, coal, manganese, and copper in Sabah. Except for petroleum, tin, iron ore, bauxite, and copper, most of the minerals are of minor economic importance. In Sabah the only mineral of great importance is copper.

Tin

Other than crude petroleum, tin is the most important mineral in Malaysia. It is found in most of the West Malaysian states and is distributed in two belts—western and eastern. The western tin belt runs along the flanks of the Main Range and along either side of the subsidiary granite ranges to the west of the Main Range. The less continuous, and also less rich, eastern tin belt runs from southern Terengganu to eastern Johor. Investigations indicate that there are large deposits of tin in the Malay Reservations (lands set aside by law for the use of Malays and which cannot be sold to others). The search for new fields has been extended to the offshore areas of the western peninsula. Iron ore deposits along the eastern and western peninsula have been largely depleted.

The production of crude petroleum is increasingly important. Malaysia is one of the largest producers of crude petroleum in the Far East.

Hydroelectric potential

The power resources of Malaysia include coal, peat, wood, petroleum, and hydroelectricity. Although there are proved reserves of coal and peat, they are not economical to mine and therefore remain only of potential use. Wood and charcoal were the traditional domestic fuels, but in the urban areas they have largely been displaced by bottled gas. Very little hydroelectric power is generated in East Malaysia, but, as on the peninsula, the abundant rainfall and steep gradients of the rivers in the interior highlands offer good hydroelectric resources. The nation's petroleum resources constitute a major source of fuel oil. For power generation, therefore, Malaysia continues to depend primarily on fuel oil.

Agriculture, forestry, and fisheries. Agriculture, mining, forestry, and fishing combined constitute the most important sector of the Malaysian economy. More than four-fifths of the work force of West Malaysia and almost two-fifths of the work force in Sarawak and one-fourth in Sabah are engaged in these pursuits. Nearly half of the economically active population continues to depend upon agriculture for a livelihood. The most important cash crops are rubber (the country's second largest foreign exchange earner), oil palm, and forest products. The main food crop raised in Malaysia is rice (called *padi* in Malay), which ranks second only to rubber in terms of the amount of land and labour devoted to its production. Cultivation systems range from the traditional shifting cultivation, still commonly practiced in East Malaysia, to the most modern forms of plantation agriculture.

The main biological resources of Malaysia are its forests and its fisheries. The extensive areas still under rain forest in both West and East Malaysia are a source of various forest products used by the indigenous peoples, as well as of timber, charcoal, firewood, poles, resins, gums, and rattans, which are produced both for local consumption and for export.

Most of the land area in Malaysia is still covered by forest. The timber industry has expanded, and the sustained export demand for tropical timber has stimulated the opening up of additional forest areas. The output of timber also has increased as a result of the clearance of large areas of forest for land development. Timber has become Malaysia's third largest foreign exchange earner, after crude petroleum and rubber.

The shallow seas off the coasts of Malaysia are the principal fishing grounds. In general the nutrient level, and hence the productivity of these seas, is low, the richer portions being those fertilized by drainage waters from the land. The annual output of the fishermen who use traditional techniques and who fish in inshore waters is low. Progress in the Malaysian fishing industry has been impressive. The long coastlines provide the Malaysian fishermen with easy access to the surrounding seas, although most of them still confine their fishing to the over-exploited shallow inshore waters. The greater use of trawlers and mechanized fishing boats has allowed the fish resources of the offshore grounds to be tapped, however.

Mining and petroleum exports

Industry. The mining and extractive industry contributes about one-third of the total export receipts. The main minerals produced are petroleum, tin, iron ore, and bauxite. West Malaysia is the world's largest producer of tin, but intensive exploitation has led to a steady depletion of known deposits, and the tin industry is consequently searching for new deposits off the coast of the western peninsula, as well as in Malay Reservations. The output of iron ore also has dropped due to the gradual depletion of high-grade deposits. Bauxite production declined after the closure of the Sarawak mine in 1965, but this was partially offset by increased output from the mines in southeastern Johor. West Malaysia and Sarawak produce most of the oil, while Sabah produces smaller amounts. The first offshore well began production in Sarawak in 1968. Malaysia's only copper mine is located in Mamut, Sabah; its output of copper concentrates has increased greatly.

Most of Malaysia's manufacturing plants are located on the peninsula. The contribution of the manufacturing sector of the economy has increased. Malaysia has attempted to broaden its economic base by industrialization. Efforts have been directed mainly at the production of hitherto imported consumer goods, and the range of manufacturing also has been widened to include production for export. Governmental efforts to promote industrialization include the granting of favoured treatment, including tax holidays, for selected industries; the establishment of tariff barriers for the protection of the home market; and the establishment of institutions to promote industrial development. The government has fostered industrial estates in the less developed states in an effort to industrialize those regions and to balance overall industrial growth.

Trade. Most of Malaysia's exports are made up of raw materials. Five commodities—crude petroleum, rubber, timber, palm oil, and tin—account for about three-fourths of the total value of all exports. The most important imports are transport equipment and machinery, manufactured goods, and food, which together constitute about two-thirds of the total value of all imports.

Economic development

Administration of the economy. The free enterprise system introduced during the British colonial era has been modified since independence. The government has embarked upon a program of planned economic development that aims at raising living standards of the population in general and of the economically weaker communities in particular. The program, however, encourages the continued participation of the private sector. Direct governmental participation is limited to the development of such infrastructural facilities as transportation, public utilities, drainage and irrigation works, agricultural settlement schemes, and industrial estates, as well as of social and administrative services. Malaysia is a member of the Association of South East Asian Nations (ASEAN), which works to integrate the economic and social development of the Philippines, Thailand, Indonesia, Singapore, and Malaysia.

Malaysia's systems of public finance—auditing and organization of accounts, parliamentary control, and revenue collection—are generally based on British principles. The country's fiscal system—rather than being employed to manipulate the pace of economic activity, the level of employment, or the level of prices—is basically a mechanism for raising revenue for governmental expenditure. Increasing dependence is being placed by the government on income and other direct taxes for revenue collection, while customs and excise duties provide a slightly smaller proportion of the total yield. The government is attempting, meanwhile, to distribute the tax burden equitably, by using such means as sliding income tax scales, higher duties on luxury goods, and lower duties on essential commodities.

More than half the total labour force of Malaysia is engaged in primary production. There is continued dependence on imports for many manufactured goods, and the industrial sector remains relatively small, though it has become the fastest growing sector of the economy. By the standards of other developing countries in Asia, Malaysia's economy is nevertheless relatively advanced, varied, and complex. The component states of Malaysia are, however, in widely different stages of economic development. West Malaysia, contributing almost 90 percent of the total GNP, is clearly the dominant area, with Sarawak and Sabah lagging far behind.

Economically, Malaysia needs to maintain a steady rate of economic growth despite a relatively rapid population increase. Underemployment, while its extent has not been accurately assessed, is widespread, especially in the subsistence agricultural sector. In Sarawak, population pressure already has led to land exhaustion and to soil erosion in many areas of shifting cultivation. Higher income agricultural industries, by contrast, have increased their employment.

In order to reduce its vulnerability to world market fluctuations, Malaysia has to diversify its economy and at the same time raise standards of living, especially in its depressed areas—in the eastern side of the peninsula and in the interior parts of Sarawak and Sabah.

Transportation. Although Malaysia's transportation system has expanded rapidly, it has failed to keep pace with the country's economic growth. The transport network in West Malaysia is well developed, especially in the tin and rubber belt. Both Sarawak and Sabah, however, have poor transport facilities, relying primarily on river transport for movement between the coast and the interior.

Peninsula roads

The road network in West Malaysia is considered the best in Southeast Asia. Most of the roads are hard-surfaced. There are highways running from Kuala Lumpur to Seremban, Kuala Lumpur to Karak, and Jerangau to Jabor. The road networks in Sarawak and Sabah are still inferior. Some of the roads in East Malaysia are bitumen-surfaced; the rest of the system, however, is generally of poor quality.

The rail transport system is only well developed on the peninsula. Sarawak has no railway, and Sabah has only a short line linking Kota Kinabalu to Tenom. Growth is notable in rail traffic, especially in bulk cargo and long-distance passenger traffic.

Traditionally, transport in both East and West Malaysia has been by water. River transport, however, is no longer important in West Malaysia except on parts of the east coast, but in East Malaysia it continues to play a major role. On the rivers of Sarawak and Sabah, which link the interior and the coast, almost all settlements are riverine, and many are frequently accessible only by water. Due to the long coastlines, the coastal and sea transport systems of Malaysia are of great importance. A number of coastal and river ports have been established at strategic locations—George Town and Port Kelang in West Malaysia; and Kuching, Sibu, Labuan, Kota Kinabalu, Sandakan, and Tawau in East Malaysia. Traffic at these ports has steadily increased.

Air transport has grown even more rapidly, with passenger traffic increasing mostly in West Malaysia. The rate of increase has been lower in East Malaysia. Regular internal services fly between Kuala Lumpur, Kuching, and Kota Kinabalu and also link the states of Malaysia. The Malaysian Airlines System was formed in 1971 and operates international services in the Pacific region in addition to its internal operations.

ADMINISTRATIVE AND SOCIAL CONDITIONS

The *yang di-pertuan agong*

Government. Malaysia is a federal constitutional monarchy with a nonpolitical head of state, or *yang di-pertuan agong* ("the supreme ruler"), who is elected from among nine state hereditary rulers for a five-year term. The legislature is composed of the Dewan Negara, or Senate, and the Dewan Rakyat, or House of Representatives. Malaysia also has a prime minister and cabinet, an independent judiciary, and a neutral civil service. The federal parliament is the supreme legislative body of the country. It also controls the finances of the government. A bill passed by both houses and sanctioned by the *yang di-pertuan agong* becomes a federal law.

The House of Representatives functions in a similar manner to the British House of Commons. It has a membership of 154, of which 114 are from West Malaysia, 24 from Sarawak, and 16 from Sabah. Members are elected to office from single-member constituencies on the basis of a simple majority. The term of office of a member is five years. The Senate has a membership of 58, of which 32 are appointed by the *yang di-pertuan agong* on the recommendation of the prime minister. The 26 elected members are made up of representatives from each of the state legislative assemblies. Voting in either house is by a simple majority, but amendments to the constitution require a two-thirds majority.

The *yang di-pertuan agong* appoints the prime minister, who must be a citizen in law as well as a member of the House of Representatives. On the advice of the prime minister, the *yang di-pertuan agong* then appoints the other ministers who make up the cabinet. The number of ministers is not fixed, but all must be members of the federal parliament.

The powers of the federal parliament are relatively wide and include the authority to legislate in matters concerned with defense, external affairs, internal security, the administration of justice, and citizenship. The state legislatures, however, retain responsibility for matters pertaining to Islāmic law and for matters pertaining to personal and family laws affecting those of the Islāmic faith, as well as for land laws. The constitution also provides that some items may be dealt with either by the federal or by the state legislature.

Each state of Malaysia has its own written constitution and its own legislative assembly. Each state also has an executive council collectively responsible to the legislative assembly and headed by a chief minister. Several Malay states—Johor, Kedah, Kelantan, Pahang, Perak, Selangor, and Terengganu—have hereditary rulers, who are entitled sultans. The raja (king) is the ruler in Perlis, and the *yang di-pertuan besar* ("the chief ruler") in Negeri Sembilan. The heads of the other states are appointed to office. They are known as governors in Melaka, Pinang, and Sarawak, while in Sabah the head of state is known as the *yang di-pertuan negara.* The ruler or governor of a state acts on the advice of the state government. The constitution provides for parliamentary elections and for elections to state legislatures, to be held at least every five years.

lections

Justice. The constitution of Malaysia, which is the supreme law of the country, provides that the judicial power of the federation shall be vested in the High Court of West Malaysia and the High Court in East Malaysia and also in subordinate courts. Above the High Courts is the Federal Court, with jurisdiction to hear and determine appeals from decisions by any High Court. The supreme head of the judiciary is the lord president of the Federal Court.

Each High Court consists of a chief justice and a number of other judges—12 in West Malaysia and four in East Malaysia. The High Court has unlimited criminal and civil jurisdiction and may pass any sentence allowed by law. Below the High Court are the subordinate courts, which consist of the Sessions Courts and the Magistrates' Courts. Both these lower courts have criminal and civil jurisdiction—criminal cases coming before one or the other court depending upon the seriousness of the offense and civil cases depending upon the sum involved. In addition, there are religious courts in those Malay states that are established under Islāmic law. These courts are governed by state and not federal legislation.

Armed forces. The Malaysian armed forces have increased in strength and capability since the formation of Malaysia in 1963. After the withdrawal of British military forces from Malaysia and Singapore at the end of 1971, a five-power agreement between Malaysia, Singapore, New Zealand, Australia, and Great Britain was concluded to ensure defense against external aggression. The ASEAN also provides additional regional security. In the early 1970s Malaysia was engaged in internal security operations against guerrilla forces of the Malaysian Communist Party on the West Malaysia–Thai border, as well as against the clandestine Communist organization in Sarawak.

:curity
range-
ents

The armed forces consist of the army, the navy, and the air force. The army, which is the most experienced and the largest of the three, includes infantry brigades, reconnaissance regiments, scout cars, artillery regiments, and various supporting arms. The Royal Malaysian Navy includes a complement of frigates, coastal minesweepers, and a number of patrol boats. The emphasis is on speed and manoeuvrability for the purpose of defending the long, indented coastlines and narrow waters of Malaysia against intruders. The Royal Malaysian Air Force has combat aircraft, as well as many transport aircraft and helicopters.

The states of Malaysia inherited from their common colonial past an internal security system based on the British model. The police force is well trained and combats not only crime but also subversive activities, including armed Communist insurrection.

Education. Governmental policy is to provide nine years of education to each child. Enrollment in primary schools has increased. Considerable progress also has been achieved at the secondary level. Institutions of higher learning include the Universiti Malaya (University of Malaya), the Universiti Sains Malaysia (University of Science, Malaysia, formerly the University of Pinang), the Universiti Kebangsaan Malaysia (National University of Malaysia), the Universiti Pertanian Malaysia (University of Agriculture), and the Universiti Teknologi Malaysia (Technological University of Malaysia). The Mara Institute of Technology and a number of teacher-training colleges and vocational training centres are other important facilities. Enrollment in higher education has also increased.

Health and welfare. The general level of health has improved. The country is free from many of the diseases that plague tropical countries, but diseases borne by carriers, such as malaria, are still a problem in rural areas. Health conditions and health facilities vary among the component states, being better in West Malaysia than in the states of Borneo. Within each state, health services are better in the towns and cities than in the rural areas. Segments of the rural population continue to rely on traditional rather than modern medicine for treatment. Most of the modern health services are provided by the government. Welfare services are, however, provided by both government and voluntary agencies and include relief programs for the needy, the aged, and the handicapped.

The multicultural character of the population of Malaysia is visibly reflected in the wide variety of houses, which range from the longhouses and stilt houses of the rural peoples to examples of modern architecture in the cities. There is an abundance of forest-derived building material in the rural areas, so that no statistically measurable housing problem is evident, even though some of the dwellings in the more remote areas are often only temporary shelters. In the larger towns and cities of Malaysia, however, slums are common. A governmental housing authority has been established to assist in developing low-cost housing.

Many people, especially in Sarawak and Sabah, live by hunting, gathering, fishing, and simple farming; wage earners form only part of the total economically active population. Because of the increasing pressure of population on the land, however, there is a growing tendency for the male labour force to seek employment in manufacturing. Persons with an upper secondary or post-secondary education have entered the labour force at a fast rate, resulting in a higher unemployment rate among skilled workers than among the unskilled. Wages in the manufacturing sector, nevertheless, continue to be higher than wages in the agricultural sector, and underemployment continues to characterize the rural economy. Increasing industrialization is expected to draw increasing numbers of workers from the countryside to the cities and to create a greater demand for skilled workers.

CULTURAL LIFE

Malaysia, with its complexity of peoples and cultures, is a melting pot of several important cultural traditions, stemming from the Malay Archipelago as well as from China, the Indian subcontinent, the Middle East, and the West. Malay culture and Bornean culture are indigenous to the area. In the first one and a half millennia AD, indigenous Malay culture in the Malay Peninsula and in other parts of Southeast Asia was strongly marked by pre-Islāmic Indian and early Islāmic influences. Indian contact with the Malay Peninsula extended over a period from about the 4th century AD to the late 14th century, exerting a profound influence upon religion (through Hinduism and Buddhism), art, and literature. Islām, introduced to Malacca (now Melaka) in the 15th century, soon became the dominant religion of the Malay. The introduction of Western cultural influences in the 19th century affected many aspects of Malay life, especially in technology, law, social organization, and economics.

Contemporary Malay culture is thus multifaceted, consisting of many strands—animistic, early Hindu, early and modern Islāmic, and, especially in the cities, Western. The collective pattern thus established is distinct from other cultures and recognizably Malay.

Unlike the early Chinese traders who settled in Malacca and George Town and were partially assimilated (at least to the extent of adopting the Malay language), the Chinese who emigrated to the Malay Peninsula in the late 19th and early 20th centuries in large numbers were usually transients who established self-contained communities. Chinese cultural influence has consequently been minimal. The Chinese immigrants themselves, moreover, did not form a homogeneous group. Their culture in Malaysia has its roots in the culture and civilization of prerevolutionary China, with modifications brought about by local circumstances and environment.

Most of the Indians and Pakistanis originally came as labourers to work in the coffee and rubber plantations. Like the Chinese, they too, until World War II, were mainly transients, living in closed communities and remaining virtually unassimilated.

The communities of Malaysia have been affected by British colonial rule and Western modernizing influences. Western cultural influence has been greatest in education and institutional forms. Traditions and cultural institutions have been least affected in the rural areas—in eastern West Malaysia and in the interior of East Malaysia—while the cities have been the focus of the most rapid cultural changes.

Art forms

External cultural influences have made the least impact in music, dancing, literature, and the decorative arts. In East Malaysia the indigenous cultural background includes no written history or literature. Architecture is little developed, and the principal art forms are dancing and handicrafts, represented notably by the textiles handwoven by the Punan tribe, cloth made by the Bajau people, patterned rattan mats and basketwork, and wood carvings. Particularly on the peninsula, the artistic manifestations of Malay culture are mainly in literature, music, dancing, and the decorative arts. Painting and sculpture are poorly developed, primarily because Islām does not encourage the representation of the human form. Examples of the Malay decorative arts include batik cloth (cloth hand dyed by using a special technique), silverware, the handmade kris (a short sword or heavy dagger with a wavy blade), wood carving, and basketwork. Malaysian Chinese culture is derived from Chinese civilization and is represented by literature, drama, music, painting, and architecture. Some Malaysian artists, of Malay, Chinese, and Indian origin, also have begun to produce new art forms, especially in painting and architecture, that represent a synthesis that is distinctively Malaysian in character.

The newspapers are all privately owned and vary greatly in circulation, quality of reporting, and news coverage. Among the educated groups, the press is the principal source of information. In remote rural areas the radio is relied upon. Television is, however, the most popular medium among all language groups. (O.J.B.)

For statistical data on the land and people of Malaysia, see the *Britannica World Data* section in the BRITANNICA WORLD DATA ANNUAL.

History

WEST MALAYSIA

Malaya has been inhabited for at least 6,000 years. Archaeologists have unearthed evidence of several Stone Age cultures as well as an Early Bronze Age civilization. These probably existed contemporaneously with more advanced settlements stimulated by the arrival of adventurers from India beginning in the 2nd century BC.

Rise of Indianized states. The Malay Peninsula, and indeed most of Southeast Asia, experienced more than 1,000 years of Indian or Indianized influence. Indian blood and Indian culture blended with indigenous elements to produce an amalgam, which found political expression in the creation of a number of states and several great empires. Lacking the fertile plains of Java or Cambodia, the peninsula was unable to support a great empire. Although knowledge is scanty and is based largely on Chinese written sources, it does appear that at least 30 small states, nearly all along the east coast, flourished during this millenium. These included Tun Sun, in the northeastern part of the peninsula, where in the 3rd century AD there were more than 500 Indian merchants and priests. Ch'ih-t'u ("Red Land," Tanah Merah in Malay) may have been established in the 6th century on either the Kelantan or Trengganu River. The most important of the Indianized states was Lankasuka, whose control extended at times across the whole northern part of the peninsula. The most significant complex of Indianized temple ruins in Malaya, on the slopes of Kedah Peak opposite Pulau Pinang (former Penang Island), was at one time part of this state. Kedah, the name of a modern Malay state, was once thought to be Kalah, described by Arab sailors of the 9th through the 14th centuries. This state, it is now thought, was north of the peninsula, in present-day Thailand.

Indian influence

These small states, now long forgotten, have left a living legacy, traces of which are to be found even in the political structure of modern Malaysia. In religion, too, traces of Hinduism and Buddhism persist in the ritual, language, and practice of the Malay Muslims. A similar influence is seen in many of the arts and crafts still flourishing on the peninsula.

Much of the influence was imposed indirectly from India by the great Indianized empires of Southeast Asia. The small states of the peninsula, it would appear, were usually under the hegemony of one or the other of these empires, either in Siam or Cambodia to the north or Sumatra and Java to the south. On Singapore Island, for example, Tumasik, an Indianized settlement, owed allegiance in the 14th century first to Sumatra-West Malaya, under Śrivijaya; then to Majapahit, the great Javanese empire; and finally to Ayutthaya (Siam). Paramesvara, Tumasik's Sumatran ruler near the end of the 14th century, unsuccessfully attempted to win independence for Tumasik from Ayutthaya. Tumasik was sacked, but Paramesvara found refuge at the small fishing village of Malacca (now Melaka), where he established himself as ruler in about 1400.

Advent of Islām. From this precarious beginning, Malacca rose during the 15th century to become a political power of the first rank, the most important commercial mart in Southeast Asia, and the main centre for the diffusion of Islām. A fortunate set of circumstances combined to permit this. Soon after he won control of Malacca, Paramesvara secured the support of China. Chinese settlers at that time were beginning to move into the Nanyang ("South Seas"), or general Southeast Asian region. The Chinese admiral Cheng Ho called at Malacca in 1409 and again in 1414. In return for annual tribute, once taken to Peking by Paramesvara himself, the Chinese emperor promised his protection to Malacca. For three decades, China's guardianship deterred other Southeast Asian states, notably Siam, from crushing Paramesvara's settlement and permitted it to develop. Its location, in addition, was attractive to traders from India, Arabia, China, and the Indonesian islands as a central, independent meeting place not for the purchase of Malaccan goods, of which there were few, but for barter and exchange. A good government and a sensible nonrestrictive policy encouraged the traders to frequent the peaceful bay in increasing numbers. The fishing village thus grew into a bustling and famous city where many of the commodities of Asia were traded.

Emergence of Malacca

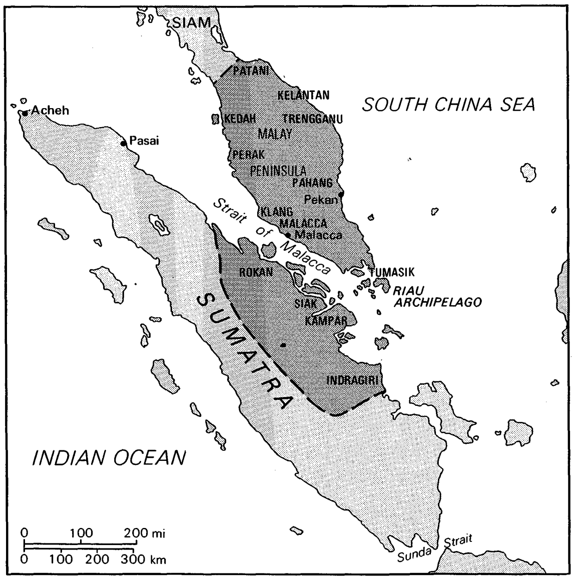
Malacca's most important role, however, was not commercial but religious. It replaced Sumatran ports as the main centre for the transmission of Islām throughout Southeast Asia. Originally disseminated by peddlers and traders from east Bengal and other areas of the Indian subcontinent, Islām had secured a toehold in north Sumatra and east Malaya during the 14th century or earlier. With the growth of Malacca in the 15th century, however, the new faith was conveyed most effectively along Malaccan trade routes by missionaries and merchants. The ancient Hindu beliefs survived only on Bali; elsewhere, the courts and the peoples accepted the new faith and helped it to spread, mainly as a major movement for peace, to the coastal areas of West Irian and the Philippines.

On the Malay Peninsula itself, the spread of Islām was linked closely to the territorial expansion of the Malacca sultanate, which began during the reign of Muzaffar Shah (1446–59), Paramesvara's great-grandson. In 1456 Muzaffar Shah appointed an outstanding man, Tun Perak, as *bendahara* ("prime minister"). In that year, also, Tun Perak became the saviour of Malacca, defeating the Siamese in a fierce naval battle off South Johore. For some 40 years (1456–1498), although he never became sultan, he was the effective ruler of Malacca and promoted its expansion. The state of Perak already had become Muslim and in 1460 became a part of Malacca; Pahang, a tributary of Siam, was conquered. Islām followed and moved north to Trengganu. South Malaya also was made part of the growing sultanate. The river states in Sumatra that had been commercially linked to Malacca were acquired in 1477. In 1474 Kedah, Malaya's one granary of importance, also received a Muslim sultan.

Early European intrusions. Long before the end of the 15th century the fame of Malacca had reached Europe, attracting those interested in the spice trade. The Portuguese, who for a century had been dedicated to the national endeavour of reaching the Orient by sea, arrived at Malacca in 1509. It was the beginning, as one Asian historian has called it, of "the Vasco da Gama Era," the period of European dominance in Asia. Although the Portuguese and subsequently the Dutch did not exert the dominance exercised by the British in the 19th century, the European intrusion of the 16th century was, nevertheless, the beginning of a political movement that grew in strength for over 400 years and has only just ended.

In 1511 Malacca was captured by a Portuguese fleet led by Afonso de Albuquerque. Along with the Moluccas (Maluku) and Goa (in India), it became one of Portugal's major trading outposts in Asia. The Portuguese were, however, few in number, and although they retained defensive possession of Malacca until 1641, Malaya was more deeply influenced in the 16th century by Acheh, an aggressive state in Sumatra.

When the Achinese first attacked Malacca in 1537, the city was saved by its new stone fort. Subsequent assaults came in 1539, 1547, 1568, 1570, and 1573. Acheh's target was not merely Malacca but all the Malay states as well. The story is complex and confused, for the allies of one decade were often enemies in the next. The Malay states and the Portuguese sometimes fought one another, sometimes joined together. Acheh remained the aggressor until well into the 17th century. By that time the few hundred Portuguese at Malacca were a spent force. They had to face a new opponent, the Dutch, who had secured a base in 1596 on Java. In 1607 the Dutch allied themselves with Acheh, while the Portuguese in Malacca joined with Johore. The Achinese sacked Johore, 1613–1615; Pahang in 1617; Kedah in 1619; and Perak in 1620. Acheh thus secured suzerainty over the Malay states for 20 years.



Malacca Empire in 1500.

The Dutch

In Java the Dutch, in the meantime, had grown steadily in strength. They began actively meddling in the affairs of the Malay states in 1633, when they laid siege to Malacca. Year after year the few Portuguese waited in vain for relief, surrendering finally in 1641. Malacca had become a moribund town, its bustle and importance long vanished. Despite all efforts by the Dutch to revive the port and to channel the trade in tin from the west Malay states to it, Malacca never recovered its earlier glory.

The Minangkabau and the Buginese. The last 50 years of the 17th century witnessed the migration of many Minangkabau people from Sumatra. These sturdy agriculturalists settled in south Malaya, which until then had been virtually unpopulated. They brought with them an almost uniquely matrilineal outlook by which property and authority descended through the female. Their leader was freely elected (their method of election has served as the basic model for selecting the ruler of modern Malaysia), and in the late 18th century, they formed a confederation of small states, which came to be known as Negri Sembilan, or "Nine States."

During the 18th century the Dutch found that profits from dealings with Malacca could not compare with the benefits to be derived from trade with Java and the Spice Islands. Malacca was therefore neglected, a situation that permitted large-scale penetration of Malaya by the "Norseman of the East," the Buginese rovers from the Celebes (or Sulawesi). Just as the well-governed state of Normandy arose as the Norseman's contribution to Atlantic civilization, so are Selangor and Johore perhaps the best examples of Buginese occupation and organization in Southeast Asia. Hardly a member of the Johore civil service today does not have Buginese blood in his veins, and Johore's independent progress in the 19th and 20th centuries is attributable largely to the migrations of the 18th century. Although the Buginese had traded for many centuries on the Java Sea, they did not begin to settle in the relatively weak Malay Peninsula until the late 17th century. In Selangor they established their ruler as sultan as early as 1700. The Buginese leader, Raja Haji, seized Perak and sacked Alor Star, capital of Kedah, in 1770. By that time the Buginese had developed a new trading centre south of Singapore, in the Rhio (Riau) archipelago, where a policy of free trade and peaceful government attracted Asian and European merchants. Malacca was nearly captured by the Buginese in 1784.

The Buginese

Dutch influence in the Malay Peninsula was weakened not only by the Buginese invasions but also by the arrival of the British. Seeking a source for goods to be sold in China, the East India Company acquired Penang Island in 1786. In 1795 the British captured Malacca, and in 1819 Sir Thomas Stamford Raffles established a settlement on Singapore Island. Britain's entrenched position was clarified beyond dispute by the Anglo-Dutch treaty of 1824. Malacca, Singapore, and Penang Island, the three "Straits Settlements," came under the direct control of the British Colonial Office in 1867. Singapore, a free port, grew rapidly.

With the opening of the Suez Canal in late 1869, the full effect of European technological superiority swept over Southeast Asia. The Malay states were little prepared. With the exception of Johore, which was effectively ruled by Abu Bakar, the states were generally poorly administered and failed to cope with their mounting problems, chief of which was the steady immigration of Chinese in search of tin. From the mid-19th century to the present, the Chinese and the Malays have been two elements of an inadequately integrated community, prone to racial clashes and communal disorder. The clashes, which included disputes between rival Chinese groups on the newly discovered tin fields of Perak, became sufficient in 1874 to provoke British intervention.

Chinese immigration

This intervention by a European power into the internal affairs of weak Asian states was at first crude. Malay reaction saw the murder of J.W.W. Birch, the first British resident (chief officer) in Perak (1875). Lack of understanding and incompetence also characterized the first British residents in Selangor and Sungei Ujong, one of the Negri (Negeri) Sembilan ("Nine States"). They were followed, however, by more talented men. Frederick Weld (1823–91) expanded the policy of intervention to encompass all the Negri Sembilan and Pahang on the east coast. He selected competent officers to work as residents in the various states. Among these was Frank Swettenham, resident in Selangor, whose service in the Malay Peninsula culminated in his appointment as first resident general, when the four states of Perak, Selangor, Negri Sembilan, and Pahang were brought together in a federation in 1896, with Kuala Lumpur as its capital. The other sultanates in the peninsula were also induced to accept British officers as residents or "advisers" but managed to avoid joining the federation. They saw too great a loss of individuality and control despite the British policy of preserving the sultanates, the Muslim faith, and the Malay way of life. Between 1902 and 1909 Trengganu (Terengganu), Kelantan, and Kedah were extricated from the Siamese sphere of influence. In the south, Johore (Johor) accepted a British adviser in 1914.

The genesis of the modern state of Malaya. British political control over the whole peninsula radically transformed Malaya economically and socially. As with the Romans in Britain, the British in Malaya devoted much attention to transport. Railways appeared, first from the coast to the tin fields, and then north to south down the peninsula until connections with Singapore were completed in 1924. A road network was constructed and paved to withstand the rains. Penang (Pinang), Singapore, Port Swettenham (Klang), and other ports were improved and given constant attention by the British.

Tin and rubber industries

Development of transport assisted in the rapid growth of the important tin and rubber industries. During the 19th century technological developments in the U.S. and Europe created an ever-increasing demand for tin. The Malays were indifferent to the opportunities created by this demand, and thousands of Chinese migrated to Malaya during the second half of the century to develop tin mining and manufacture. Vast new deposits were found, and the industry remained in Chinese hands through the 19th century. The Chinese, however, lacked capital and technology. Their open-cut, labour-intensive methods of mining were replaced by the mining dredge introduced by British-financed companies in the early 20th century. Two major tin smelters were established at Penang and Singapore, and "Straits tin" became known throughout the world. The tax on exported tin became one of the two main sources of revenue for Malaya; the other was a similar levy on rubber.

The first strain of rubber, introduced in the 1870s, had proved unsuccessful. The planters grew coffee until the development in the 1890s of a superior strain of rubber. From 1896 through the early 20th century, funds from British financiers provided for the clearing of thousands of acres of virgin jungle and the establishment of new rubber plantations. Rubber production was increased also with the allotment of acreage to individual Malay and Chinese smallholders.

Whereas Chinese immigrants had been attracted in great numbers by the tin industries of Perak and Selangor in central Malaya, between 1880 and 1920 thousands of south Indians moved to the west coast of Malaya to work on the rubber estates. A plural society was thus created and was maintained largely by the educational system. One school system taught the Malays in Malay. Missionary schools used English to teach the Christian Chinese, Indians, and Malays. Still other schools, financed by Chinese and supported by the Chinese Communist and Kuomintang parties, instructed Chinese residents in Chinese. The status quo of the plural society, preserved by Britain, was unchallenged until 1941. The few attempts toward self-government were discouraged.

In the field of public health, the British eliminated malaria, beriberi, and other previously incurable diseases.

Japanese invasion and occupation

By the close of 1941 the relations among East Asian nations had reached a critical stage. Japan, locked in a struggle with China since 1937, and increasingly desperate for raw materials such as oil, faced an embargo on raw materials from its major supplier, the U.S. To reach the oil and rubber of Java, Japan had to neutralize the great British naval base on Singapore Island. Attacks on Pearl Harbor on December 7, 1941, and on the Philippines on December 10 eliminated the U.S. from the Pacific, at least for several decisive months. On December 8 the Japanese had invaded Malaya and then headed south to Singapore.

The campaign was the greatest disaster the British had experienced since Cornwallis' surrender at Yorktown; the Japanese victory was complete. Singapore was captured within 70 days. Within three days Japanese aircraft had sunk the major British battleships "Prince of Wales" and "Repulse." Within a week they had swept the skies clear of the antiquated British aircraft. The troops of Gen. Yamashita Tomoyuki outfought and outmanoeuvred Gen. A.E. Percival and his British, Indian, Australian, and Malay troops. The Japanese scored further successes at Jitra in north Kedah, December 11–12, 1941, and at Slim River in Perak on January 7, 1942. Singapore fell on February 15. The Japanese lost 15,000 men. The total loss to the British Commonwealth was approximately 166,600 men.

Birth of Malayan nationalism

Following the expulsion of the Japanese and the return of British rule in late 1945, an attempt by London to organize Malaya into one state deeply offended the Malays, who prized their treaty arrangements with Britain and were determined not to become a mere colony. A tremendous upsurge of Malay political feeling, led by Dato Onn Bin Jaafar, resulted in the creation in 1946 of the United Malays National Organization (UMNO). In 1948 a federation joining the Malay Peninsula and Penang was formed.

In that year also the Malayan Communist Party (MCP), formed before the war among the Chinese community, rose in armed revolt and opposed the British from enclaves in the jungle. An emergency was proclaimed, as Britain endeavoured to suppress the MCP by military means and by eliminating the political causes of its agitation. These aims finally were achieved by Gen. (later Field Marshal) Sir Gerald Templer, who arrived in 1952. Templer invigorated military measures already planned and, as high commissioner, encouraged the attainment of *merdeka* ("independence") through political cooperation among the races. Under a new leader, Tunku Abdul Rahman, the UMNO in 1952 joined with the Malayan Chinese Association (MCA), led by Tan Cheng Lock and his son Tan Siew Sin. Both parties were anti-Communist and anticolonial. In the national election of 1955 the UMNO–MCA Alliance, led by Abdul Rahman, captured 51 of 52 seats. Britain responded by relinquishing its powers in Malaya. On August 31, 1957, Malaya secured its *merdeka,* and the emergency period ended three years later.

Postwar Malaya has been marked above all by the political drive of the Malays, expressed through UMNO. Only by associating with this powerful organization and its leader, Tunku Abdul Rahman, were other elements in the community able to achieve some of their goals. Malay dominance has caused bitterness, and racial antagonism found expression in rioting in Kuala Lumpur on May 13, 1969. Tun Abdul Razak succeeded the Tunku in 1970.

Tun Abdul Razak continued the peaceful progress of the country until his death in 1976. The third prime minister of Malaysia, Datuk Hussein Onn (son of Onn Bin Jaafar), faced with a deteriorating racial situation, was forced to circumscribe some of the civil liberties, processes of parliamentary government, and economic freedom of the Chinese in order to sustain the economically underprivileged Malays (or *bumiputras*). These steps toward a more traditional Asian leadership produced some friction. The country continued to progress, however, buttressed by new discoveries of petroleum offshore. (K.G.T.)

SABAH

Although there is evidence of Chinese trade from the 7th century onward, Sabah's contacts were confined to the Philippines for centuries. It was known to the Sulu pirates as the Land Below the Wind, because of its position below the typhoon belt.

After the rise of Portuguese power and the capture of Malacca on the Malay Peninsula by the Portuguese, many Muslim merchants moved to Brunei, which became an important centre of Islām. It was one of the great sea states of the Malay Archipelago and exercised suzerainty over Sabah.

During the 16th century, North Borneo was visited frequently by the Portuguese, who established trade with Sabah. At this time the Spaniards were competing with the Portuguese for hegemony, as well as with the Dutch and British. One of the main reasons for Dutch and, particularly, British intervention in North Borneo was the prevalence of piracy in the area and their need as shipping powers to control piracy.

In 1762 a British expedition sent to capture Manila was ceded the Sabah Peninsula, which then became known as North Borneo, by the Sultan of Sulu in gratitude for his release from Spanish captivity. The British also occupied the island of Labuan at this time. North Borneo came under full British control at the end of the 19th century. In 1872 the Sultan of Sulu gave William Cowie, a gunrunner, permission to establish a base on Sandakan Bay. Cowie was also given other trading privileges in North Borneo, but he ran into opposition from an American, Joseph William Torrey, who claimed trading privileges for the area under an older grant from the Sultan of Brunei. Torrey joined with Baron Gustavus de Overbeck, who represented a private syndicate headed by two English brothers, Alfred and Edward Dent, and although the Sultan of Brunei had no right to cede the territory, and despite the protests of the Spanish and Dutch governments, who had claims in the area, the Dents became the possessors of a large part of North Borneo, for which they paid an annual rent to the Sultan of Brunei. They then formed a limited provisional association, which became the British North Borneo Company in 1882, incorporated by royal charter. In 1888 the state was made a British protectorate, and, for the next 20 years, territorial consolidation took place, the final boundaries being defined in 1905.

After Japanese occupation in World War II, Sabah was granted crown colony status, in 1946.

SARAWAK

Little was known of Sarawak's rich prehistory until 1947, and nothing until then was excavated. Since 1947, energetic excavational and related fieldwork has been carried out by the Sarawak Museum throughout Sarawak, the adjacent state of Brunei, and, on a smaller scale, Sabah.

The classic site to date has been excavated on a considerable scale. This is the great cave at Niah, 400 miles northeast of Kuching, the capital of Sarawak, and 15 miles inland along a small river behind the South China Sea. Among the most significant discoveries at Niah is a radiocarbon dating of 40,000 years at what is only a medium level in the deposit, much of which remains to be excavated. From 40,000 years ago and more, there is an almost unbroken succession of human frequentations and occupations of the cave continuing after the arrival of a massive Chinese trade starting in the T'ang and Sung dynasties (AD 618 to 1279).

The Niah site provides a unique sequence of the evolution of stone tools from early hand-chopping axes, which are massive pebbles simply struck on one side to a point; through the development of flake tools; then finely edge-ground tools of the Mesolithic; on into the beautiful black polished adzes and axes of the Neolithic, or Late Stone Age. A fine set of wall drawings, provisionally attributed to the Early Iron Age, has also been found in one of the caves. These, too, are so far exceptional in Southeast Asia. They are associated with decorations showing the journey of the "ship of the dead" to the spirit world. Underneath the drawings, on the floor of the cave, more than 200 feet above the valley floor, beautifully carved model boats have been used as coffins for the deceased.

The Early Iron Age materials at the Niah Cave are closely paralleled by those found elsewhere in the Sarawak river delta, at the southwest corner of the island. There, about 30 places have been mapped as prehistoric monuments, reflecting what must have been an enormous trade between the Asian mainland and western Borneo reaching back more than 1,000 years. Traders from the mainland seem to have brought mainly ceramics, metal, and probably clothing. In return they received edible bird's nests, rhinoceros horn, hornbill ivory (*ho-ting;* one of the most valuable materials for carving in ancient China and to be obtained only from this area), camphor, spices, woods, bezoar stones, rattans, and other goods. With whom they traded is not yet clear. The nature of an earlier people is overlaid by many generations of interbreeding with existing groups. (T.Hn./Ed.)

By the mid-14th century, Sarawak came under the control of the great Hindu–Javan state of Majapahit. With the decline of Majapahit in the 15th century, Sarawak became the southern province of the sultanate of Brunei, one of the most important centres of Islām and one of the great sea states of the Malay Archipelago.

In 1839 the Dayak and the Malays in Sarawak rebelled against the oppressive rule of the Sultan's governor. At this time James Brooke, an English adventurer and a former officer of the East India Company, arrived in Sarawak, and in return for his services in putting down the rebellion, he was promised the title of raja and control of a large area of Sarawak. After some difficulty in getting the Sultan to honour this agreement, Brooke was installed as raja of Sarawak (1841) over the sector from Tanjong Datu to the Batang (River) Samarahan. The next few years he endeavoured to suppress piracy and head-hunting, often with the aid of British naval forces.

During the wars between the Dutch and the Kongsis in the 1850s, there was considerable migration of Chinese into Brooke's domain. The Chinese, aroused by the second Anglo-Chinese war then in progress, formed a secret society intent on taking control from Brooke and establishing an independent state. In 1857 the Chinese seized the capital, Kuching, and forced Brooke to flee, but the Malays and the Dayak of Sarawak put down the revolt and restored order. The Chinese remained in Sarawak, and, aided by migrants, they formed one of the important elements in the country. Chinese also continued to migrate to Brunei, and their migration to Sabah was encouraged in the 19th century by the British North Borneo Company in order to obtain a supply of labour for the exploitation of the country.

Under the provocation of the pirates, who kept moving their bases of operation eastward, Brooke extended his control through purchase and annexation over all of Sarawak west of Rajang River by 1861. By stages, the Brooke family extended control over more Brunei territory, under the rationale that the Sultan's control was either ineffective or oppressive. By 1890 the Sultan of Brunei controlled only a small portion of his original domain. In 1850 the United States recognized Sarawak as an independent domain, and recognition was given by the British in 1864. In 1888 an agreement with the British government placed Sarawak under British protection, but the Brooke family maintained control over all domestic affairs.

The Brooke "dynasty" continued to rule until World War II. In 1941 the third "white raja," Sir Charles Vyner Brooke, abrogated his powers by enacting a constitution designed to establish democratic self-government, but the effort was delayed by the Japanese occupation (1942–45). The territory, devastated by war, was ceded by Brooke to the British crown in 1946. (L.A.P.G./Ed.)

UNIFICATION AND THE FEDERATION OF MALAYSIA

Singapore secured full self-government in 1959. Its leader, Lee Kuan Yew, immediately pressed for "independence

Conflict with Singapore

through merger" with Malaya. Singapore shared much with the peninsula, and Lee felt that its future could be assured only by uniting with the mainland. Tunku Abdul Rahman, the leader of independent Malaya, agreed to the merger in order to forestall the growth of Communist influence among Singapore's Chinese population. To offset the inclusion of the 1,000,000 Chinese of Singapore, however, Malaya suggested in 1961 that Sarawak and Sabah (North Borneo), the other British colonies in Southeast Asia, which had substantial Malay populations, join with Malaya and Singapore in the merger. Both Sabah and Sarawak became independent in September 1963. After considerable discussion, and increasing opposition from the Philippines and Indonesia, the four states established a new, independent federation of Malaysia on September 16, 1963. (August 31, the date of Malaya's founding in 1957, remains the national day.)

Within two years, however, although Malaysia was able to withstand the external pressures of Indonesian armed opposition and "confrontation," internal disagreements worsened between Kuala Lumpur and Singapore. Abdul Rahman believed that civil war was imminent, and Malay political supremacy seemed challenged. On August 9, 1965, Singapore was separated from Malaysia and established as an independent republic.

Indonesian hostility took the form of raids across the borders of Sarawak and Sabah by guerrillas operating from Indonesian Borneo. The raids continued sporadically until 1966, when the two countries signed an agreement. In the late 1960s the Philippines unsuccessfully revived its longstanding claim to the territory.

In the 1970s Malaysia, along with nonaligned Thailand, worked to suppress Communist forces. With the resurgence of Islāmic fundamentalism in the late 1970s, many Muslim Malays joined the Dakwah (missionary) movement that was especially popular among young educated Muslims. Their increasing militancy led to desecration of Hindu temples and racial tension among the ethnically diverse Malaysians.

Race problems were further exacerbated by the arrival of boat people from Vietnam (1979), most of whom were ethnic Chinese. Prodded by Malay protests that the ethnic balance of Malaysia would be upset, the government sought to forcibly expel the boat people and even issued orders that any caught landing on Malaysian shores should be shot on sight, a decision that provoked international criticism.

Racial tensions

Malaysia, as a federation of disparate ethnic, economic, and political components, has had difficulties and problems from the moment of its inception in 1963. The most intractable of these problems—the strained relations between West Malaysia and Singapore—was solved by a political separation. Other problems, no less serious, remain. Racial riots between the Malays and the Chinese in May 1969—which led to the suspension of parliamentary rule and to the declaration of a state of emergency that lasted until February 1971—were an indication of the underlying tensions stemming from ethnic imbalances. These tensions persisted through the 1980s and early 1990s, with Malays constituting a majority of the population in five of Malaysia's six poorest states. The Malays, virtually all of whom are Muslim, have been increasingly attracted to radical forms of Islām, thus further exacerbating racial tensions.

There are also other imbalances—between West and East Malaysia, between urban and rural areas, and between modern and traditional ways of life. There is very little trade between East Malaysia and West Malaysia, indicating the need for closer economic integration between the states. Malaysia also faces the threat of Communist aggression on the Thai and Sarawak borders. Within the country itself, however, the fundamental problem remains that of eliminating ethnic disparities in income while fostering the growth of a Malaysian consciousness and sense of identity among the people. (K.G.T./Ed.)

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

MYANMAR

Myanmar is an independent socialist republic in Southeast Asia, with an area of 261,228 square miles (676,577 square kilometres). It is bordered by Thailand and Laos on the east, by China (the autonomous region of Tibet and the province of Yunnan) on the north and northeast, by India on the northwest, by Bangladesh and the Bay of Bengal on the west, and by the Gulf of Martaban and the Andaman Sea on the southwest and south.

In 1989 the country's official English name was changed from the Union of Burma to the Union of Myanmar (Burmese: Pyeidaungzu Myanma Naingngandau); in the Burmese language the country has long been known as Myanmar. The change was made because the name Burma connotes Burman, the country's largest ethnic group, although there are several minority ethnic groups. The name of the capital, formerly Rangoon, was changed to Yangôn; many other geographic names were also changed. In this article the name Burma is used for the country during the period of British rule, 1885–1948; the name Myanmar is used in all other contexts.

Myanmar stretches from latitude 10° N to about 28° 30′ N; almost half the country is, therefore, situated outside the tropics, but, because of its configuration, it is generally considered to be a primarily tropical country. The country has roughly the shape of a diamond, with a long tail—running south along the Malay Peninsula—attached to it. Its total length from north to south is 1,275 miles (2,052 kilometres); its width at the widest part, across the center of the country about the latitude of Mandalay, is 582 miles (931 kilometres) from east to west.

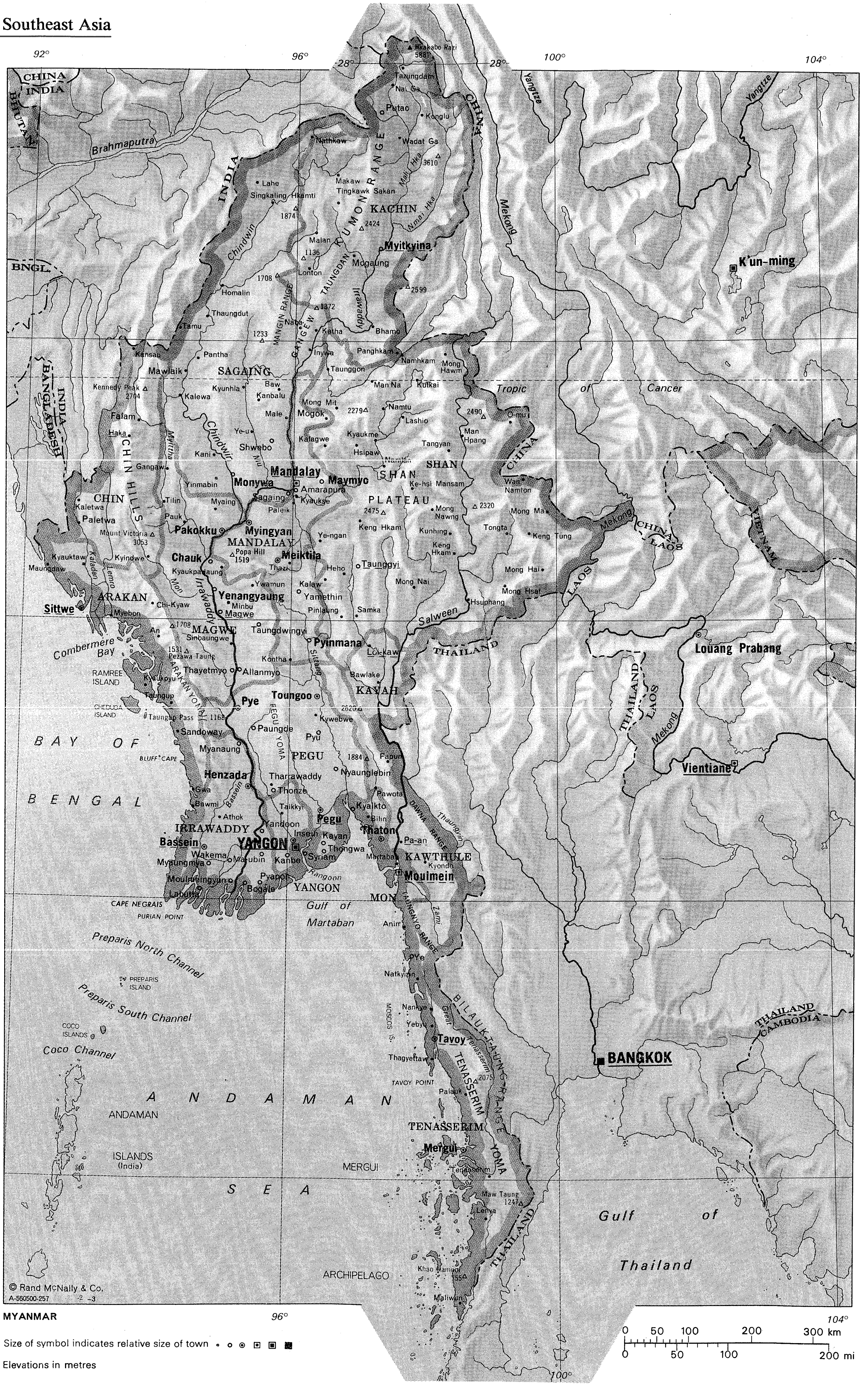
In ancient and medieval times, Myanmar was regarded as a gateway to the Indochinese Peninsula, as well as to China; it was known as the Golden Land, because of its abundant natural resources and the wealth that its control of trade routes generated. In the 11th century it became the homeland of Theravāda Buddhism (one of the two great schools of Buddhism, the other being Mahāyāna) at a time when that faith was being suppressed in other countries. After a 60-year series of Anglo-Burmese wars, during which portions of the country were conquered, Myanmar (Burma) was annexed by the British in 1886 and was administered as a province of India until 1937. Burma fell under Japanese rule for a period during World War II; British rule ended in 1948, when the country regained its sovereign independence.

Myanmar is individualistic in its social structure and nationalistic in its outlook. As one of the original nonaligned countries—maintaining independence from both Eastern and Western power blocs—Myanmar achieved some international importance in the first decade of its independence; thereafter, however, it tended to remain aloof and isolated from the international community. The land and air battles that were waged throughout the country during World War II destroyed many towns and villages and dislocated the economy. Although rice exportation was resumed, Myanmar failed to regain its status as the foremost rice-exporting country in the world. The wounds of war had not completely healed, partly because of lack of capital and partly because of insurgency and political unrest. Since independence, some modest industrialization programs have been embarked upon, but Myanmar remains a primarily agricultural country.

Physical and human geography

THE LAND

Relief. The country slopes from north to south, from an elevation of 19,296 feet (5,881 metres) at Hkakabo Razi peak in the extreme north to sea level at the Irrawaddy and Sittang deltas. The mountain ranges are longitudinal,



© Rand McNally & Co.
A-560500-257 -2 -3

MYANMAR

Size of symbol indicates relative size of town ◦ ○ ⊙ ▣ ▣ ■

Elevations in metres

MAP INDEX

Political subdivisions

Arakan	19 00 N	94 15 E
Chin	22 00 N	93 30 E
Irrawaddy	17 00 N	95 00 E
Kachin	26 00 N	97 30 E
Kawthule	17 30 N	97 45 E
Kayah	19 15 N	97 30 E
Magwe	23 00 N	95 00 E
Mandalay	22 00 N	96 00 E
Mon	16 30 N	97 35 E
Pegu	18 00 N	96 30 E
Sagaing	24 00 N	95 00 E
Shan	22 00 N	98 00 E
Tenasserim	13 00 N	99 00 E
Yangôn	16 50 N	96 10 E

Cities and towns

Allanmyo	19 22 N	95 13 E
Amarapura	21 54 N	96 03 E
An	19 47 N	94 02 E
Anin	15 40 N	97 46 E
Athok	17 12 N	95 05 E
Bassein	16 47 N	94 44 E
Baw	23 19 N	95 50 E
Bawlake	19 11 N	97 21 E
Bawmi	17 19 N	94 35 E
Bhamo	24 16 N	97 14 E
Bilin	17 14 N	97 15 E
Bogale	16 17 N	95 24 E
Chauk	20 54 N	94 50 E
Chi-Kyaw	20 17 N	93 54 E
Falam	22 55 N	93 40 E
Gangaw	22 11 N	94 07 E
Gwa	17 36 N	94 35 E
Haka	22 46 N	93 51 E
Heho	20 43 N	96 49 E
Henzada	17 38 N	95 28 E
Homalin	24 52 N	94 55 E
Hsipaw	22 37 N	97 18 E
Hsuphang	20 18 N	98 42 E
Insein	16 53 N	96 07 E
Inywa	23 56 N	96 17 E
Kalagwe	22 31 N	96 31 E
Kalaw	20 38 N	96 34 E
Kaletwa	21 45 N	92 48 E
Kalewa	23 12 N	94 17 E
Kanbalu	23 12 N	95 31 E
Kanbe	16 42 N	96 01 E
Kani	22 26 N	94 50 E
Kansau	23 50 N	93 35 E
Katha	24 11 N	96 21 E
Kayan	16 54 N	96 34 E
Ke-hsi Mansam	21 56 N	97 50 E
Keng Hkam	21 01 N	98 29 E
Keng Hkam	21 27 N	97 03 E
Keng Tung	21 17 N	99 36 E
Konglu	27 16 N	97 56 E
Kontha	19 30 N	96 03 E
Kunhing	21 18 N	98 26 E
Kutkai	23 27 N	97 56 E
Kyaikto	17 18 N	97 01 E
Kyaukme	22 32 N	97 02 E
Kyaukpadaung	20 50 N	95 08 E
Kyaukpyu	19 05 N	93 52 E
Kyaukse	21 33 N	96 06 E
Kyauktaw	20 51 N	92 59 E
Kyindwe	20 58 N	93 51 E
Kyondo	16 35 N	98 03 E
Kyunhla	23 21 N	95 18 E
Kywebwe	18 42 N	96 25 E
Labutta	16 09 N	94 46 E
Lahe	26 20 N	95 26 E
Lashio	22 56 N	97 45 E
Lenya	11 28 N	99 00 E
Loi-kaw	19 41 N	97 13 E
Lonton	25 06 N	96 17 E
Magwe	20 09 N	94 55 E
Makaw	26 27 N	96 42 E
Malan	25 37 N	96 21 E
Male	23 02 N	95 58 E
Maliwun	10 14 N	98 37 E
Mandalay	22 00 N	96 05 E
Man Hpang	22 41 N	98 36 E
Man Na	23 27 N	97 14 E
Martaban	16 32 N	97 37 E
Ma-Ubin	16 44 N	95 39 E
Maungdaw	20 50 N	92 21 E
Mawlaik	23 38 N	94 24 E
Maymyo	22 02 N	96 28 E
Meiktila	20 52 N	95 52 E
Mergui	12 26 N	98 36 E
Minbu	20 11 N	94 52 E
Mogaung	25 18 N	96 56 E
Mogok	22 55 N	96 30 E
Mong Hai	20 46 N	99 49 E
Mong Hawm	23 51 N	98 20 E
Mong Hsat	20 32 N	99 15 E
Mong Ma	21 37 N	99 54 E
Mong Mit	23 07 N	96 41 E
Mong Nai	20 31 N	97 52 E
Mong Nawng	21 39 N	98 08 E
Monywa	22 05 N	95 08 E
Moulmein	16 30 N	97 38 E
Moulmeingyun	16 23 N	95 16 E
Myaing	21 37 N	94 51 E
Myanaung	18 17 N	95 19 E
Myaungmya	16 36 N	94 56 E
Myebon	20 03 N	93 22 E
Myingyan	21 28 N	95 23 E
Myitkyina	25 23 N	97 24 E
Naba	24 15 N	96 11 E
Nai Ga	27 48 N	97 30 E
Namhkam	23 50 N	97 41 E
Namlan	22 15 N	97 24 E
Namtu	23 05 N	97 24 E
Nankye	14 20 N	98 11 E
Nathkaw	26 53 N	96 13 E
Natkyizin	14 55 N	97 57 E
Nyaunglebin	17 57 N	96 44 E
O-mu	22 58 N	99 18 E
Pa-an	16 53 N	97 38 E
Pakokku	21 20 N	95 05 E
Palauk	13 16 N	98 38 E
Paleik	21 49 N	96 03 E
Paletwa	21 18 N	92 51 E
Panghkam	23 53 N	97 37 E
Pantha	23 49 N	94 33 E
Papun	18 04 N	97 27 E
Pauk	21 27 N	94 27 E
Paungde	18 29 N	95 30 E
Pawota	17 46 N	97 17 E
Pegu	17 20 N	96 29 E
Pinlaung	20 08 N	96 47 E
Putao	27 21 N	97 24 E
Pyapon	16 17 N	95 41 E
Pye (Prome)	18 49 N	95 13 E
Pyinmana	19 44 N	96 13 E
Pyu	18 29 N	96 26 E
Sagaing	21 52 N	95 59 E
Samka	20 09 N	96 57 E
Sandoway	18 28 N	94 22 E
Shwebo	22 34 N	95 42 E
Sinbaungwe	19 43 N	95 10 E
Singkaling Hkamti	26 00 N	95 42 E
Sittwe (Akyab)	20 09 N	92 54 E
Syriam	16 46 N	96 15 E
Taikkyi	17 19 N	95 58 E
Tamu	24 13 N	94 18 E
Tangyan	22 29 N	98 24 E
Taungdwingyi	20 01 N	95 33 E
Taunggon	23 38 N	96 32 E
Taunggyi	20 47 N	97 02 E
Taungup	18 51 N	94 14 E
Tavoy	14 05 N	98 12 E
Tazungdam	28 02 N	97 35 E
Tenasserim	12 05 N	99 01 E
Thagyettaw	13 45 N	98 09 E
Tharrawaddy	17 39 N	95 48 E
Thaton	16 55 N	97 22 E
Thaungdut	24 26 N	94 42 E
Thayetmyo	19 19 N	95 11 E
Thazi	20 51 N	96 05 E
Thongwa	16 46 N	96 32 E
Thonze	17 38 N	95 47 E
Tilin	21 42 N	94 04 E
Tingkawk Sakan	26 04 N	96 44 E
Tongta	21 20 N	99 16 E
Toungoo	18 56 N	96 26 E
Wadat Ga	26 00 N	97 00 E
Wakema	16 36 N	95 11 E
Wan Namton	22 03 N	99 33 E
Yamethin	20 26 N	96 09 E
Yandoon	17 02 N	95 39 E
Yangôn	16 47 N	96 10 E
Ye	15 15 N	97 51 E
Yebyu	14 15 N	98 12 E
Yenangyaung	20 28 N	94 52 E
Ye-ngan	21 09 N	96 27 E
Ye-u	22 46 N	95 26 E
Yinmabin	22 05 N	94 54 E
Ywamun	20 31 N	95 25 E

Physical features and points of interest

Andaman Sea	12 00 N	95 00 E
Arakan Yoma, *mountains*	19 00 N	94 40 E
Bassein, *river*	15 56 N	94 18 E
Bengal, Bay of	18 00 N	93 00 E
Bilauktaung Range	13 00 N	99 00 E
Bluff Cape	18 00 N	94 26 E
Cheduba Island	18 48 N	93 38 E
Chindwin, *river*	21 26 N	95 15 E
Chin Hills	22 30 N	93 30 E
Coco Channel	13 45 N	93 00 E
Coco Islands	14 05 N	93 18 E
Combermere Bay	19 37 N	93 34 E
Dawna Range	16 50 N	98 15 E
Gangew Taungdan, *mountains*	24 55 N	96 35 E
Great Tenasserim, *river*	12 24 N	98 37 E
Hkakabo Razi, *mountain*	28 20 N	97 32 E
Irrawaddy, *river*	15 50 N	95 06 E
Kaladan, *river*	20 09 N	92 57 E
Kennedy Peak	23 19 N	93 45 E
Kumon Range	26 30 N	97 15 E
Lemro, *river*	20 25 N	93 20 E
Mali Hka, *river*	25 43 N	97 29 E
Mangin Range	24 20 N	95 42 E
Martaban, Gulf of	16 30 N	97 00 E
Maw Taung, *mountain*	11 39 N	99 35 E
Mekong, *river*	20 28 N	100 10 E
Mergui Archipelago, *islands*	12 00 N	98 00 E
Mon, *river*	20 20 N	94 54 E
Moscos Islands	14 00 N	97 45 E
Mu, *river*	21 56 N	95 38 E
Myittha, *river*	23 12 N	94 17 E
Namnoi, Khao, *mountain*	10 36 N	98 38 E
Negrais, Cape	16 02 N	94 12 E
Nmai Hka, *river*	25 42 N	97 30 E
Pegu Yoma, *mountains*	19 00 N	95 50 E
Pezawa Taung, *mountain*	19 33 N	94 31 E
Popa Hill	20 55 N	95 15 E
Preparis Island	14 52 N	93 41 E
Preparis North Channel	15 27 N	94 05 E
Preparis South Channel	14 40 N	94 00 E
Purian Point	15 50 N	94 24 E
Ramree Island	19 06 N	93 48 E
Salween, *river*	16 31 N	97 37 E
Shan Plateau	22 00 N	98 00 E
Sittang, *river*	17 10 N	96 58 E
Taungnyo Range	15 38 N	97 56 E
Taungup Pass	18 40 N	94 45 E
Tavoy Point	13 32 N	98 10 E
Tenasserim Yoma, *mountains*	13 00 N	98 45 E
Thaungyin, *river*	17 50 N	97 42 E
Victoria, Mount, *mountain*	21 14 N	93 55 E
Yangôn, *river*	16 29 N	96 21 E
Zami, *river*	16 09 N	97 58 E

running from north to south. The country as a whole can be divided into five landscape regions—the northern mountains, the western mountains, the eastern plateau, the central basin, and the coastal strips.

The five landscape regions

The northern mountains consist of a series of ranges forming a complex knot, with peaks of nearly 20,000 feet. It is in this region that the sources of such great rivers as the Irrawaddy, Salween, Mekong, and Yangtze arise in swirling torrents. They flow through deep gorges within a few miles of each other, separated by peaks rising sheer into the sky. It was across this wild and forbidding region that the ancestors of the present-day Burmans travelled from their original homeland in Tibet.

The western ranges, which originate in the northern mountain knot, continue southward as far as Cape Negrais, the southern tip of the Arakan Peninsula, where they run under the sea, reappearing as the Andaman Islands. Their average height is about 6,000 feet, although some peaks rise to 10,000 feet and more. They are folded mid-Tertiary ranges (formed about 40,000,000 to 20,000,000 years ago), with a core of old crystalline rocks surrounded by hard, tightly folded sedimentary rocks on either side. Forming the border between India and Myanmar, they are given different names according to locality—being known successively as the Patkoi, Lushai, Naga, Manipur, and Chin hills. The southern portion of these mountains lies entirely within Myanmar, forming the Arakan Yoma (*yoma* in Burmese means "main bone"). They separate the Arakan coastal strip from the central plain.

The Shan Plateau in the east rises abruptly from the central basin, often in a single step of 2,000 feet. Occupying the whole of the east of the country, it is deeply dissected, with an average height of 3,000 feet. The plateau was formed during the Mesozoic Era (245,000,000 to 66,400,000 years ago) and thus is much older than the western mountains. But the plateau also shows intensive folding, with north–south longitudinal ranges with heights of 6,000 to 8,600 feet rising abruptly from the plateau surface. Northward, the plateau merges into the northern ranges, and southward it continues into the Tenasserim Yoma, a series of parallel ranges with narrow valleys. The central basin, lying between the Arakan Yoma and the Shan Plateau, is structurally connected with the folding of the western mountains. The basin was deeply excavated by the predecessors of the Irrawaddy, Chindwin, and Sittang rivers; the ancient valleys are now occupied by these rivers, which cover the ancient soft sandstones, shales, and clays with their new alluvial deposits. In the deltaic regions formed by the Irrawaddy and Sittang (total area 12,000 square miles), the landscape is absolutely flat, and the monotony is relieved only by a few blocks of erosion-resistant rocks, never more than 60 feet high. The basin is cut into two unequal halves, the larger Irrawaddy Valley and the smaller Sittang Valley, by the complex folded range of Pegu Yoma.

In the centre of the basin and structurally connected with the Pegu Yoma and its northern extension is a line of extinct volcanoes with small crater lakes and eroded cones, the most impressive being Popa Hill, 4,984 feet.

The coastal strips consist of the narrow Arakan and

Tenasserim coastal plains, which are backed by the high ranges of the Arakan Yoma and Tenasserim Yoma and are fringed with numerous islands of varying sizes.

he rivers

Drainage and soils. Like the mountains, the rivers run from north to south. About two-thirds of Myanmar's surface is drained by the Irrawaddy and its tributaries. Flowing through the entire length of Myanmar, it is navigable for 965 miles. At the apex of its delta, it breaks up into a network of streams and empties into the Andaman Sea through nine mouths. Its great tributary, the Chindwin, drains the western region. The Bassein River drains the southern Arakan Yoma, and the Yangôn (Rangoon) River drains the Pegu Yoma, both entering the Irrawaddy at the delta. The Sittang flows into the Gulf of Martaban of the Andaman Sea and, in spite of its comparative shortness, has a large valley and a huge delta. The Shan Plateau is drained by the Salween, which enters Myanmar from Yunnan in China and empties into the Gulf of Martaban south of the Sittang. It is deeply entrenched and crosses the plateau in a series of deep gorges. Many of its tributaries are more than 300 miles long and enter the Salween in cascades. The Arakan coastal plains are drained by short, rapid streams, which, after forming broad deltas, flow into the Bay of Bengal. The Tenasserim plains are drained also by short and rapid rivers, which enter the Gulf of Martaban.

The highland regions of Myanmar are covered with laterite (red soil, leached of silica and containing iron oxides and hydroxide of aluminum). Protected by the forest cover, it absorbs heavy rain, but, once the forest is cleared, it erodes quickly. The lowland regions are covered with alluvial soil—mainly silt and clay. Low in potash, lime, and organic matter, it is improved by fertilizers. In the central-region dry belt, the alluvial soil develops into a black soil rich in calcium and magnesium. In the same region, however, when the soil has a low clay content, it becomes saline under high evaporation and is recognizable by its yellow or brown colour.

Climate. Myanmar belongs to the monsoon (rain-bearing wind) region of Asia, but its climate is greatly modified by its geographical position and its relief. Although the cold-air masses of Central Asia bring snow to the northern tip for two months of the year, the mountain wall prevents them from moving farther south, so that Myanmar lies primarily under the influence of the monsoon winds. The north and south alignment of ranges and valleys forces the northeast monsoon to blow from the north. When the southwest monsoon winds blow from the Indian Ocean, this alignment of ranges and valleys creates alternate zones of heavy and scanty rainfall.

Elevation and distance from the sea affect both the temperature and the rainfall; although it is a tropical country, temperatures in Myanmar are not high. The daily range of temperature is negligible everywhere, and no locality in Myanmar has a continental type of climate (*i.e.*, characterized by large seasonal differences in average temperature). Even at Mandalay, in the centre of the dry belt, the annual range is only 22° F (12° C), compared with 11° F (6° C) at Yangôn and 7° F (about 4° C) at Moulmein; the average temperature at Mandalay is 82° F (28° C), compared to 81° F (27° C) at Yangôn near the coast, 79° F (26° C) at Sittwe (Akyab) in Arakan, and 71° F (22° C) at Lashio on the Shan Plateau.

There are three seasons: the cool dry season, from October to February; the hot dry season, from March to the middle of May; and the rainy season, from May to October. The coastal regions and the mountains receive annually 200 inches (5,100 millimetres) of rain, while the delta regions receive nearly 100. The central region is not only away from the sea but also in the rain shadow of the Arakan Yoma. Rainfall gradually decreases northward until in the dry zone it is only between 20 and 40 inches. The Shan Plateau, because of its elevation, receives about 65 inches.

he forests

Plant and animal life. Even after centuries of rice cultivation involving clearing of forest, about one-half of the area of Myanmar is covered with forests of various types, depending on elevation and the amount of rainfall. Above the frost line at 3,000 feet, evergreen forests of oak and pine are to be found. In the northern mountains at heights above 6,000 feet are forests of rhododendrons. In regions with more than 80 inches of rainfall, there occur evergreen tropical forests. The trees are of hardwood but of little commercial value. In regions where the rainfall is between 40 and 80 inches, there appear monsoon forests, the trees of which shed their leaves in the hot season. They produce valuable cabinet wood, including teak. Where rainfall is less than 40 inches, the forests gradually merge into scrubland. There are no real grasslands, but, when a patch of forest has been cleared, bamboo, bracken, and coarse grass grow. In the Irrawaddy and Sittang deltas are tidal forests of mangrove trees, growing as high as 100 feet and supplying firewood and charcoal, as well as bark for tanning.

In the jungles of Myanmar are the homes of pheasants, parrots, peacocks, wild fowl, and grouse. The Asiatic two-horned rhinoceros, wild buffalo, bison, and various kinds of deer were once plentiful but are now reduced in numbers and protected. Elephants are numerous, a number of them being caught annually to be trained to work. Tigers, leopards, and wild cats are still common. Bears are found in hilly regions, and gibbons and monkeys of various kinds inhabit the thicker parts of the forests. Snakes include pythons, cobras, and vipers, and crocodiles are found in the deltas. Turtles live in coastal regions, and edible fishes abound in every stream.

Settlement patterns. *Traditional regions.* The Burmans live in the plains, and the tribal peoples in the hills. In the plains the division between northern and southern Myanmar (Upper and Lower Burma) dates from early history, not only because of differences in the geography but also because the Mons, now a small minority, lived in southern Myanmar. The division became more marked during the period from 1852 to 1885, when southern Myanmar became British Burma. The Arakan and Tenasserim coastal plains are traditional regions; the people there speak Burmese in a dialect retaining archaic features.

The villages

Urban and rural settlement. Myanmar is a land of villages. Except for Yangôn, Mandalay, Henzada, Pegu, Myingyan, and Moulmein, there are few large cities, and the towns are merely overgrown villages. The hill peoples, although practicing a shifting agriculture, have settled in villages perched near the tops of mountains and at some distance from the fields. On the Shan Plateau and the plains, the fields adjoin the villages. Older villages are circular in shape, but along the banks of the delta streams and along the railway line the villages are rectangular. Houses are built of timber and bamboo, the roofs thatched or tiled. In the past, houses were typically built on piles, the original purpose being protection from wild animals or floods. The style persists in many villages, especially those on the hills, and farm animals are kept under the houses at night. In small towns the piles have given way to a supporting brick structure with cement flooring, the upper story still being made of timber. Houses entirely of brick were few in number before 1942, but many later sprang up in Yangôn, Mandalay, and larger towns on the rubble of buildings destroyed during World War II. Life in villages and towns is still communal because of tribal custom, the influence of Buddhism, and the "classless" nature of Myanmar society. Even once-cosmopolitan Yangôn regained its native atmosphere after independence.

THE PEOPLE

Language groups

Linguistic and other groups. More than 100 indigenous languages—as distinct from mere dialects—are spoken in Myanmar, most of them by the hill peoples. The common language is Burmese, spoken by both the people of the plains and the hill peoples. All of these languages belong to only three groups. The Burmese language itself, and most of the other languages, belong to the Tibeto-Burman subfamily of the Sino-Tibetan family. The Shan language belongs to the Tai family. Languages spoken by the Mons of southern Myanmar and by the Was and Palaungs of the Shan Plateau are members of the Mon-Khmer subfamily of the Austro-Asiatic family.

Until colonial times, only the Burmans, Mons, and Shans had written languages; writing systems for Karen, Kachin, and Chin were developed later. The archaic Burmese spoken in Arakan and Tenasserim, when written down, shows no difference from the Burmese spoken in the plains. To the majority of the hill peoples, Burmese is the second mother tongue.

During the colonial period, English became the official language, but Burmese survived as the second official language. Both English and Burmese were made compulsory subjects in schools and colleges, and although a knowledge of English became an asset, and many learned to speak it, no English-speaking elite emerged. Burmese remained the language of commerce, and Burmese literature also survived. After independence, English ceased to be the official language and lost its importance in schools and colleges, although an elementary knowledge of English is still required.

Ethnic groups. The original home of the Burmans in the dry zone established the ethnic character of the entire Irrawaddy Valley and the coastal strips. These areas hold the majority of the population.

The Irrawaddy and Sittang deltas were once peopled by the Mons, who entered the country from their kingdoms in the Chao Phraya River Valley in Thailand. They were conquered in the 11th century by the Burmans, a less cultured race at the time. The Mons attempted from time to time to throw off the Burman yoke, but by the end of the 18th century they had been largely absorbed by the Burmans, mainly through intermarriage but partly through suppression. A few hundred thousand still remain in the Sittang Valley and in Tenasserim; although they still call themselves Mon, many of them have Burman blood and no longer speak their original language.

The tribal peoples

In the western hills and the valley of the Chindwin are various tribes called by the comprehensive name of Chin. The Upper Irrawaddy Valley and the northern hills are occupied by tribal groups under the comprehensive name of Kachin. The names Chin and Kachin are Burmese and indicate, respectively, "friend and companion" and "dancing friend and companion." These tribal peoples were indeed friends and companions, for they lived with the Burmans on the southeastern slopes of Tibet and followed them on the long migration into Myanmar.

The Shans of the Shan Plateau have no racial affinity with the Burmans, and their society, unlike that of the Tibeto-Burman peoples, was not democratic.

The Was and the Palaungs are probably Mon-Khmer by race, but, because of the smallness of their numbers and because of their long residency on the plateau, they are usually classified with the Shans. In the same way, the Nāgas on the Myanmar side of the Indo-Myanmar frontier are classified with the Chins, and the Lolo-Muhsos in the northeast are grouped with the Kachins.

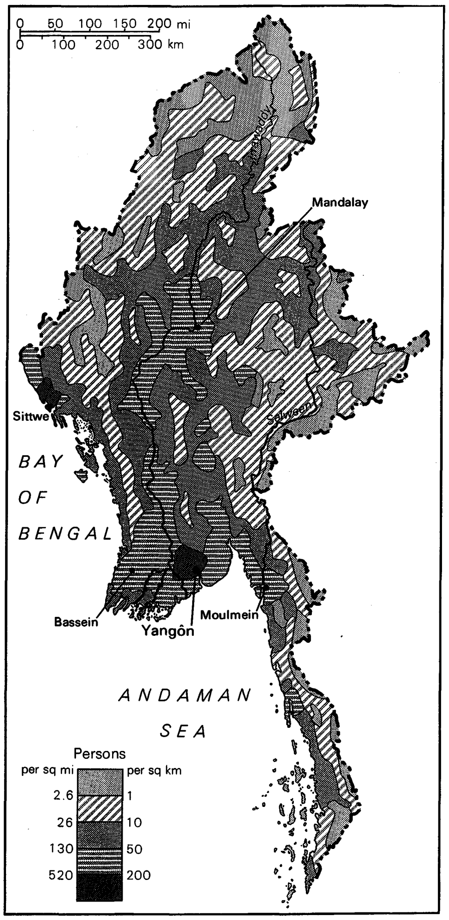
The Karens are the only hill people who have come down to the plains. Although racially and linguistically they are Tibeto-Burman, they were closely associated with the Mons. They are found in the deltas among the Burmans, on the Pegu Yoma, and along both sides of the lower Salween River.

The Kayahs, who live on the southern edge of the Shan Plateau, were known as Red Karens, or Karenni, from their red robes. Although racially and linguistically Karens, they tend to have their own identity.

The number of Indians, Pakistanis, Bengalis, and Chinese in Myanmar decreased as a result of an exodus at the outbreak of World War II and of a second exodus in 1963, when commerce and industry were nationalized.

Religious groups. The vast majority of the population is Buddhist. The Burmans are Buddhists except for minimal numbers of Christians and Muslims; the Shans are also Buddhists. Among the Karens, there are twice as many Buddhists as Christians, and the remainder are animists. The other hill peoples are animists except for a small number of Kayah Buddhists and Kachin and Chin Christians.

Demography. The density of population in each region depends upon its suitability for agriculture. The most populous regions are therefore the Irrawaddy Delta and the dry zone, with the highest densities found in the upper delta, between Yangôn and Henzada. The populations of the Sittang Delta, the sedimented hinterland of Akyab, and the regions of both sides of the lower Chindwin are moderately dense. Arakan (except the Sittwe [Akyab] region), the west bank of the Irrawaddy behind Arakan Yoma, Tenasserim, and the more inaccessible parts of the western and northern mountains and the Shan Plateau are sparsely to minimally inhabited.



Population density of Myanmar.

THE ECONOMY

Agriculture. Myanmar's economy is basically agricultural; more than two-thirds of the people derive their livelihoods directly from agricultural pursuits; of the nonagricultural workers who are employed in the other sectors of the economy, many are indirectly involved in agriculture through activities such as transportation, processing, marketing, and exporting of agricultural goods.

Myanmar may be divided into three agricultural regions: the delta, where paddy rice culture predominates; the dry zone, an area of mixed agriculture; and the hill country, where forestry and shifting agriculture have been the rule.

Although the dry zone was Myanmar's most important agricultural region in the past, it is the rice culture of the delta that now provides roughly half of the country's export earnings and the staple diet of the country's people and on which the short- and long-term health of the economy depends. About one-half of the land in agricultural use is devoted to rice culture, and, despite a climate that would permit much more extensive double-cropping, only about one-tenth of the land is actually so managed. The delta's traditional agriculture consisted primarily of rice in normal years, with the substitution of millet in drier years when there was insufficient moisture for rice; both grains yielded good returns on the alluvial soils of the delta. After Myanmar was officially annexed to British India in 1886, colonial policy called for a more commercially oriented and extensive cultivation of rice; since the indigenous labour force was insufficient to support this new economy, officially encouraged immigration of Indian and Chinese labourers and their families took place during the early decades of the 20th century. By 1942, first-generation immigrants composed about 13 percent of the total population. Despite relatively low growth in rice production after World War II, rice remained both the basic food and the basic export of Myanmar.

Agriculture in the dry zone is much more diversified than that of the delta. Its major products include, in addition to rice, other grains (wheat, millet, and corn, or maize), peanuts (groundnuts), sesame, legumes, tea, and rubber. However, to cultivate much of this land successfully, irrigation is required. The construction of the earliest known irrigation works, in Kyaukse and Minbu districts, can be dated from the 11th century, and, although their maintenance has not been a continuing concern of governments throughout Myanmar history, many are still in active service. As in the delta, the arrival of the British led to increased commercial and public works activities; British authorities renewed and extended many of these ancient systems and, in addition, built new ones during the early years of the 20th century, including the Mandalay (1903), Shwebo (1906–12), Mon (1912), and Salin (1926). Most of Myanmar's irrigated land is in the dry zone, and even here almost all is cultivated in rice. The portions of the dry zone that are not irrigated are utilized for the production of crops that are less sensitive to the seasonality or irregularity of rainfall than rice. In addition to the crops mentioned above, cotton and sugarcane are cultivated in this zone, although neither is of considerable significance. Cattle are also raised in the dry zone.

rigation

The third zone of agriculture, the hill country, occupies perhaps two-thirds of the area of Myanmar, and, although it is of much less economic significance than the other two, it is the home of many of the country's minority (non-Burman) ethnic groups. Their agriculture has been, and to a considerable extent still is, a type of slash-and-burn agriculture called *taungya,* although more sedentary traditional modes also exist and others are imposed with the advance of agricultural technology and central planning. Outside the forest areas of these highlands, the principal foods raised are rice, yams, millet, with swine and poultry ubiquitous. Possibilities, largely unrealized, exist for commercial agriculture in the more fertile river and lake basins.

rincipal
vestock

Bullocks and buffalo are used as beasts of burden, and goats, pigs, and poultry are raised for food in all parts of the country.

Fisheries. The second most important element in the diet after rice is fish—fresh, preserved, or prepared as *ngapi,* a sort of paste that is eaten with rice. Marine fisheries are not well developed, although the reported commercial catch is more than three times as great as the reported catch from inland waters. The latter, however, provides much private, noncommercial fishing in virtually every type of permanent, seasonal, or artificial body of water of any size. Two nonindigenous fish, the European carp and the tilapia of Thailand, have been imported and breed very well in impounded waters.

Forestry. Forestry is particularly important to economic planners; Myanmar has been estimated to contain up to three-quarters of the world's exploitable teak supplies. The trees are found in the mixed deciduous forests of the hills and commonly comprise about 12 percent of forests of which they are a part. The forests are state-owned, and the responsibility for their management and exploitation belongs to the State Timber Board.

Resources. Development of Myanmar's rich mineral deposits began in 1975. Deposits of silver, lead, and zinc, located in the northern Shan Plateau; tin and tungsten, found in Tenasserim; and barite, from the Maymyo area, are mined. Copper mining at Monywa began in 1982. Ruby and sapphire mines in the northern Shan Plateau were heavily exploited during colonial times. Jade is mined in the northern mountains. Petroleum and natural gas are produced to meet domestic needs. Coal of an inferior quality is found in the upper Chindwin Valley.

Manufacturing and power. There was little industrialization until after independence, when a limited program began. Yangôn, Myingyan (in the dry zone), and Arakan State were selected to become the new industrial centres. There are textile factories at Yangôn and Myingyan and one near Paleik in the central region. Oil refineries are located at Chauk, Syriam, and Mann. Yangôn also has steel-processing and pharmaceutical plants, and there is a paper mill in Arakan. Existing food-processing plants (mainly rice mills) and lumber mills have been improved and expanded. Cottage industries are encouraged by subsidies.

Oil
refining

After independence the government built a large thermal power plant in Yangôn and undertook the development of hydroelectric power plants on the Balu Chaung, a tributary of the Salween; at Taikkyi near Pegu; and in northern Arakan. The Balu Chaung plant serves Yangôn and many surrounding villages.

Administration of the economy. All large industrial enterprises, the banking system, insurance, foreign trade, domestic wholesale trade, and 90 percent of the retail trade were nationalized in 1962 and 1963. Small-scale industry (consisting mainly of food and beverage processing, miscellaneous manufacturing, and cottage industries), agriculture, and fishing were left in the private sector. Foreign investment was permitted to resume in 1973, and in 1975–76 the government placed nationalized corporations on a commercial basis, instituted a bonus system for workers, and began to accept financial aid from the Asian Development Bank and the International Bank for Reconstruction and Development.

Enterprises remaining in the private sector after nationalization account for only a small fraction of the nation's tax income. The balance is collected from the public sector. The principal sources of revenue are income, commercial, and customs taxes. There is also a land tax, applicable to the private sector only, and, applicable to both, excise, forest, mineral, fishery, and miscellaneous taxes.

The overall economic objective is the establishment of a fully socialized state. Policy measures include the diversification of agricultural production to promote exports; the substitution, wherever feasible, of essential commodities for imports; the utilization of existing industrial plants to full capacity by providing spare parts and adequate raw materials; and the assignment of priority to mining and petroleum exploration both on land and offshore. The pace of development is based on the availability of foreign exchange. Improvements in the rice market and mineral production enabled Myanmar to achieve a balance of exports and imports.

Transportation. *River transport.* The Irrawaddy River is the backbone of the transport system. Trade in rice is dependent on water transport. There has never been any need for grain elevators, as small country boats can reach any rice field that has surplus rice to sell. The Irrawaddy is navigable up to Bhamo, a distance of 875 miles from the sea; the next stretch of another 90 miles to Myitkyina is navigable only during the dry season when there are no rapids. The Chindwin is navigable for 380 miles from its confluence with the Irrawaddy below Mandalay. The many streams of the Irrawaddy Delta, with a total length of 2,000 miles, are navigable, and there is a system of connecting canals with a total length of 60 miles. The Sittang, in spite of its silt, is usable by country boats; the Salween, in spite of its rapids, is navigable up to 74 miles from the sea. Small steamers and country boats also serve the coasts of Arakan and Tenasserim.

Navigable
streams

Railways. The first railway line, running from Rangoon (Yangôn) to Pye (Prome) and built in 1877, followed the Irrawaddy Valley. At that time Pye was just a few miles from the frontier between British Burma and the Kingdom of Myanmar in the north. The line was not extended to Mandalay; instead, after 1886 a new railway from Rangoon up the Sittang Valley was constructed, meeting the Irrawaddy at Mandalay. From Mandalay it crosses the river and, avoiding the Irrawaddy Valley, goes up the Mu Valley to connect with the Irrawaddy again at Myitkyina. A short branch line connects Naba to Katha on the Irrawaddy below Bhamo.

The Yangôn–Mandalay–Myitkyina railway is the main artery, and from it there are branch lines connecting the northern and central Shan Plateau with the Irrawaddy. Other branches run from Pyinmana across the Pegu Yoma to Kyaukpadaung, and from Pegu to Moulmein to Ye.

The Pye–Yangôn railway has a branch line crossing the apex of the delta to Henzada and Bassein.

Roads. The road system was built without plan or policy, and, until independence, it was confined to the Irrawaddy and Sittang valleys, duplicating the railway route. A road goes from Pye along the Irrawaddy to the oil fields. Government policy is to improve and extend existing highways and to construct new ones.

There were originally three international roads in use during World War II—the road from Lashio to K'unming in China, the road between Myitkyina and Assam in India, and the road between Kentung in the southern Shan Plateau and northern Thailand. The road between Assam and Myitkyina has disappeared, and the other two roads have been neglected and are only in partial use.

Airways and ports. Myanma Airways runs frequent domestic flights from Yangôn and regular flights to Hong Kong, Bangkok, Dacca, Calcutta, and Kāthmāndu. Yangôn and Mandalay have international airports. Yangôn, as the terminus of road, rail, and river transport systems, is an international port with up-to-date equipment and facilities. Bassein, Moulmein, and Sittwe are also international ports.

ADMINISTRATIVE AND SOCIAL CONDITIONS

Government. *Constitution of 1974.* A constitution was approved by a referendum in December 1973 and came into force on January 4, 1974, the 26th anniversary of Myanmar's independence. General elections were held in January and February, and a new government was established on March 2, 1974, the 12th anniversary of the coup d'état that had established a military government.

The 1974 constitution provided for seven states (Arakan, Chin, Kachin, Karen, Kayah, Mon, and Shan) and seven divisions of Myanmar proper (Irrawaddy, Magwe, Mandalay, Pegu, Yangôn, Sagaing, and Tenasserim), with "local autonomy under central leadership." The supreme power was the Pyithu Hlutdaw, which exercised legislative, executive, and judicial authority. The adoption of the name Hlutdaw (or Hluttaw; literally "Place of Release") was significant. Under the Burman kings the Hlutdaw was a supreme council discharging legislative, executive, and judicial functions, but as the historical Hlutdaw was a royal institution the new designation was Pyithu ("People's") Hlutdaw. Thus, Pyithu Hlutdaw was the People's Council at the national level, as distinct from the People's Council at the division or state, township, and ward or village levels. Elections to the Hlutdaw and to the local council were held simultaneously, members being elected for four-year terms.

Pyithu Hlutdaw

Central government structure. The organs of the 450-member Pyithu Hlutdaw were the Council of State, the Council of Ministers, the Council of People's Justices, the Council of People's Attorneys, and the Council of People's Inspectors. The Council of State consisted of an elected representative from each state and division, an equal number elected by the Pyithu Hlutdaw as a whole, and the prime minister as an ex officio member. The Council of State elected its own chairman, who was ex officio president of the republic, and its own secretary. The president and the secretary were also, respectively, the chairman and the secretary general of the Burma Socialist Programme Party (BSPP). The Council of State appointed senior civil servants and deputy ministers and submitted lists of names for election by the Hlutdaw of the prime minister, the Cabinet, and the councils of justices, attorneys, and inspectors. The number of ministers could fluctuate, but the three councils had five members each. The Council of People's Justices was the equivalent of a supreme court in a parliamentary democracy; the Council of People's Attorneys was comparable to an attorney general (but members did not need legal qualifications or experience), and the Council of People's Inspectors to an auditor general.

Local government structure. The local People's Councils followed the pattern of the Hlutdaw; every council had an Executive Committee and a Judicial Committee. All but the village or ward councils also had a Committee of Inspectors. The Committee of Judges, not necessarily composed of lawyers, sat as the local court, exercising criminal and civil jurisdiction.

Political parties. Until 1988 the official political party was the BSPP. Membership was by election after preliminary service as a "friend of the party" and then as an "auxiliary." It was at first restricted to cadres, but the party held its first national assembly in 1971, when its membership was widened so that it became a national party. Civil servants and members of the armed forces, as well as workers and peasants, were members, and senior military officials and civil servants were included in the party's hierarchy.

One-party system

In September 1988 the armed forces took control of the government, creating a new ruling body, the State Law and Order Restoration Council (SLORC). All state organs, including the Pyithu Hlutdaw, the Council of State, and the Council of Ministers, were abolished. The law maintaining the BSPP as the sole political party was abrogated, and new parties were encouraged to register for general elections scheduled for May 1990. By February 1989, 233 parties had registered; of these the most important were the ruling BSPP, which changed its name to the National Unity Party (NUP), and the main opposition party, the National League for Democracy (NLD).

Armed forces. Myanmar's fighting forces consist of an army, a navy, and an air force. The Defense Service Academy is located at Maymyo. The police force, although armed and equipped and often used as a branch of the army in emergencies, remains essentially civilian in character and regional in organization; it deals with the day-to-day maintenance of law and order.

Health and welfare. In social welfare measures, the government has given special attention to workers and peasants and to the hill peoples. The National Housing Board was retained after the coup in 1962, and, in spite of a shortage of imported building materials, the housing problem was stabilized somewhat. A high mortality rate, especially among infants, has been lowered substantially, and every village has a health unit and access to a hospital.

Education. The literacy rate has always been at a high level and continues to improve. Education is free and compulsory between ages five and nine, and in secondary schools and universities fees are nominal. The University of Rangoon and its branch at Mandalay were separated in 1958 into autonomous units now known as the Rangoon Arts and Science University and the Mandalay Arts and Science University. Various institutes and colleges teach medicine, technology, agriculture, economics, etc.

CULTURAL LIFE

Myanmar's traditional culture is essentially a folk culture. Buddhism has pervaded life in Myanmar since the 11th century and has blended harmoniously with folk culture. The emphasis given by Buddhism to equality among all men and women and to the liberty and dignity of the individual fits in with Tibeto-Burman social tradition. The Feast of Merit in Chin animism, for example, has the same folk background as that of a native Buddhist almsgiving ceremony; the dancing, singing, music, and general gaiety of a Kachin sacrificial celebration are reflected in the Myanmar ceremony initiating a boy as a novice in the Buddhist order of monks; and the animals, monsters, trees, flowers, and abstract designs carved on the central post of a Kachin men's club are similar to those carved on the doors of the palace of the kings. Some observers maintain that the culture of Myanmar is animistic, others that it is Buddhist, but all agree that it is unique.

In 1886 the traditional drama appeared to be dying with the passing of the old kingdom. As it was essentially a folk drama, however, it survived and with the regaining of independence gathered new strength.

Woodcarving, lacquerwork, goldwork, silverwork, the sculpture of Buddhist images and mythological figures, and temple architecture also survived under colonial rule; there has been a revival of these indigenous art traditions under the patronage of the Ministry of Culture. The arts of bronze casting among the Burmans and of making bronze drums among the Karens, however, both disappeared. The traditional marionette show also declined, although occasionally there are attempts to revive it. The cinema is the one Western art form that has been accepted in the cultural life of Myanmar.

iterature

Burmese literature remained alive throughout the colonial period, and, both in verse and in prose, native traditions continue to prevail. A later development is biography, which has become more popular than fiction. The Burma Translation Society, which is sponsored by the government, annually awards substantial cash prizes for the best translation, the best novel, and the best biography. There are state schools of dance, music, drama, and fine arts at Yangôn and Mandalay. The National Museum is at Yangôn, and there are regional museums at Mandalay and at the state level. For statistical data on the land and people of Myanmar, see the *Britannica World Data* section in the BRITANNICA WORLD DATA ANNUAL. (M.H.Au./Ed.)

History

Myanmar lies on the western edge of the peninsula of Indochina and forms part of the region of Southeast Asia. In prehistoric times it was traversed, both along its coasts and down its river valleys, by band after band of migrating peoples on their way farther eastward and southward. During the first 10 centuries of the Christian Era the overland trade route between China and India passed through Myanmar's borders, and merchant ships from India, Sri Lanka, and even farther west converged on its ports, which were also the termini of the portage routes across the narrow isthmus from what is now the Gulf of Thailand. Thus, Myanmar was the gateway of Southeast Asia.

The Indian merchants brought with them not only precious cargoes but also great cultural gifts in the form of religious, political, and legal ideas; and within a few decades Indian cultural traditions had remolded native society, native thought, and native arts and crafts. Although the new cultural treasures passed through to Southeast Asia, Myanmar never allowed its own native culture to be overwhelmed.

Surrounded on three sides by mountain walls and on the fourth by the sea, Myanmar has always been insular; as a consequence its culture, in spite of the many Indian influences and in spite of its close affinity with the cultures of the other countries of Southeast Asia, has remained distinct and separate. For example, it rejected the Hindu Indian theory of a divine king and insisted that men and women are born free and equal.

The first area in Southeast Asia to receive Buddhism, because of its proximity to India, Myanmar became the centre of Theravāda Buddhism almost overnight in the 11th century, when the faith was being assailed not only in India and Southeast Asia but in the whole of Asia itself. The Buddhist doctrine of the equality of all men and women was superimposed with grace and harmony on the native tradition of social equality.

The unique nature of Myanmar society and culture was noticed at the time of the British conquest (late 19th century) even by humble privates, whose conception of Burma as a "cleaner, greener land" was embodied by the British author Rudyard Kipling in "Mandalay." The historian G.E. Harvey wrote in *British Rule in Burma* that "there is, in Burmese life, not only a beauty that delights the eye but also a dignity which makes one proud of the human race."

MYANMAR UNTIL THE ADVENT OF BRITISH RULE (TO 1885)

The origins of civilization in Myanmar. The Irrawaddy River, flowing southward through the entire length of Myanmar, divides the country in two, and its valley forms the central plain. The first human settlements in Myanmar appeared some 5,000 years ago in the middle portion of the Irrawaddy Valley. The stone and fossilized-wood tools used by those Paleolithic and Proto-Neolithic people have their own distinguishing features, and anthropologists have given their culture its own name—Anyathian. A discovery in 1969 by workers from the government's Department of Archeology of some cave paintings and stone tools in the eastern part of Myanmar's Shan State shows that that area, too, had Paleolithic and Proto-Neolithic settlements. Crude shards and ring stones found at the site appear to have been attached to stonecutting tools, to make them more suitable for digging; and the woodcutting tools in the find were probably used in clearing patches of the forest for cultivation, which would indicate that the shift from gathering to agriculture had already begun.

he first iman set- ements in yanmar

The Anyathian people probably migrated down the Irrawaddy River, and the Shan cave people down the Salween. Thus, even during the stone ages, the general pattern of human migration and settlement for both Myanmar and the rest of mainland Southeast Asia had been fixed—tribal groups coming from the north over the mountains and then down the great rivers: the earlier or stronger groups settling in the valley to learn the art of wet-rice cultivation and the later or weaker groups being pushed into the mountainous parts to learn only the method of slash-and-burn rice growing. In those early times the Irrawaddy had not yet formed its great delta; and at its huge mouth and along the coast, the country was stormy, wild, and desolate.

The Stone Age Anyathians stayed far from the sea. The earliest arrivals on Myanmar's coast were Negritos, a fact remembered in folktales, which tell of dark-skinned, short-statured, fuzzy-haired, and human-flesh-eating ogres who dwelled at the mouths of great rivers and on islands in the sea; a Buddhist legend describes how monk-missionaries sent from India saved the people of Thaton from ogres who lived on nearby islands and raided the city in quest of human flesh. As to the Anyathians, even folk memory no longer knows them, and it cannot be said whether they died out or merged with the later arrivals in the Irrawaddy Valley.

Origin of the Mons

The Mons. The first people in Myanmar to leave definite traces of their settlement were the Mons. Speaking an Austro-Asiatic language, they were close cousins of the Khmers, with whom they originally came down the Mekong River. Later the Khmers followed the river southeastward to what is now Kampuchea (Cambodia), and the Mons veered southwestward toward the source of the Chao Phraya River (Mae Nam Chao Phraya) and went down its valley. By at least the 3rd century BC the Mons not only occupied the entire Chao Phraya Valley but had also spread to the Sittang Valley, doubtless through the gaps in the eastern mountain wall of Myanmar. Their port capital was Thaton, not far from the isthmian portage routes; and through this window to the sea they saw India, in its full glory, united and peaceful under the emperor Aśoka and a flourishing centre of Theravāda Buddhism.

Introduction of Indian culture. Aśoka sent a mission of Buddhist monks to Suvarnabhūmi, "the Golden Land"; although a few scholars still remain unconvinced, it seems definite that the Golden Land was the three Mon regions of Upper Chao Phraya, Lower Chao Phraya, and the Sittang Valley. The ancient monastic settlement of Kelasa, situated a few miles from Thaton and claimed by Burmese and Mon chronicles to have been founded by Aśoka's missionaries, was mentioned in early Sinhalese records as being represented at a great religious ceremony held in Sri Lanka in the 2nd century BC.

With the expansion of Indian commerce in Southeast Asia in the 1st century AD, Thaton's prosperity and importance increased. The expansion was sudden and revolutionary but peaceful, and the Indian merchants and seamen came to Thaton as friends rather than as conquistadores or colonists. Their numbers were never great, and their settlements were only temporary. Until the 8th century, when there was a scramble for the new lands of the Irrawaddy Delta, no conflicts developed between the Indians and the Mons. As a result, the Indians' cultural gifts were found acceptable.

Blending of Indian and native cultures. The Mons saw to it that their native culture was not abandoned or displaced, and they worked for a harmonious blending of the old and new cultures. They brought many of their native animistic beliefs into the fold of Theravāda Buddhism. They enhanced the power and prestige of their king by adopting the Hindu ritual of coronation. They developed a new art of sculpture by blending the native tradition of wood carving with the Greco-Indian conventions of making images of the Buddha. They built *stūpa*s on the Indian model and then developed new forms of temple architecture by a mixture of native and Indian traditions.

Mon culture

Within a few decades the Mons became the most advanced people in Southeast Asia, and they assumed the role of teachers to their neighbours, spreading Theravāda Buddhism and their new culture over the entire region. Their cousins, the Khmers, were the first to benefit, followed by the Burmans. Even in the 13th century, when their glory had passed and they were a conquered people in the Chao Phraya Valley, the Mons freely shared their cultural heritage with the new arrivals, the Tais.

The Tibeto-Burman invasions. About the time the Mons were reaching the Chao Phraya Valley, some Tibeto-Burman tribes were leaving their homeland on the southeastern slopes of the Tibetan Plateau. They had reached quite a high level of culture, but their way of life and their social values of individual equality were being threatened by the rival kingdoms of Tibet and China. Preferring physical hardship to political bondage, they started a long trek across the northern mountains of Myanmar; they entered the Upper Irrawaddy Valley, perhaps by the time of the Buddha (*c.* 500 BC). By Aśoka's time their spearhead, the Pyus and allied tribes, had founded their city-kingdoms, perhaps at Tagaung and certainly at Halingyi. They gradually moved southward, leaving along the way settlements that later became city-kingdoms. They stopped for a while at Taungdwingyi, and some tribes crossed the river and the mountains to Arakan. They finally reached the vicinity of modern Prome, only a few miles from what was then the apex of the Irrawaddy's mouth. At about the time of Christ, they founded the city of Śrī Kṣetra near modern Prome.

The Pyu state. There had long been a trade route between China and India that passed through northern Myanmar and then across the Chindwin Valley. The Pyus had gained full control of this route and had even provided an alternative routing—down the Irrawaddy to Śrī Kṣetra and then by sea to India and also to the portage routes and the islands. Because the Pyus were in control of the whole Irrawaddy Valley, Roman embassies to China in AD 97 and 121 chose the overland route through Myanmar for their journey.

Doubtless in the beginning the Pyus learned much from the Mons, but later they came into direct contact with Indian culture. They were so prosperous and powerful that they became the overlords of the Mon kingdom of Thaton and the portage routes. Chinese historical records noted that the Pyus claimed sovereignty over 18 kingdoms, many of them to the south of Myanmar.

Pyu life

The same Chinese records emphasized the humane nature of Pyu government and the elegance and grace of Pyu life. Fetters, chains, and prisons were unknown, and punishment of criminals was a few strokes with the whip. Gaily dressed in blue, the men wore gold ornaments on their hats and the women wore jewels in their hair. The Pyus lived in houses built of timber and roofed with tiles of lead and tin; they used golden knives and utensils and were surrounded by art objects of gold, green glass, jade, and crystal. The walls, the palace, and the monasteries were of glazed brick. Amid this luxury, the Pyus were devout Theravāda Buddhists. Their architects evolved the vaulted temple, which later found its golden age at Pagan. Their sons and daughters were disciplined and educated in monasteries or convents as novices.

Despite their power over the neighbouring regions and the elegance of their life, the Pyus remained a loosely knit medley of tribes. Retaining their old tribal values, the Pyus remained more democratic than the Mons, who possessed the institution of a hereditary chieftain (later king) and a hereditary nobility; but as a consequence, the Pyus were less united and disciplined. There were conflicts between Pyu tribal groups, resulting in changes of dynasties at Śrī Kṣetra.

The huge delta became silted up, and the Pyu city swiftly became inaccessible from the sea. The Mons—politically organized as the confederacy of Ramanna and embracing the three kingdoms of Thaton, Dvaravati, and Haripunjaya—asserted their claim to the newly emerged island of Pegu, in the mouth of the Irrawaddy, by attacking both the Pyus and the Indian merchant ships in the vicinity. In the 6th century a tribal war forced the Pyus and their loyal allies to abandon Śrī Kṣetra. There was a fresh merger of tribes, and in the process the Burmans, a people subject to the Pyus, became prominent for the first time.

Advent of the Pagan Burmans. The Pyus and their allies moved to northern Myanmar, where they set up a new capital. The Burmans, still loyal, established a small settlement of their own. Early in the 9th century the Tai-Shans from their kingdom of Nanchao in South China mounted a series of raids on the cities of Indochina and even besieged Hanoi. The Mon and Khmer cities held firm, but the Pyu capital fell. The leadership of the Tibeto-Burman tribes passed to the Burmans, and in 849 they founded their own city of Pagan and built a wall around it.

The founding of the city of Pagan

The Pagan kingdom (849–1287). By that time the Mons had become supreme in southern Myanmar. They had occupied the whole of the Irrawaddy Delta, building the port of Bassein in the west and founding the city of Pegu in the centre. They could have stepped into the vacuum caused by the destruction of the Pyu kingdom, but they were not politically ambitious and perhaps did not relish going up the river into arid country.

The unification of Myanmar. The Mons' reluctance allowed the infant Burman kingdom to survive and grow. The Burmans had learned much from the Pyus, but they were cut off from the fountain of Buddhist culture, now transferred to Sri Lanka. Theravāda Buddhism had disappeared from India, and in its place were Mahāyāna Buddhism, its offshoot Tantrism, and a resurgent and aggressive Hinduism. The Burmans still controlled the overland route; along it from China and Nanchao came Mahāyāna Buddhism, and from Bengal came Tantrism. There was also a return to native animistic beliefs.

In 1044 Anawrahta came to the throne at Pagan, and his accession coincided with the seemingly final defeat of Theravāda Buddhism in Southeast Asia; the islands and the Khmers went over completely to Hinduism, and the Theravāda Mons were pressured by their neighbours into accepting certain Hindu beliefs. An orthodox monk, Shin Arahan, fled in protest to Pagan, where he was able to convert the young king to the Theravāda faith. The people also accepted it because there was still a vague remembrance that the Pyus had been Theravādins and because the faith's emphasis on the equality and dignity of the individual fitted their society and their prevailing mood. Anawrahta, declaring himself the champion of Theravāda Buddhism, challenged and subdued the Mahāyānist kingdom of Nanchao, annexed the animistic Shan Plateau, and conquered the Theravāda Mons. He united the whole of Myanmar into a single kingdom and founded the Pagan dynasty. His work was continued by his great commander Kyanzittha, who was elected king; and the latter was followed by another great ruler, his grandson (Anawrahta's great-grandson), Alaungsithu.

Anawrahta's subjugation of the Shans checked the flood of Tai-speaking mercenaries and migrants into Southeast Asia, and the strength of his kingdom brought peace to the region, now divided into two great empires, Khmer and Burman. Anawrahta's dynasty of kings lasted until the 13th century. By that time, their great temples had been built and their message of Theravāda Buddhism had been carried not only to the Shans but also to the Khmers.

Destruction by the Mongols

A malaise seemed to have set in by the middle of the 13th century. Then came a demand for submission and tribute by the Mongol Kublai Khan. The proud people of Pagan refused and made a valiant effort to stem the tide of invasion that followed, but it was unavailing, and Pagan fell to the Mongol armies in 1287.

Cultural changes in the Pagan period. Pagan was a fabulous kingdom even to its contemporaries, and Marco Polo was impressed by its splendour. Its many temples and monasteries testified to the prosperity of its people and the richness of its culture. The conquest of the Mon kingdom of Thaton was the foundation of both its economy and its culture. All the ports of the country were in Burman hands, and Pagan's temples were built by free artisans, who were paid their wages not only in money but also in kind and whose clothing, shelter, health, comfort, and safety were the responsibility of their employers (as evidenced by contemporary inscriptions, giving the monetary details connected with the construction and endowment of a temple).

Mon craftsmen, artisans, artists, architects, goldsmiths, and wood carvers—captured at Thaton—were taken to Pagan to teach their skills and arts to the Burmans, whose artistic traditions, inherited from the Pyus, had been lying dormant. They learned quickly and soon were able to stand shoulder to shoulder with Mon and Indian craftsmen. Mon monks and scholars taught the Burmans the Pāli language and the Buddhist scriptures, and the Burmans soon became scholars themselves, making Pagan the centre of Theravāda learning. Some of their religious commentaries came to be accepted as part of the Pāli canon by all Theravādins. The women of Pagan took part in all these activities, their freedom and equality doubly guaranteed by tribal usage and by Buddhist doctrine.

Devout and orthodox, the people of Pagan made Buddhism their way of life but retained those animistic beliefs that were found compatible with Buddhism. They set a pattern of religion, government, and society that later generations followed almost without change. Temple frescoes show the Pagan people to have been simply dressed. Their utensils and implements were not of gold, and so their life was not so elegant as that of the Pyus.

The period of Shan dominion (1287–1531). During the reigns of the last kings of Pagan, many Shans entered royal service as mercenaries. After the fall of Pagan these Burmanized Tai-speaking soldiers of fortune drove out the Mongols; but instead of restoring the fallen dynasty, they founded small kingdoms for themselves. Before invading Myanmar, Kublai Khan had conquered Nanchao and turned it into the Chinese province of Yunnan; when his armies had withdrawn, he encouraged the Shans to leave their old homeland and attack the Burmans and the Khmers. The consequence was the rise of a number of Tai-speaking kingdoms in the Chao Phraya Valley and the restoration of the kingdom of Pegu by the Mons of southern Myanmar. In 1368 northern Myanmar was unified under one kingdom, that of Ava, whose kings were half-Burman and half-Shan and regarded themselves as Anawrahta's successors. They encouraged scholarship and learning, making their period a great age of Burmese literature. They wasted their resources, however, by waging a long war against the Mons.

The Shans of the Chao Phraya Valley fought both Pegu and Ava. In the meantime the small settlement of Burman refugees at Toungoo, on the Pegu–Ava frontier, sought survival by playing Mon against Shan. The Shan refugees from Nanchao who were waiting in the north of Myanmar for an opportunity to win back their old kingdom decided to conquer Ava instead. They had not been won over to Buddhism; and in 1540 there was a massacre of monks, resulting in an exodus of Burmans to Toungoo. In the meantime the Mons at Pegu were achieving another golden age under wise rulers. Under Dhammazedi, Pegu became the centre of Theravādin scholarship and it also entered into close commercial relationship with Malacca, on the Malay Peninsula, an Islāmic kingdom before 1511 and a Portuguese possession thereafter.

The Toungoo dynasty (1531–1752). In 1531 Tabinshwehti became king of Toungoo, and within a few years he conquered northern Myanmar from the Shans and southern Myanmar from the Mons. Transferring his capital to Pegu, he made a great attempt to unite Burmans, Mon, and Shan into a single nation. He died in 1551 and was succeeded by his brother-in-law Bayinnaung. The Shans in the Chao Phraya Valley had consolidated their power under the kingdom of Ayutthaya (Siam), and they cast their eye on the Shan Plateau and southern Myanmar. Bayinnaung marched on to Ayutthaya and conquered the entire Chao Phraya Valley. He used Portuguese mercenaries and Portuguese cannon, but he was determined to keep the Portuguese out of mainland Southeast Asia.

.ule of
ayin-
aung

Bayinnaung's wars had exhausted Myanmar's resources, and after his death the kingdom broke up. Manipur (which had been subjugated in 1560) declared itself a free kingdom. Then the Mons revolted, the Siamese not only regained their independence but also ravaged southern Myanmar, and the Portuguese founded a kingdom at Syriam. Bayinnaung's grandson, Anaukpetlun, by 1613 overcame all three and reunited the kingdom. His successor, Thalun, was peaceful, however, and he abandoned the policy of expansion and moved the capital back from Pegu to Ava. There were no more Mon rebellions, and the kingdom remained at peace. Eventually, the kings became weak and power passed to the ministers.

The fall of the Ming dynasty in China (mid-17th century) resulted in occupation of Myanmar territory by the armies of the fallen emperor and also by the pursuing armies of the new emperor. The raja of Manipur sent his raiders into Myanmar for loot. The Mons rose in rebellion, encouraged by the French in India. Assailed from all sides, Ava fell to the Mons in 1752; and the whole of Myanmar passed under Mon rule.

The Alaungpaya dynasty (1752–1885). It was soon proved, however, that only the king and his ministers had been defeated, not the Myanmar people; before the year had ended, a popular leader, Alaungpaya, drove the Mons out of northern Myanmar and regained the Shan States. By 1758 he had regained Manipur and defeated the Mons and their French garrisons. The Thai became alarmed and attempted to rouse the Shan chiefs to rebel. Alaungpaya retook Tenasserim, the site of the old portage kingdoms, and invaded Thailand. Although his invasion failed and he himself died during the retreat in 1760, the Myanmar now felt that, unless Thailand were conquered, their kingdom would be insecure. The defeats suffered in 1752 had embittered them, and they decided to pursue a policy of aggrandizement and repression. Alaungpaya's son and successor, Hsinbyushin, sent his armies into Thailand in 1766; they captured Ayutthaya in 1767. China, alarmed over the growing power of Myanmar, invaded the country four times during the period 1766–69, without success.

Myanmar then reconquered Arakan and occupied Assam, thus coming face to face with the British power in India. The result was the First Anglo-Burmese War (1824–26), in which Thailand fought on the British side. Myanmar eventually had to sue for peace and lost Assam, Manipur, Arakan, and Tenasserim.

The Anglo-
Burmese
wars

The Second Anglo-Burmese War (1852) was provoked by the British, who wanted to close the gap in their coastline stretching from Calcutta to Singapore; it resulted in the British annexation of Pegu Province. As the British became more and more interested in trade with China through its back door, they waited for a suitable pretext and in 1885 declared war on Myanmar for the third time. To meet international criticism of their action, the British gave the excuses that the last independent king of Myanmar, Thibaw (ruled 1878–85), was a tyrant and that he was intriguing to give France greater influence over the country. Neither of these charges seems to have had much foundation.

The administration of traditional Myanmar. The king was the chief executive and the final court of appeal, but there were checks on his power. He could not make laws, only edicts that would lapse with his death. Custom was the only recognized source of law, and even the king could be sued for civil wrongs. In addition, following the old tribal requirement that each chieftain was to be elected by the council of elders, the king had to be elected by the previous king's council of ministers. With the advent of strong kings, generally the election was merely a formality; but in times of crisis there were real elections.

The king was also the patron of Buddhism and was expected to act as a humane ruler. Buddhist monks were not organized into a church, but there was a primate who, although appointed by the king, often proved to be his sternest critic. Although monks were outside the sphere of politics, they had the right to give sanctuary in that they could ask for the life of a condemned criminal. In practice this right was exercised only when there was some miscarriage of justice. Because monasteries also served as schools for all children, the monks were the teachers of the people and molded public opinion regarding the king's justice.

As a corollary to the long-held belief that all men and women were born free and equal, there were no hereditary nobles. The king's officials were appointed, at least in theory, on their character and talent alone; and their appointments lapsed with the king's death.

The centre of government

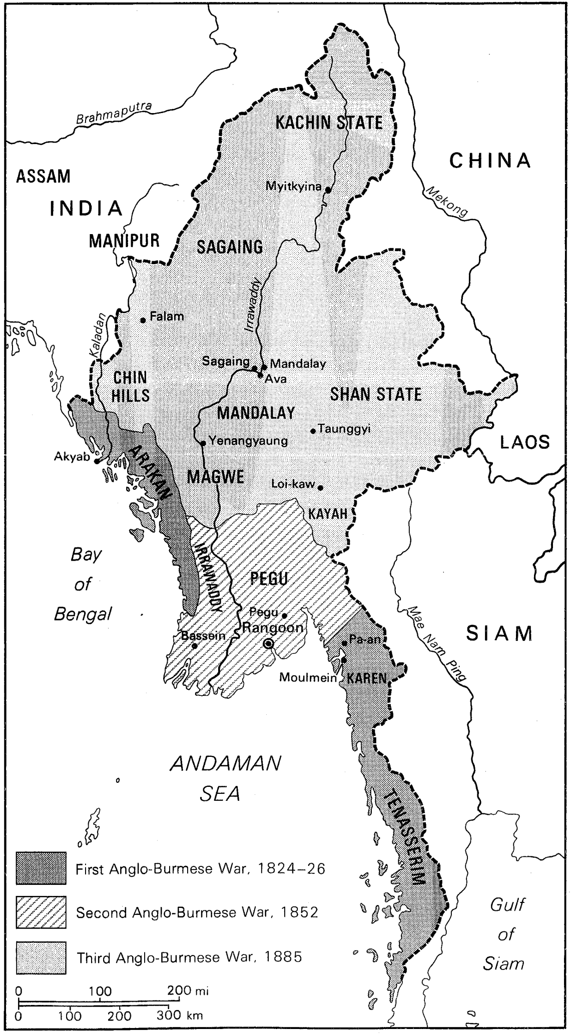
The High Court of Justice was the centre of government. It had three wings—the treasury, the executive, and the judiciary. The judicial wing was the final court of appeal; in theory and often in practice the king presided over its deliberations. All proclamations and appointments that were made by the king became valid only when orders giving effect to them were issued by the executive wing.

Every province had a governor, to whom were delegated certain powers by the High Court of Justice; but there was always a right of appeal against all his decisions to the High Court. Local government was in the hands of "group-village headmen." Villages were grouped together according to their size and geography. Each village in a group was under an elected elder, and these elders together elected the headman, who was then given his appointment by the king. The institution of headman was almost hereditary: because of the need for continuity and experience, the elders usually gave preference to the son, brother, daughter, or even the widow of the former headman. The headman, like a governor, had the right of immediate audience with the king.

THE BRITISH IN BURMA (1885–1948)

The Third Anglo-Burmese War lasted but a week; the Myanmar people never expected that hostilities would break out, and, realizing the hopelessness of the situation because of the disparity in arms, they offered no resistance. They also believed that the British aim was merely to replace King Thibaw with a prince who had been sheltered and groomed in India for the throne; and this belief seemed to be confirmed when the British commander called upon the High Court of Justice to continue to function. The British finally decided, however, not only to annex the whole kingdom, which they called Burma, but also to make it a province of India (effective January 1, 1886).

Adapted from F. Trager, *Burma, From Kingdom to Republic* (1966); Praeger Publishers, Inc.



British territorial acquisitions in Burma.

Anti-Indian feelings

This was a bitter blow to the Myanmar. They had always been prejudiced against the Indians, calling them *kala* ("caste people"). In addition, they remembered that, in Myanmar's three wars with the British, Indian troops had fought for their British masters. It was true that, after the war of 1824, Arakan and Tenasserim had been placed under British commissioners responsible to the British government of India and that, after the province of Pegu had been annexed in 1852, the whole territory of British Burma had been placed under a chief commissioner also responsible to the government of India. But these were temporary and makeshift arrangements, to be reconsidered in the light of experience. Burma seemed too big a territory to be attached to India, and moreover the people were different from the Indians in race, language, society, way of life, attitude, and temperament. The Buddhist culture the Myanmar had received from India had been modified and transformed so that it had become Myanmar-Buddhist culture, while in India Buddhism had been suppressed and discarded many centuries before.

The effects of colonialism. The Myanmar refused to accept the British victory as final, and they resorted to guerrilla warfare against the British army of occupation. There were spontaneous uprisings all over the country, led by former officers of the disbanded royal army, former officials (including village headmen), and princes of the blood. They considered themselves soldiers still fighting the Third Anglo-Burmese War; but to the British the war had ended with the annexation of the kingdom and the fighters were rebels and bandits. For the next four years, although no martial law was proclaimed, the British military officers acted as both judge and jury in dealing with captured guerrillas. Villagers who aided the rebels were also sternly punished. There were mass executions and even cases of wholly massacred villages.

British attempts to end the rebellion

As the guerrillas fought on, the British tried to force the people into submission. Villages were burned; families who had supplied villages with their headmen were uprooted from their homes and sent away to Lower Burma (the southern part of Myanmar, which had been under British control since the second Anglo-Burmese war). Strangers were appointed as headmen for the new villages the British had set up. The guerrillas resorted to cruel and desperate measures against the new village officials. Finally, the might of the British army prevailed, and by 1890—with thousands of guerrillas killed or executed—the struggle was over.

The British did not wish to favour any one religion over another, and they were reluctant to patronize Burmese Buddhism. Queen Victoria, as empress, had proclaimed strict British neutrality in religious affairs of the Indian peoples. The patronage of Burmese Buddhism did not mean financial support, of which Burmese Buddhist monks were in no need (they could own no property, and such simple needs as robes, food, books, and medicine were met by donations from the public). Patronage involved authority; and although the king had been required to kneel before even a freshly ordained monk visiting him, it had been the king's right to appoint the primate, who exercised supervision and discipline among the ranks of the clergy throughout the kingdom. The king had also had the right to attach two royal officials to the primate—the commissioner of ecclesiastical lands and the ecclesiastical censor. The duty of the commissioner had been to see that ecclesiastical lands were exempted from payment of taxes; but at the same time, he had seen to it that bogus and illegal endowments did not escape taxation. The duty of the censor had been to maintain a register of monks, which had given the king an indirect control over the clergy. The power of disrobing a wayward monk had lain only with the primate, but the same result could have been achieved by the censor scratching the monk's name from the register. This arrangement was designed to prevent unworthy persons from taking advantage of the people's respect and devotion for the clergy.

ducation
nder the
ritish

The British refusal of a plea by the clergy and the elders to continue this arrangement resulted in the appearance of pseudo-monks, who lowered the prestige of the Buddhist clergy and contributed to the failure of an attempt by Sir Arthur Phayre (chief commissioner of Lower Burma, 1862–67) to build a modern educational system using the monastic schools as its foundation. Phayre's successors recommended the foundation of government schools, but only a few were actually established. The government of India preferred to encourage foreign Christian missions to found schools by offering them "grants-in-aid."

Many mission schools were founded to which parents were forced to send their children because there were no alternatives. Because the teachers were missionaries, the lessons they gave were marked by repeated criticism of Buddhism and its culture. In the government schools the first teachers, British and Indian, were mere civil servants, unable and unwilling to continue the older traditions.

The British impact on the Burmese economy proved disastrous. This was not so much due to the usual economic exploitation of a country by a foreign ruler (the British were milder in this respect than were other colonial powers) as it was to the British policy of laissez-faire and to Burma's being made an Indian province. The British dream of a golden road to China through Burma could not be realized, but the opening of the Suez Canal in 1869 was equally golden, for it increased international demand for Burma's rice, which had never before been an important item of export. This development came in the wake of a change in Burma from a barter economy to a money economy. The Burmese peasant was both psychologically unprepared for and inexperienced in the use of cash; and under Burmese law, capital accumulations were almost impossible because the Burmese possessed no right of disposing property by will, so that a family estate had to be divided after the death of the parents among all sons and daughters.

The Irrawaddy Delta was swiftly cleared of its mangrove forests to become covered with rice fields. Even in 1857 the price of rice increased 25 percent; by 1890 the 1857 price had more than doubled and it continued to increase until the world depression of the 1930s. Tempted by the prevailing rice prices, the Burmese flocked to the delta; but in order to prepare the land for cultivation they had to borrow capital from Indian moneylenders from Madras at an interest rate of some 120 percent. The British banks would not grant loans on mortgage of rice land, and the British government had no policy for establishing land-mortgage banks or for making agricultural loans. Prevailing prices were high in the international market, but the local price was kept down by a handful of British firms that controlled wholesale trade and by Indian and Chinese merchants who controlled retail trade. Although the Burmese economy developed rapidly from 1890 to 1900, the Burmese people did not benefit from it. A railway had been built through the entire valley of the Irrawaddy, and hundreds of steamboats plied the entire length of the river; but the railway and the boats belonged to British companies. Roads had been built by the government, but they were meant for the swift transport of troops during frequent rebellions. A British company worked the ruby mines until they became nearly exhausted. The extraction of oil and timber was monopolized by two British firms. The balance of trade was always in favour of Burma; but this was only on paper because there was no control or even record of foreign exchange remittances to their home countries by the British and Indian troops; the British, Indian, and Chinese merchants; the British and Indian civil servants; and the Indian labourers who came in thousands without any restrictions because they were merely travelling from one part of the Indian empire to another. With land values and rice prices soaring, the Indian moneylenders foreclosed mortgages at the earliest opportunity. The dispossessed farmers could not find employment even on their lost lands because, with a higher standard of living, they could not compete with Indian labourers. Burmese villagers, unemployed and lost in a disintegrating society, took to petty theft and robbery and soon acquired the reputation of being lazy and undisciplined.

The new
Burmese
leaders

The emergence of nationalism. Those Burmese who attended the new schools managed to gain admission to the clerical grades of government service, but even in those lower grades there was competition from Indians. Because science courses were not available, the professions of engineering and medicine were closed to the Burmese. Those who advanced to the government liberal arts college at Rangoon entered the middle grades of the civil service; a few went on to London to study law. When these young barristers returned to Burma they were looked upon by the people as their new leaders. Their sojourn in the liberal atmosphere of London had convinced these new leaders that some measure of political independence could be regained by negotiation. They first gave their attention to the national religion, culture, and education; in 1906 they founded the Young Men's Buddhist Association (YMBA) and established a number of schools supported by private donations and government grants-in-aid (the movement was not antigovernment). That same year the British, attempting to pacify the Indian National Congress, introduced some constitutional reforms in India. Only a few inferior changes were made in the Burmese constitution, but these confirmed the young leaders' faith in British liberalism. In 1920, however, when it was found that Burma would be excluded from new reforms introduced in India, the barristers led the people in a nationwide protest, which involved a boycott of British goods. In 1886, to humiliate the Burmese, the British had made Burma a part of India; they now found that they could not exclude Burma from sharing benefits granted to India.

The following December, Rangoon College was raised to the status of a full university; but because its control was vested in a body of British professors and government nominees, its students went on strike. Younger schoolboys and schoolgirls followed suit. The strikers camped in the courtyards of monasteries, reviving memories of days when education was the concern of the monks. The general public and the Buddhist clergy gave full support to the strike. The University Act was amended, and the strike was settled; but many strikers refused to go back to mission and government schools. The YMBA schools, now calling themselves "National" schools, opened their doors to the strikers.

The
Thakin
movement

Constitutional reforms were finally granted in 1923; but the delay had split the leaders, some of whom, like the masses, were beginning to doubt whether political freedom could be attained by peaceful protest. In the University of Rangoon itself, students began to resent their British professors. A radical student group began organizing protests, which came to be known as the Thakin movement (a purposely ironic name, taken from *thakin,* "masterly," by which Burmese were required to address the British).

(M.H.Au./Ed.)

In 1931 Burmese peasants, under their leader Saya San, rose in rebellion. Armed only with swords and sticks, they resisted British and Indian troops for two years. The young Thakins won the trust of the villagers and emerged as leaders in place of the liberals. In 1936 university students again went on strike, and two of their leaders, U Nu and Aung San, joined the Thakin movement. In 1937 the British government separated Burma from India, but the masses interpreted this as proof that the British planned to exclude Burma from the next phase of Indian reform.

When World War II broke out in Europe in 1939, the Burmese sympathized with Britain, but they wanted to bargain with the government before giving their support. A warrant was issued for the arrest of Aung San, but he escaped to China, where he attempted to contact radical groups for support. Japanese assistance was offered instead; he returned to Burma in secret, recruited 29 young men, and took them to Japan, where the "Thirty Comrades" received military training. The Japanese promised independence for Burma; hence, when Japanese troops reached Bangkok, Aung San announced the formation of the Burma Independence Army. Independence was not declared, however, and Aung San and his army began to conspire against Japan. When the Japanese administration learned of the plot, they disbanded the Independence Army and formed a smaller Burma Defense Army, with Aung San still as commander.

Ba Maw, the first prime minister under the 1937 constitution and later the leader of the opposition, was appointed head of state, with a Cabinet including Gen. Aung San and U Nu. In 1943, when the tide of battle started to turn against the Japanese, they declared Burma a fully sovereign state. But the Burmese government was still a mere facade, with the Japanese army ruling. Aung San contacted Lord Mountbatten, the Allied commander in Southeast Asia, offering cooperation, and in March 1945, on receiving Mountbatten's approval, Aung San and his army joined the British side.

The AFPFL

After the liberation of Rangoon, Aung San and the Thakins formed a coalition of all parties under the name Anti-Fascist People's Freedom League (AFPFL). At this critical moment the British military administration showed its old prejudices. But Lord Mountbatten hastily sent a major general, Sir Hubert Rance, to head the administration; and Rance, with a conciliatory attitude, regained for the British the trust of Aung San and the general public. When the war ended, the military administration was withdrawn and Rance was replaced by the former civilian governor, who formed a Cabinet consisting of older and more conservative politicians. The new administration arrested Aung San on a criminal charge. Surprised and angered, the Burmese people prepared for rebellion; but the British government in London wisely replaced the governor with Rance, now retired from the army. Rance formed a new Cabinet, including Aung San, and discussions for a peaceful transfer of power began. These were concluded in London in January 1947, when the British agreed to Burma's independence and its leaving the Commonwealth.

Independence

The Communist and conservative wings of AFPFL were dissatisfied with the agreement. The Communists broke away and went underground, and the conservatives went into opposition. In July Aung San and most members of his Cabinet were assassinated by gunmen sent by U Saw, a former prime minister and now a conservative. Rance asked U Nu to form a new Cabinet. On January 4, 1948, Burma became a sovereign independent republic.

MYANMAR SINCE INDEPENDENCE

Constitutional government, 1948–62. With its economy shattered and its towns and villages destroyed during the war, independent Myanmar needed some years of peace; a foreign policy of neutrality was decided upon, but because of internal strife no peace resulted. The Communists were the first insurgents, followed by Aung San's veterans, and then the Karens, the only ethnic minority on the plains; but the other minorities—Chins, Kachins, and Shans—who had been ruled separately by the British but who had enthusiastically joined the Union, stood firm.

A division of Chinese Nationalist troops occupied parts of the Shan Plateau after China had fallen to the Communists in 1949; and, because of the United States' general support for Nationalist China, Myanmar stopped accepting American aid and became prejudiced against all foreign aid. At the UN Myanmar endeavoured to show impartiality. It was one of the first countries to recognize Israel and also the People's Republic of China.

In 1958 Myanmar was well on the road to internal peace and economic recovery when the ruling AFPFL was divided by personal quarrels between U Nu and his closest associates. Amid rumours of a military takeover, U Nu invited the army chief of staff, Gen. Ne Win—who had been a Thakin, one of the Thirty Comrades, and Aung San's second in command—to take the premiership. Ne Win prepared the country for general elections, which took place in February 1960. The opponents were the two sections of the AFPFL, and U Nu was returned to office with an absolute majority.

Socialist takeover. In March 1962 Gen. Ne Win led a coup d'état and arrested U Nu, the chief justice, and Cabinet ministers. Suspending the constitution, he ruled the country with a revolutionary council made up of senior military officers. His stated purpose was to make Myanmar a socialist state. Land had been nationalized in U Nu's administration, and now all major commerce and industry were nationalized as well; the economy did not improve, however. A one-party (Burma Socialist Programme Party [BSPP]) system was established, and measures were introduced to decentralize the administration. In April 1972 Ne Win and other members of the revolutionary council retired from the army.

New constitution

U Ne Win promised a new constitution, and in September 1971 representatives of the party's central committee, of the country's various ethnic groups, and of other interest groups were appointed to draft a document. A referendum to ratify the new constitution was held in December 1973, with more than 90 percent of eligible voters signifying approval. Elections to the Pyithu Hlutdaw—the supreme legislative, executive, and judicial authority—and to local People's Councils were held early in 1974; the new government was established on March 2 with U Ne Win as president.

After the establishment of the new political organization, Myanmar's economy grew steadily but slowly, although in the 1980s mounting trade deficits and external debt payments increasingly hindered growth. Student and worker unrest erupted periodically, and Communist and ethnic insurgency continued in the eastern and northern parts of the country. In May 1980 U Ne Win offered full amnesty to all political insurgents inside or outside Myanmar who reported to authorities within a 90-day period. Most notable among those accepting was U Nu, who returned from India to practice as a Buddhist monk. U Ne Win retired as president and chairman of the State Council in November 1981 but remained in power until July 1988, when he resigned as chairman of the BSPP amid violent protests against his rule. A civilian government was then established, but pro-democracy demonstrations continued. On September 18, 1988, the armed forces, led by Gen. Saw Maung, seized control of the government. The military moved to suppress the demonstrations, and thousands of unarmed protestors were killed. Martial law was imposed over most of the country.

On May 27, 1990, Myanmar held its first multiparty parliamentary elections in 30 years. The result was a landslide victory for the opposition National League for Democracy (NLD). The NLD's leaders, U Tin U, a former colleague of U Ne Win, and Daw Aung San Suu Kyi, the daughter of the nationalist leader Aung San, had been under house arrest since July 1989. The military government had announced before the elections that it would retain power until the new assembly wrote a constitution and formed a strong government. (Ed.)

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

For coverage of related topics in the *Macropædia* and *Micropædia,* see the *Propædia,* sections 937, 969, and 976, and the *Index.*

SINGAPORE

The Republic of Singapore is a city-state situated at the southern tip of the Malay Peninsula, about 85 miles (137 kilometres) north of the Equator. It consists of Singapore Island, which is diamond-shaped and 221 square miles (572 square kilometres) in area, and some 50 other islets, many of which are uninhabited, which make up another 18 square miles (46 square kilometres). To the north the main island is separated from West Malaysia by Johor Strait, a narrow channel crossed by a road and rail causeway more than half a mile long. The southern limits of the state run through Singapore Strait where outliers of the Riau-Lingga Archipelago, which forms a part of Indonesia, extend to within 10 miles (16 kilometres) of the main island. Most of Singapore's population is Chinese.

ze and
gional
portance

The largest port in Southeast Asia, and one of the largest ports in the world, Singapore owes its growth and prosperity to its focal position at the southern extremity of the Malay Peninsula, dominating the Strait of Malacca that leads from the Indian Ocean to the South China Sea. Once a British colony, and now a member of the Commonwealth of Nations, Singapore joined the federation of Malaysia on its formation in 1963. When friction subsequently developed between the Malay-dominated central government and the Chinese-dominated Singapore administration, Singapore accepted Malaysia's invitation to leave the federation and become an independent state, which it did on August 9, 1965.

Since 1965, despite a large and rapidly growing population concentrated on a minute geographical base, Singapore has reduced its birth rate, rehoused many of its people, initiated a program of industrial expansion, and raised national income to the second highest per capita level of any Southeast Asian nation. Singapore was formerly the principal British naval and air base in Asia, but its strategic role has been eroded by developments in nuclear warfare, as well as by the withdrawal of British power from Asia.

Physical and human geography

THE LAND

Relief. Nearly two-thirds of the main island is less than 50 feet (15 metres) above sea level. Bukit Timah, the highest summit, is only 531 feet (162 metres) high; with other peaks, such as Bukit Panjang and Bukit Mandai, it forms a block of rugged terrain in the centre of the island. To the west and south are lower scarps with marked northwest-southwest trends, such as Bukit Faber Ridge. The eastern part of the island is a low plateau cut by erosion into an intricate pattern of hills and valleys. These physical units reflect their geological foundations: the central hills are formed from granite rocks, the scarplands from highly folded and faulted sedimentary rocks, and the eastern plateau from uncompacted sands and gravels.

Drainage and soils. A dense network of short streams drains the island, but floods are locally severe, owing to low gradients and to excessive runoff of water from cleared land. Many streams, especially those draining northwards, have broad mangrove-fringed estuaries that extend far inland. None of the soils is even reasonably fertile, but those derived from the granites tend to be better than most. Soils developed from the sedimentary rocks are variable, but many contain hard pans (compacted layers) that restrict plant roots and impede soil drainage. The soils of eastern Singapore are extremely infertile. All have suffered extensive degradation (loss of material through erosion) as a result of generations of careless exploitation.

limate

Climate. Singapore, being in an equatorial region, experiences uniformly high temperatures throughout the year, with the mean monthly average varying from about 81° F (27° C) in June to 77° F (25° C) in January. The daily range is much greater, averaging 12° F (7° C), but the maritime location and constant humidity keep maximum temperatures relatively modest: the highest temperature ever recorded was only 95° F (35° C). As temperatures are so uniform, the incidence of rainfall defines the seasons; rainfall in turn depends on air-mass movements, which are affected by the presence of the Asian and Australian land masses. Between December and March, for example, the northeast trade winds that blow at that season are reinforced by outflows of surface air from Asia, bringing the strongest winds and heaviest precipitation of the year; rainfall averages more than 10 inches (250 millimetres) a month in both December and January. From May to September surface winds from the southern hemisphere are strengthened by the Australian high-pressure belt and reach the Singapore-Malaya region. Winds are generally light during this season, however, and rainfall drops to about six inches in June and July. From April to May and from October to November sluggish air movements are common, and intense afternoon showers bring monthly rainfall averages of seven to 10 inches. Altogether, the annual rainfall averages about 100 inches, and rain falls somewhere on the island every day of the year.

Plant and animal life. Little remains of the original vegetation or animal life, except for a few thousand acres of second growth rain forest preserved around catchment areas. Some mangrove vegetation survives in the Kranji estuary on the northwest side of the island, but elsewhere tracts of scrub, or of *Imperata* grass called lalang, are common. Many exotic plants have been introduced for ornamental use. The largest native animals are the long-tailed macaque (an Asian species of monkey), the slow loris (a large-eyed tailless nocturnal lemur), and the scaly anteater. Birds are numerous, especially those, such as the Indian mynah bird, the brahminy kite (a kite with reddish-brown plumage and a white head and breast), and the house swallow, which have adapted to a symbiotic relationship with man. Reptiles, such as cobras and lizards, are also common. Fringing coral reefs with their associated fish and wildlife occur around many parts of the coast.

Settlement patterns. The city of Singapore stands on the southern part of the main island. Urbanization has reduced differences between city and country. Today built-up areas cover a large fraction of the main island. In the older parts of the city (along Singapore River, in the Rochore-Kallang floodplain, and in the low hills behind the harbour), rows of narrow-fronted Chinese shops, with their upper floors subdivided into living cubicles, line the streets. Similar shops stand along the main roads that radiate from the city. Behind their facades lie extensive suburbs composed partly of bungalows, semi-detached residences, and apartments, and partly of rural types of dwellings, such as wooden huts roofed with *atap* (thatch made from the *Nypa* species of palm tree) or corrugated iron. Such huts are common in all the rural districts, and in squatter settlements in the city. The traditional Malay *kampong*s (villages built on stilts) are seldom seen, except in fishing villages on some of the smaller islands.

The port of Singapore occupies the southern coast of the main island south of the Singapore River. It consists of two partly sheltered areas of water—the Eastern Roads, which are subdivided by an offshore breakwater, and the Western Roads. The two roadsteads are connected by the deepwater channel of Keppel Harbour, with its miles of berths, docks, warehouses, and other facilities, which runs between the main island and islands to the south. A few miles further west is the new port, Jurong.

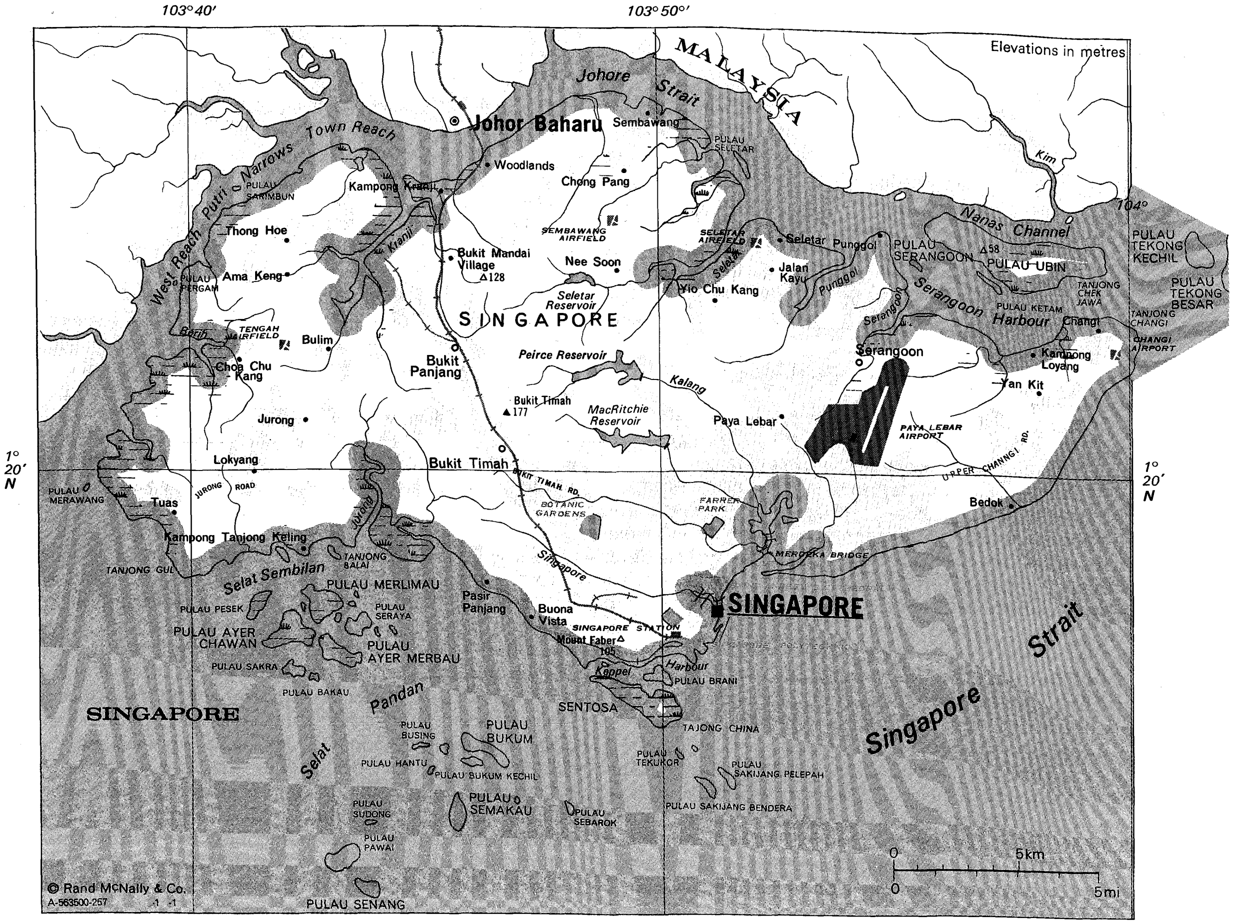
THE PEOPLE

As a result of past immigration, the population is diverse. Chinese predominate, forming about three-fourths of the total. Malays form the next largest ethnic group and Indians the third. None of these three major communities is homogeneous. Among the Chinese, almost half originate from Fukien Province and speak the Amoy dialect; about one-fourth are Teochews from Swatow, and roughly one-fifth are from Kwangtung. The Chinese community as a whole, therefore, uses mutually incomprehensible dialects. Linguistic differences are less pronounced among the Malays, but the group includes Indonesians speaking Javanese, Boyanese, and other dialects. The Indian group is most diverse, consisting of Tamils (more than half), Malayalis, and Sikhs; it is also considered to include the Pakistani and Sinhalese communities.

Official
languages

Because of this diversity, no fewer than four official languages are recognized—English, Mandarin, Malay, and Tamil. English remains the main medium for administration, commerce, and industry; more than half of all schoolchildren attend English-language schools. Mandarin, the national language of China, transcends dialect barriers; one-third of the school population is taught in that language. Malay, like English, is widely used for interracial communication and plays an especially useful role in view of the close ties between Singapore and Malaysia.

Religious affiliations reflect ethnic patterns. About three-fourths of all Chinese profess some degree of attachment to Confucianism, Buddhism, or Taoism, or to some combination thereof. Virtually all Malays, and some Indians, adhere to Islām, which is the formal religion of about



SINGAPORE

MAP INDEX

Cities and towns

Ama Keng	1·24n	103·42e
Bedok	1·19n	103·57e
Bukit Mandai Village (Mandai)	1·25n	103·45e
Bukit Panjang	1·23n	103·46e
Bukit Timah	1·20n	103·47e
Bulim	1·23n	103·43e
Buona Vista	1·16n	103·47e
Changi	1·23n	103·59e
Choa Chu Kang	1·22n	103·41e
Chong Pang	1·26n	103·50e
Jalan Kayu	1·24n	103·52e
Jurong	1·21n	103·42e
Kampong Kranji	1·26n	103·46e
Kampong Loyang	1·22n	103·58e
Kampong Tanjon Keling	1·18n	103·42e
Lokyang	1·20n	103·41e
Mandai, see Bukit Mandai Village		
Nee Soon	1·24n	103·49e
Pasir Panjang	1·17n	103·47e
Paya Lebar	1·22n	103·53e
Punggol	1·25n	103·55e
Seletar	1·25n	103·53e
Sembawang	1·27n	103·50e
Serangoon	1·22n	103·54e
Singapore	1·17n	103·51e
Thong Hoe	1·25n	103·42e
Tuas	1·19n	103·38e
Woodlands	1·27n	103·46e
Yan Kit	1·22n	103·58e
Yio Chu Kang	1·23n	103·51e

Physical features and points of interest

Ayer Chawan, Pulau, *island*	1·17n	103·42e
Ayer Merbau, Pulau, *island*	1·16n	103·43e
Badminton Stadium	1·18n	103·53e
Bakau, Pulau, *island*	1·16n	103·43e
Balai, Tanjong, *cape*	1·19n	103·47e
Berih, *river*	1·23n	103·40e
Besar, Pulau Tekong, *island*	1·24n	104·03e
Botanic Gardens, *park*	1·19n	103·48e
Brani, Pulau, *island*	1·15n	103·50e
Bukit Timah Race Course	1·20n	103·48e
Bukum, Pulau, *island*	1·14n	103·47e
Bukum Kechil, Pulau, *island*	1·14n	103·46e
Busing, Pulau, *island*	1·14n	103·45e
Changi Airport	1·22n	103·59e
Changi Prison	1·22n	103·58e
Changi, Tanjong, *point*	1·23n	104·00e
Chek Jawa, Tanjong, *cape*	1·24n	104·00e
China, Tanjong, *point*	1·14n	103·51e
Faber, Mount, *mountain*	1·16n	103·49e
Farrer Park	1·19n	103·51e
Fort Canning, *army installation*	1·18n	103·51e
Government House, *building*	1·18n	103·51e
Gul, Tanjong, *point*	1·17n	103·39e
Hantu, Pulau, *island*	1·14n	103·45e
Haw Par Villa, *building*	1·16n	103·47e
Jalan Besar Stadium, *sports arena*	1·18n	103·52e
Johore Strait	1·28n	103·48e
Jurong, *river*	1·18n	103·44e
Kalang, *river*	1·19n	103·52e
Kechil, Pulau Tekong, *island*	1·25n	104·01e
Keppel Harbour	1·16n	103·50e
Ketam, Pulau, *island*	1·24n	103·57e
Kranji, *river*	1·26n	103·45e
Kranji War Memorial, *monument*	1·26n	103·45e
MacRitchie Reservoir	1·21n	103·50e
Merawang, Pulau, *island*	1·20n	103·38e
Merdeka Bridge	1·18n	103·53e
Merlimau, Pulau, *island*	1·17n	103·42e
Nanas Channel	1·25n	103·58e
Nanyang University	1·21n	103·41e
National Museum	1·18n	103·51e
Pandan, Selat, *strait*	1·15n	103·44e
Pawai, Pulau, *island*	1·12n	103·43e
Paya Lebar Airport	1·21n	103·54e
Peirce, Reservoir	1·22n	103·49e
Pergam, Pulau, *island*	1·24n	103·40e
Pesek, Pulau, *island*	1·17n	103·41e
Punggol, *river*	1·25n	103·54e
Putri Narrows, *strait*	1·27n	103·42e
Saint Andrew's Cathedral	1·18n	103·51e
Sakijang Bendera, Pulau, *island*	1·13n	103·51e
Sakijang Pelepah, Pulau, *island*	1·13n	103·52e
Sakra, Pulau, *island*	1·16n	103·42e
Sarimbun, Pulau, *island*	1·26n	103·41e
Sebarok, Pulau, *island*	1·13n	103·48e
Seletar, *river*	1·25n	103·52e
Seletar, Pulau, *island*	1·27n	103·52e
Seletar Air Field, *airport*	1·16n	103·53e
Seletar Reservoir	1·24n	103·48e
Semakau, Pulau, *island*	1·12n	103·46e
Sembawang Airfield, *airport*	1·25n	103·49e
Sembilan, Selat, *strait*	1·18n	103·42e
Senang, Pulau, *island*	1·11n	103·44e
Sentosa, *island*	1·15n	103·50e
Serangoon, *river*	1·24n	103·56e
Serangoon, Pulau, *island*	1·25n	103·56e
Serangoon Harbour	1·23n	103·57e
Seraya, Pulau, *island*	1·16n	103·43e
Singapore, *island*	1·23n	103·48e
Singapore, *river*	1·17n	103·51e
Singapore, University of	1·19n	103·49e
Singapore International Airport, see Paya Lebar Airport		
Singapore Polytechnic, *university*	1·16n	103·51e
Singapore Station, *rail terminal*	1·17n	103·50e
Singapore Strait	1·10n	103·55e
Sudong, Pulau, *island*	1·13n	103·44e
Sultan Mosque	1·18n	103·52e
Tekukor, Pulau, *island*	1·14n	103·50e
Tengah Airfield, *airport*	1·23n	103·42e
Timah, Bukit, *hill*	1·21n	103·45e
Town Reach, *channel*	1·28n	103·44e
Ubin, Pulau, *island*	1·24n	103·58e
Victoria Memorial Hall, *building*	1·17n	103·51e
West Reach, *channel*	1·22n	103·38e

one-sixth of the population. The remaining population is either Christian or Hindu.

THE ECONOMY

With an urban population engaged largely in trade, finance, and industry, Singapore's economy differs markedly from that of most Southeast Asian countries, which are primarily based on the production of raw materials. The monetary unit is the Singapore dollar.

Resources. The state has few natural resources, apart from small amounts of brick clays, rock aggregates, and building sand. The natural vegetation provides only firewood and a few poles. Local catchment areas are small, obliging Singapore to use water from Johor state in West Malaysia. The low-lying relief does not permit hydroelectric development, so that energy has to be generated from imported fuels.

Agriculture and fisheries. About one-fourth of the total area of Singapore is devoted to agriculture. Although more than half the agricultural acreage is used for the cultivation of rubber and coconuts, most of the farms produce vegetables, fruit, and livestock for local markets. Singapore is self-sufficient in pig, poultry, and egg production. Large amounts of marine fish are caught in local waters, and ~~supplies are augmented by fish raised in freshwater ponds~~ on farms. Total fish production nevertheless accounts for a fraction of total consumption.

gricul-
ral
oduction

Industry. The manufacturing sector, on the other hand, has expanded rapidly. Manufacturing plants produce transport equipment and petroleum products, and process food. The ship-repairing and ship-building industry is growing rapidly.

Finance. Financial services are available both from modern banks and insurance companies and from traditional thrift societies and moneylenders. So far, most of all bank credit has been advanced for general commerce and trade, but the proportion devoted to industry is growing, with encouragement from such official bodies as the Development Bank of Singapore. Other sources ~~of finance include finance companies, building societies,~~ insurance firms, and co-operatives. Singapore also has a thriving stock exchange. Commodity markets, especially for rubber, pepper, and tin, form an important element in the economy.

Trade. Singapore continues to perform its traditional financial-intermediary function of shipping raw materials from the region in exchange for finished products from the industrial nations. Malaysia, the U.S., the United Kingdom, China, and Vietnam are—after Japan—Singapore's principal trading partners. Rubber and petroleum together account for about one-fourth of the value of all trade. Crude rubber, imported from West Malaysia and Indonesia, is shipped to Japan, the U.S., China, the Soviet Union, and Britain. Petroleum from the Middle East is refined in Singapore, and the resulting fuel oils, diesel oils, gasoline, and other products are exported to Vietnam, Thailand, Japan, and Hong Kong. Other significant items in foreign trade include timber, spices, and rice. The export of machinery, vehicles, and textiles is also increasing.

Administration of the economy. The economy of Singapore is basically of the free-enterprise type. The private sector therefore has an essential entrepreneurial role, and is expected to generate the increased productivity necessary for progress. At the same time the government is strongly committed to developmental goals, and its influence is pervasive.

urces of
venue

The main sources of government revenue are excise duties (mainly on liquor, tobacco, motor vehicles, and petroleum products) and income tax (mostly from companies). Tax relief may be granted to new manufacturing and export enterprises; to encourage the entrepôt trade, free trade zones have been established.

Only one-fourth of all workers are trade unionists. Most of those employed in the public sector and most employees of the larger, foreign-directed firms are unionized. There are numerous separate unions, led by the National Trades Union Congress, whose main objectives are to secure welfare benefits for workers by established procedures. The government regulates collective bargaining and compulsory arbitration, and it also provides a wide range of services, including training courses and a workers' superannuation scheme. Organized management is represented by two bodies, and there are, in addition to these, chambers of commerce representing different ethnic groups.

The country's limited area, a rapid population growth, and a weakening role as financial middleman have virtually obliged Singapore to develop its manufacturing industries. Despite advances in increasing the number of manufacturing establishments, the economy still depends on external factors over which the state has no control, such as fluctuations in the price of rubber on the world market. The internal problem of providing full employment for an expanding population has yet to be solved.

Transportation. As one of the largest ports in the world, Singapore is a major transportation nucleus. Roughly one-half of all ships handled are engaged in the coastal and interisland trade conducted between Singapore and Indonesia, but the extent of this barter trade is poorly documented. Railroads connect Singapore with West Malaysia. Roads are numerous. Both international and pan-Malaysian airlines use the airport at Changi.

ADMINISTRATIVE AND SOCIAL CONDITIONS

~~**Government.** Singapore is a parliamentary democracy~~ based on the Westminster model. The government consists of a head of state, a president with nominal powers, and a fully elected, unicameral Parliament of 58 members. The electorate includes every adult citizen who is a registered voter, and voting is compulsory. The parliamentary majority selects the prime minister and cabinet from its own ranks. In each constituency there is a Citizens' Consultative Committee, designed to act as a link between government and people. Each committee explains official policies and transmits demands or grievances to the government. Similar exchanges are conducted with organized labour, which has six seats in Parliament.

Close liaison is maintained between the political and administrative arms of government. The administra~~tive structure comprises various ministries and statutory~~ boards. These are staffed by civil servants whose technical competence and integrity are monitored by a Public Service Commission, an independent body responsible for controlling standards, appointments, and discipline.

Justice. A similar body maintains the status and assures the independence of the judiciary. Justice is administered by the Supreme Court and by courts of minor jurisdiction such as magistrates' and criminal district courts. Appeals can be made from the lower to the higher courts, with final appeal to the judicial committee of the Privy Council in London. Legal assistance in civil actions is available for the poor. The Shariah Court has jurisdiction in matters of Islāmic law.

Armed forces. The armed forces of Singapore consist of infantry battalions with supporting artillery, armour, engineer, and signal units. The training of fighter pilots and of personnel for local missile defenses has begun. There is also a flotilla of fast naval patrol boats. Military conscription for 18-year-olds was introduced in 1967. Those not selected for the statutory two-year period of service with the regular forces are obliged to serve on a part-time basis with either the Peoples' Defence Force, the Special Constabulary, or the Vigilante Corps (a paramilitary force largely made up of youths).

Education. Primary education in any one of the four official languages is free for children of Singapore citizens, and the learning of a second language is compulsory. Secondary education is available in technical, commercial, and academic subjects; courses are combined to prevent premature specialization. Existing school facilities are sufficient to meet anticipated future demands. There are also technical institutes. The National University of Singapore, founded in 1980 by a merger of the University of Singapore and Nanyang University, provides academic and professional courses to postgraduate levels. There are separate institutes to train teachers, nurses, radiographers, and seamen; many centres offer adult education.

Health and welfare. The range and quality of medical services is notably high, with a large number of doctors

Medical services

and dentists. There are both government and private hospitals, while nonhospital care is dispensed from numerous out-patient clinics and mobile centres. Official campaigns to keep Singapore clean are conducted periodically. Singapore's most significant achievement is in family planning, a distinction that it shares with few other nations; it resulted in a drop in the birth rate.

A housing and development board, established by the government, is responsible for building homes for families with low incomes. The units being built and that have been built have ameliorated somewhat the severe housing shortage that was endemic in Singapore.

The police force comprises regulars and members of the Special Constabulary. The force patrols the islands by land and water in order to maintain internal security. Traffic management, crime prevention, and political subversion are also matters under police jurisdiction. A police academy trains recruits and serving personnel.

Health conditions in Singapore compare favourably with those of the advanced nations. With the reduction in communicable diseases, cardiovascular diseases and malignant neoplasms account for an increased proportion of deaths. The government and voluntary associations, headed by the Council of Social Service, provide welfare services for the aged, sick, and unemployed.

CULTURAL LIFE

Cultural activities in Singapore are largely derivative, springing from one or another of the major civilizations of China, India, Indonesia, or the West. Traditional Chinese and Indian music, painting, and drama are practiced by numerous cultural societies and professional groups. Popular culture, based on modern mass media, is far more widespread. Malay music, which has adopted the rhythms of Western orchestras, has a general appeal. Musical films that popularize Hindustani and Tamil songs have a considerable following, as do films from Hong Kong, Taiwan, and the United States. Several Chinese, English, Indian, and Malay newspapers serve a largely literate population. Magazines published in the West, Hong Kong, and Japan have wide appeal. (Rt.H.)

For statistical data on the land and people of Singapore, see the *Britannica World Data* section in the BRITANNICA WORLD DATA ANNUAL.

History

The island, originally inhabited by fishermen and pirates, served as an outpost for the Sumatran empire of Śrivijaya. In Javanese inscriptions and Chinese records dating to the end of the 14th century, the more common name of the island is Tumasik, or Temasek, from the Javanese word *tasek* ("sea"). The Cōla ruler Rājendra attacked the island in 1025, and there was another Cōla raid in 1068. In 1275 the Javanese king Kritanagara, when he raided Pahang, probably attacked Temasek. According to a Chinese traveller, Wang Ta-yuan, just before 1349 about 70 Siamese war boats besieged Temasek for a month but had to withdraw. The Javanese *Nagarakritagama* (written 1365) includes Temasek among the conquests of the Javanese empire of Majapahit. At the end of the 14th century, Temasek fell into decay and was supplanted by Malacca (now Melaka). Yet in 1552 it was still a port of call from which St. Francis Xavier dispatched letters to Goa, and João de Barros, in *Décadas da Asia* (1553), described it as "a resort not only of Indian shipping but of traders from China, Siam, Champa, Cambodia, and the Malay Archipelago." Rājendra may have named the city Singapore (City of the Lions), or the name may have been bestowed in the 14th century by Buddhist monks, to whom the lion was a symbolic character. In the *Sejarah Melayu,* a Malay chronicle, the city was founded by the the Śrivijayan prince Sri Tri Buana; he glimpsed a tiger but mistook it for a lion, and thus called the settlement Singapura. (Rt.H./Ed.)

EAST INDIA COMPANY

On January 28, 1819, Sir Thomas Stamford Raffles of the East India Company, searching for a factory site, forestalled by the Dutch at Riau (Riouw, Rhio), and, having found the Carimon Islands unsuitable, landed at Singapore. He found only a few Chinese planters, some aborigines, and a few Malays and was told by the hereditary chief, the *temenggong* (direct ancestor of the sultans of modern Johore), that there were no Dutch there and that the company could purchase land. The *temenggong,* however, was a subordinate of his cousin Abdul Rahman, sultan of Riau, Lingga, Pahang, Johore, and Singapore, who was under Dutch surveillance. Abdul Rahman was a younger son and not a sultan de jure. Raffles, disobeying instructions not to offend the Dutch, withdrew his own recognition of Abdul Rahman as sultan of Singapore and installed Abdul Rahman's elder brother, Husain, to validate the purchase of land there on behalf of the company. The Dutch protested. In London the court of directors, though it decided Raffles had contravened instructions, took no action.

British ru

In 1824 an Anglo-Dutch treaty left Malaya and Singapore in the British sphere, and on August 2 the whole of Singapore Island was ceded to the British for a money payment. In 1826 Singapore, Penang, and Melaka were combined as the Straits Settlements to form an outlying residency of India. In 1830 they were reduced to a residency under Bengal, and two years later Singapore became their capital. When the East India Company lost its monopoly of the China trade (1833), it also lost its interest in Malaya. The settlements were transferred to the direct control of the governor-general of India in 1851. In 1867 they were made a crown colony under the Colonial Office.

DEVELOPMENT OF THE PORT

Meanwhile, the trade of Singapore had suffered from British development after 1842 of a rival port, Hong Kong, as later it was to suffer from the French occupation of Indochina, the development of Saigon and Haiphong, and from the establishment of Dutch ports and shipping lines in the Netherlands East Indies. With the opening of the Suez Canal in 1869 and the advent of steamships, however, an era of prosperity began that led eventually to the construction of three miles of wharves at Tanjong Pagar and finally, in 1921, to a naval base. The economic growth of the Malay states after they became British protectorates enlarged transit trade. (R.O.Wt.)

It was the demand of the industrial West for tin and rubber that made Singapore one of the greatest ports in the world. After World War I steps were taken to modernize Malayan defenses and, with the lapsing of the Anglo-Japanese alliance, to build a large naval base in Singapore.

WORLD WAR II

In World War II the Japanese landed in Kelantan and Thailand on December 7, 1941 (December 8 west of Hawaii). By the end of January 1942 they had overrun the peninsula and were opposite the island of Singapore. The sinking of the battleship "Prince of Wales" and the battle cruiser "Repulse" had given them command of the sea, and a superior air force covered the advance down the peninsula. The Japanese crossed the Johore Strait on February 8, 1942, and the British command surrendered the island and city on February 15. Singapore remained in Japanese occupation until September 1945.

POLITICAL DEVELOPMENTS AFTER THE WAR

British political plans for Malaya excluded Singapore from the Malay Union and later from the Federation of Malaya mainly because Singapore's predominantly Chinese population would be a racial obstacle to a common citizenship. As a separate crown colony (from 1946) Singapore made constitutional progress despite the Communist insurrection in Malaya. Elected ministers and a Legislative Assembly with an elected majority assumed government responsibility in 1955, except for defense and foreign affairs. In 1959 the official and nominated elements were eliminated and Singapore became self-governing, although Britain still retained control of defense and foreign affairs.

Part of Malaysia

Singapore joined the Federation of Malaysia on its formation in September 1963. The ruling People's Action Party (PAP) led by Lee Kuan Yew had refused in 1959 to form a government until eight extreme left-wing leaders who had been detained by the colonial authorities were released. This left wing opposed the concept of Malaysia, broke away to form the Barisan Sosialis Party, and was accused of being a Communist front organization amenable to the control of the Indonesian Communist Party. The PAP faced fresh dangers of subversion when Indonesian opposition to Malaysia took the form of military and economic confrontation (1964). There were a number of attempts to blow up military installations and public utilities.

Confrontation ended in 1966, but Singapore had left Malaysia in 1965, at the invitation of the Malaysian government, because of political friction between the state and central governments. This conflict had racial overtones and continued to affect relations between Singapore and Malaysia. In May 1969 there was some rioting as a result of the racial disturbances in Malaysia after the general election there, but order was quickly restored.

In January 1968 the British government announced that all British defense forces would be withdrawn from the Far East (except Hong Kong) by the end of 1971. In April the unprepared major opposition parties boycotted an election called seven months before it was due. The ruling PAP termed its sweep of all 58 seats (a number newly increased) a mandate for its plans for lessening the economic effects of the British military withdrawal.

At midnight on October 31, 1971, 152 years of British military presence came to an end. The 14-year-old Anglo-Malayan treaty, which also covered Singapore and which had committed Britain to the defense of the region, terminated, and in its place a five-power defense arrangement—involving Britain, Australia, New Zealand, Malaysia, and Singapore as equal partners—came into force.

Since the 1970s, Singapore has experienced phenomenal economic success, its annual rate of economic growth often exceeding 10 percent. Singapore is politically stable, although some critics of the government feel that it is perhaps too stable. The country has been headed by Lee Kuan Yew and his People's Action Party since 1959, with PAP members completely dominating Parliament.

(A.Ke./Ed.)

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

THAILAND

A nation wedged like a keystone into the heart of Southeast Asia, Thailand in the last decades of the 20th century holds a critical position in the attempts of the countries of this area to achieve and maintain a political, economic, and social stability. In 1932 the absolutist monarchy was overthrown and the nation was organized as a constitutional monarchy with a representative legislature. In 1939 the official name was changed from Siam to Prathet Thai, or Thailand—literally, the "Land of the Free." The several ethnic and religious groups represented among Thailand's people are characteristic of the cultural diversity that for centuries has spread across this portion of the continent, a mêlange of influences from the two Asian giants, China and India.

e
ıdscape

Thailand's landscape is one of high mountains at the edge of the Himalayan chain, of fertile, alluvial plains dotted with rice paddies, and of sandy beaches and tropical forests set amid the latitudes of the Asian monsoons. The main body of the country is surrounded by Myanmar on the west and northwest, Laos on the northeast and east, Cambodia on the southeast, and the Gulf of Thailand (Gulf of Siam) on the south. From the southwest corner, part of Thailand stretches southward down the Malay Peninsula as far as Malaysia. This peninsula cuts off shipping using Thailand's capital and chief port, Bangkok, from points westward; Bangkok is, nevertheless, the international air hub of Southeast Asia.

Contacts of long standing between Thailand and the West have affected the forms if not the realities of Thailand's political and economic life. The relative political stability Thailand has maintained in the face of continual Communist guerrilla warfare inside its borders since World War II is largely the result of Western support. Thailand's free-enterprise economy participates vigorously in worldwide commerce yet remains based on primary products and relies on imports for manufactured goods. In addition, the nation's economic balance is maintained by aid from the United States and other nations.

Physical and human geography

THE LAND

Three main geological regions cover most of Thailand's 198,455 square miles (514,000 square kilometres)—the folded mountains in the north, the Khorat uplift in the east, and the Chao Phraya depression, comprising much of the central plains. In addition, the maritime southeast and the long, slender peninsula in the southwest constitute separate physical regions. The monsoon climate prevailing across most of the country has a distinct influence on the landscape, its vegetation and animal life, and its human uses.

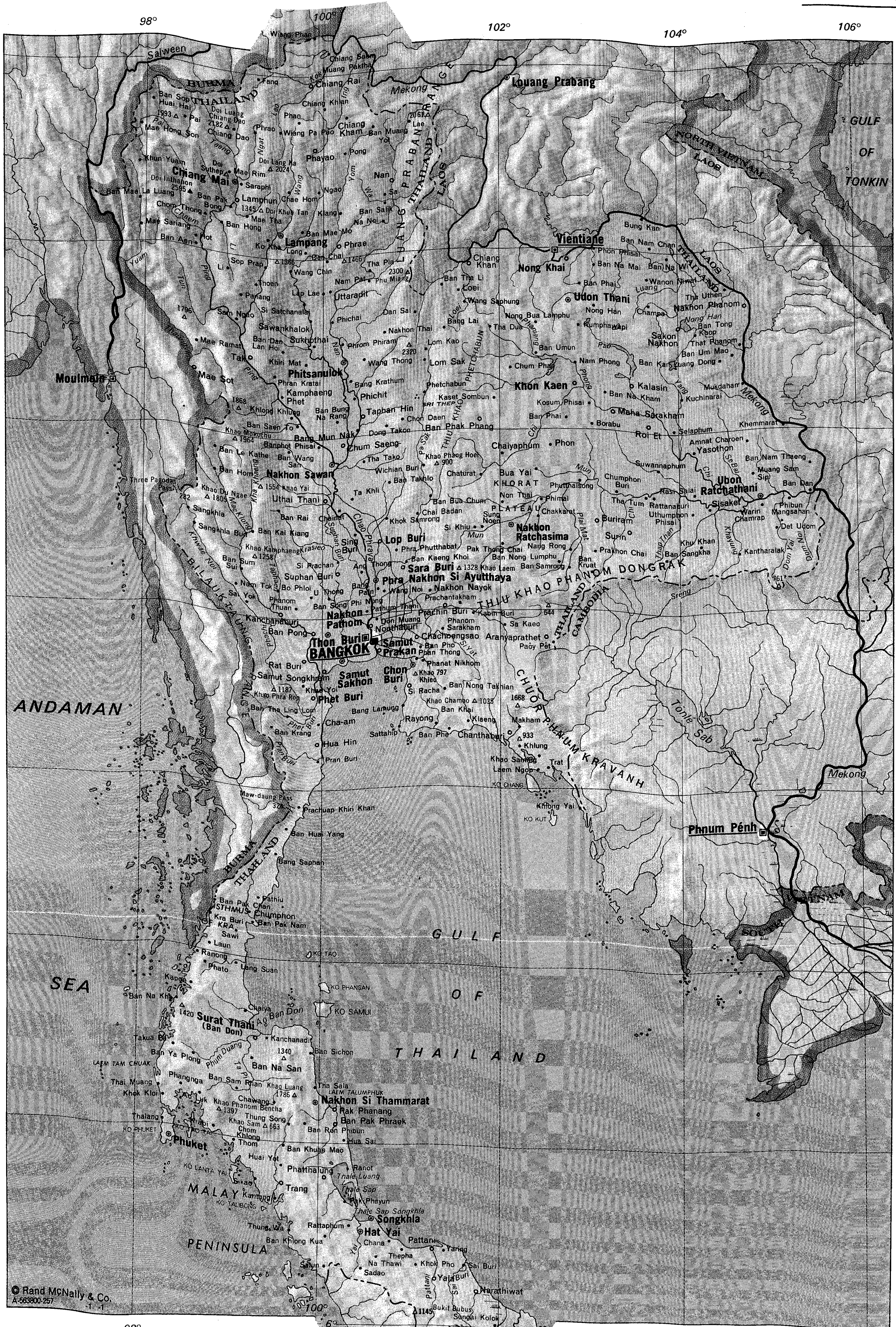
Relief. The mountains, a continuation of the Himalayan system from India, Myanmar, and China, extend far southward in Thailand along the western border and the peninsula and into Malaysia. Long granitic ridges were formed when great masses of molten rock forced their way upward through the older sedimentary strata. Peaks average about 5,200 feet above sea level. Doi Inthanon, at 8,514 feet (2,595 metres) the highest in the kingdom, is southwest of the ancient city of Chiang Mai (Chiengmai), which is overshadowed by the peak of Doi Suthep, a tourist attraction and site of the royal resort palace, Phu Phing Ratchanivet. Many of the rugged limestone hills contain caves from which remains of prehistoric man have been excavated.

The lower reaches of the mountains are rich in teak and other commercial timber, while the upper slopes are dotted with tea plantations. The rivers, emptying through narrow valleys into the central plain, are glutted by the monsoons, sweeping along in their steep descent great quantities of sediments that have produced vast, fanshaped heaps of alluvial deposits along the flood plains at the foot of the mountainous regions.

The
Khorat
region

The Khorat region, enfolded on three sides by Laos and Cambodia, consists of a two-sided geological fault that has been tilted rather than a uniform uplift of the underlying sedimentary rocks. At the western edge, the tilting produced north–south ranges whose escarpments look westward over the central plain. To the south, the tilting produced east–west ranges along the Cambodian border; these form steep escarpments that overlook the Cambodian plain. Surface elevations in the Khorat region are about 650 feet; the terrain is rolling, and the hilltops generally slope in conformity with the tilt of the land.

The generally rolling countryside of the southeast has high hills in the centre and along the eastern boundary with Cambodia. Notable peaks are Khao Khieo (2,614 feet [797 metres]), visible from the top of Phu Khao Thong (the Golden Mount) in Bangkok, and Khao Soi Dao ("Reaching for the stars"), which attains a height of 5,471 feet (1,668 metres). The mountains, reaching nearly to the sea, create a markedly indented coastline fringed with many islands, some of which are popular tourist resorts. The short, seaward-running streams have built up small alluvial basins and deltas along the coast, while the mouths of larger rivers consist of tidal flats and mangrove swamps. Long stretches of sandy beach make Chon Buri, Rayong, and some of the islands year-round resorts. On higher grounds plantations produce rubber, sugarcane, pineapples, and cassava. The central plains, which form the Thai heartland, consist of two regions—the heavily dissected rolling plains in the north and the Chao Phraya River Delta.



THAILAND

Size of symbol indicates relative size of town

• ○ ◉ ▣ ▣ ■

Elevations in metres

MAP INDEX

Cities and towns

Amnat Charoen	15·51n	104·38e
Ang Thong	14·35n	100·27e
Ao Luk	8·24n	98·44e
Aranyaprathet	13·42n	102·32e
Ban Aen	18·02n	98·35e
Ban Bua Chum	15·15n	101·15e
Ban Bung Na Rang	16·10n	100·11e
Ban Dan	15·19n	105·30e
Ban Dan Lan Hoi	16·59n	99·35e
Ban Don, see Surat Thani		
Bangkok	13·45n	100·31e
Bang Krathum	16·36n	100·20e
Bang Lai	17·06n	101·25e
Bang Lamung	12·55n	100·56e
Bang Mun Nak	16·02n	100·23e
Bang Pa-In	14·21n	100·38e
Bang Saphan	11·14n	99·31e
Ban Hom	15·33n	98·46e
Ban Hong	18·19n	98·51e
Ban Huai Yang	11·32n	99·38e
Ban Kaeng Khoi	14·33n	100·59e
Ban Kai Kiang	14·55n	99·15e
Ban Kan Luang Dong	16·47n	104·29e
Ban Khai	12·48n	101·20e
Ban Khlong Kua	6·59n	100·07e
Ban Khuan Mao	7·58n	99·38e
Ban Krang	12·45n	99·29e
Ban Kruat	14·24n	103·04e
Ban Le Kathe	15·49n	98·48e
Ban Mae La Luang	18·28n	97·57e
Ban Mae Mo	18·19n	99·47e
Ban Muang Yot	19·23n	100·35e
Ban Na Kha	9·26n	98·28e
Ban Na Kham	16·24n	103·14e
Ban Na Mai	17·47n	103·12e
Ban Nam Chan	18·03n	103·44e
Ban Nam Thaeng	15·34n	105·30e
Ban Na San	8·53n	99·17e
Ban Na Wi	17·56n	103·55e
Ban Nong Lumphu	14·40n	102·43e
Ban Nong Takhian	13·06n	101·25e
Ban Pak Bong	18·30n	98·57e
Ban Pak Chan	10·30n	98·51e
Ban Pak Nam	10·24n	99·16e
Ban Pak Phraek	8·13n	100·13e
Ban Phai	16·07n	102·47e
Ban Phai	17·33n	103·00e
Ban Phak Phang	16·06n	101·48e
Ban Phe	12·39n	101·38e
Ban Pho	13·34n	101·09e
Banphot Phisai	15·55n	99·58e
Ban Pong	13·50n	99·55e
Ban Rai	15·01n	99·35e
Ban Ron Phibun	8·11n	99·51e
Ban Saen To	16·09n	99·44e
Ban Salik	18·32n	100·44e
Ban Sam Phan	8·34n	99·02e
Ban Samrong	14·27n	102·50e
Ban Sangkha	14·36n	103·53e
Ban Sichon	8·57n	99·54e
Ban Song Phi Nong	14·18n	100·06e
Ban Sop Huai Hai	19·33n	98·05e
Ban Sum Sui	14·30n	99·08e
Ban Takhlo	15·26n	100·46e
Ban Takua Pa	8·52n	98·21e
Ban Tha Li	17·35n	101·24e
Ban Tha Ling Lom	12·53n	99·39e
Ban Tong Khop	17·04n	104·16e
Ban Um Mao	16·50n	104·37e
Ban Umun	16·51n	102·27e
Ban Wang San	15·52n	99·39e
Ban Ya Plong	8·50n	98·40e
Betong	5·47n	101·03e
Bo Phloi	14·17n	99·32e
Borabu	16·04n	103·05e
Bua Yai	15·34n	102·24e
Bung Kan	18·21n	103·38e
Buriram	14·59n	103·08e
Cha-am	12·48n	99·58e
Chachoengsao	13·42n	101·06e
Chae Hom	18·44n	99·37e
Chai Badan	15·05n	101·04e
Chainat	15·10n	100·10e
Chaiya	9·24n	99·12e
Chaiyaphum	15·48n	102·02e
Chakkarat	15·02n	102·27e
Champa	17·24n	103·42e
Chana	6·55n	100·44e
Chanthaburi	12·37n	102·09e
Chaturat	15·29n	101·49e
Chawang	8·25n	99·32e
Chiang Dao	19·23n	98·59e
Chiang Kham	19·31n	100·18e
Chiang Khan	17·50n	101·39e
Chiang Khian	19·39n	100·01e
Chiang Mai (Chiengmai)	18·46n	98·58e
Chiang Rai	19·54n	99·50e
Chiang Saen	20·14n	100·05e
Chom Thong	18·23n	98·41e
Chon Buri	13·22n	100·59e
Chon Daen	16·09n	100·54e
Chum Phae	16·40n	102·10e
Chumphon	10·32n	99·13e
Chumphon Buri	15·19n	103·20e
Chum Saeng	15·55n	100·15e
Dan Sai	17·18n	101·05e
Den Chai	17·59n	100·04e
Det Udom	14·54n	105·06e
Dong Takon	16·04n	100·46e
Don Muang	13·55n	100·39e
Fang	19·56n	99·14e
Hat Yai	7·01n	100·27e
Hot	18·06n	98·35e
Hua Hin	12·34n	99·58e
Huai Yot	7·45n	99·37e
Hua Sai	8·02n	100·17e
Kabin Buri	13·57n	101·43e
Kalasin	16·29n	103·31e
Kamphaeng Phet	16·26n	99·33e
Kanchanaburi	14·02n	99·33e
Kanchanadit	9·06n	99·26e
Kantang	7·23n	99·32e
Kantharalak	14·38n	104·41e
Kapoe	9·35n	98·38e
Kaset Sombun	16·20n	101·58e
Khao Saming	12·21n	102·28e
Khao Yoi	13·14n	99·49e
Khemmarat	16·03n	105·11e
Khiri Mat	16·50n	99·49e
Khlong Khlung	16·14n	99·40e
Khlong Thom	7·56n	99·09e
Khlong Yai	11·47n	102·53e
Khlung	12·30n	102·15e
Khok Kloi	8·19n	98·18e
Khok Pho	6·42n	101·08e
Khok Samrong	15·03n	100·44e
Khon Kaen	16·26n	102·50e
Khu Khan	14·42n	104·12e
Khun Yuam	18·47n	97·57e
Klaeng	12·46n	101·39e
Klang	18·28n	100·14e
Ko Kha	18·11n	99·24e
Kosum Phisai	16·16n	103·01e
Krabi	8·05n	98·53e
Kra Buri	10·24n	98·47e
Kuchinarai	16·23n	104·04e
Kumphawapi	17·08n	102·55e
Lae	19·25n	101·00e
Laem Ngop	12·12n	102·26e
Lampang	18·16n	99·34e
Lamphun	18·35n	99·01e
Lang Suan	9·55n	99·07e
Lap Lae	17·35n	100·01e
Laun	10·06n	98·47e
Li	17·50n	98·55e
Loei	17·32n	101·34e
Lom Kao	16·55n	101·10e
Lom Sak	16·47n	101·12e
Long	18·05n	99·50e
Lop Buri	14·48n	100·37e
Mae Hong Son	19·16n	97·56e
Mae Ramat	16·57n	98·36e
Mae Rim	18·54n	98·56e
Mae Sariang	18·10n	97·59e
Mae Sot	16·40n	98·35e
Mae Tha	18·28n	99·08e
Maha Sarakham	16·12n	103·16e
Makham	12·40n	102·12e
Muang Sam Sip	15·35n	104·48e
Mukdahan	16·31n	104·42e
Nakhon Nayok	14·12n	101·13e
Nakhon Pathom	13·49n	100·06e
Nakhon Phanom	17·22n	104·46e
Nakhon Ratchasima	14·57n	102·09e
Nakhon Sawan	15·42n	100·06e
Nakhon Si Thammarat	8·26n	99·58e
Nakhon Thai	17·05n	100·44e
Nam Pat	17·40n	100·33e
Nam Phong	16·45n	102·52e
Nam Tok	14·21n	99·00e
Nan	18·48n	100·46e
Nang Rong	14·39n	102·49e
Na Noi	18·19n	100·43e
Narathiwat	6·25n	101·48e
Na Thawi	6·45n	100·42e
Ngao	18·45n	99·59e
Nong Bua Lamphu	17·11n	102·25e
Nong Han	17·21n	103·06e
Nong Khai	17·52n	102·45e
Nonthaburi	13·51n	100·34e
Non Thai	15·10n	102·03e
Pai	19·25n	98·26e
Pak Phanang	8·21n	100·12e
Pak Phayun	7·22n	100·19e
Pak Thong Chai	14·44n	102·03e
Panang	17·33n	99·05e
Pathiu	10·41n	99·20e
Pathum Thani	14·01n	100·32e
Pattani	6·51n	101·16e
Phan	19·30n	99·45e
Phanat Nikhom	13·27n	101·10e
Phangnga	8·28n	98·32e
Phanom Sarakham	13·46n	103·23e
Phanom Thuan	14·07n	99·44e
Phan Thong	13·26n	101·07e
Phato	9·51n	98·45e
Phatthalung	7·38n	100·04e
Phayao	19·11n	99·55e
Phet Buri	13·06n	99·56e
Phetchabun	16·24n	101·11e
Phibun Mangsahan	15·14n	105·15e
Phichai	17·16n	100·06e
Phichit	16·24n	100·21e
Phimai	15·13n	102·29e
Phitsanulok	16·49n	100·15e
Phon	15·50n	102·35e
Phon Phisai	18·00n	103·06e
Phrae	18·07n	100·11e
Phra Nakhon Si Ayutthaya	14·21n	100·33e
Phran Kratai	16·41n	99·30e
Phrao	19·22n	99·15e
Phra Phutthabat	14·41n	100·51e
Phrom Phiram	17·01n	100·14e
Phuket	7·54n	98·24e
Phutthaisong	15·25n	102·55e
Pong	19·07n	100·16e
Prachantakham	14·04n	101·34e
Prachin Buri	14·03n	101·25e
Prachuap Khiri Khan	11·48n	99·47e
Prakhon Chai	14·49n	103·06e
Pran Buri	12·21n	99·59e
Ranong	9·59n	98·40e
Ranot	7·46n	100·20e
Rasi Salai	15·17n	104·11e
Rat Buri	13·32n	99·49e
Rattanaburi	15·16n	103·45e
Rattaphum	7·05n	100·22e
Rayong	12·43n	101·20e
Roi Et	16·05n	103·42e
Sa	18·37n	100·44e
Sadao	6·39n	100·31e
Sai Buri	6·42n	101·37e
Sai Yok	14·08n	99·10e
Sa Kaeo	13·50n	102·04e
Sakon Nakhon	17·10n	104·01e
Sam Ngao	17·16n	99·01e
Samut Prakan	13·35n	100·40e
Samut Sakhon	13·31n	100·15e
Samut Songkhram	13·25n	100·03e
Sangkhla	15·05n	98·32e
Sangkhla Buri	14·44n	98·38e
Sara Buri	14·30n	100·55e
Saraphi	18·43n	99·03e
Sattahip	12·41n	100·54e
Satun	6·39n	100·03e
Sawankhalok	17·19n	99·54e
Sawi	10·16n	99·06e
Selaphum	16·01n	103·59e
Sikao	7·34n	99·21e
Si Khiu	14·54n	101·44e
Sing Buri	14·53n	100·25e
Si Prachan	14·37n	100·11e
Si Racha	13·10n	100·56e
Sisaket	15·07n	104·21e
Si Satchanalai	17·30n	99·48e
Songkhla	7·13n	100·34e
Sop Prap	17·53n	99·21e
Sukhothai	17·01n	99·49e
Sungai Kolok	6·02n	101·58e
Sung Noen	14·54n	101·50e
Suphan Buri	14·29n	100·10e
Surat Thani (Ban Don)	9·06n	99·20e
Surin	14·53n	103·30e
Suwannaphum	15·35n	103·48e
Tak	16·52n	99·08e
Ta Khli	15·18n	100·20e
Taphan Hin	16·12n	100·26e
Tha Dua	17·08n	101·55e
Thai Muang	8·25n	98·16e
Thalang	8·04n	98·20e
Tha Pla	17·48n	100·34e
Tha Sala	8·35n	99·56e
Tha Tako	15·42n	100·30e
That Phanom	16·56n	104·42e
Tha Tum	15·19n	103·41e
Tha Uthen	17·34n	104·34e
Thepha	6·51n	100·57e
Thoen	17·41n	99·14e
Thon Buri	13·43n	100·29e
Thung Song	8·11n	99·41e
Thung Wa	7·01n	99·41e
Trang	7·33n	99·36e
Trat	12·14n	102·33e
Ubon Ratchathani	15·15n	104·54e
Udon Thani	17·25n	102·48e
Uthai Thani	15·20n	100·02e
U Thong	14·22n	99·54e
Uthumphon Phisai	15·06n	104·10e
Uttaradit	17·38n	100·06e
Wang Chin	17·51n	99·39e
Wang Noi	14·14n	100·43e
Wang Saphung	17·24n	101·36e
Wang Thong	16·50n	100·26e
Wanon Niwat	17·41n	103·41e
Warin Chamrap	15·09n	104·54e
Wiang Pa Pao	19·24n	99·32e
Wichian Buri	15·39n	101·10e
Yala	6·32n	101·19e
Yaring	6·49n	101·28e
Yasothon	15·46n	104·12e

Physical features and points of interest

Andaman Sea	11·00n	97·00e
Ban Don, Ao, *bay*	9·20n	99·25e
Bilauktaung Range	13·00n	99·00e
Bukit Bubus, *mountain*	6·12n	101·06e
Chaem, *river*	18·10n	98·40e
Chamao, Khao, *mountain*	12·58n	101·45e
Chang, Ko, *island*	12·05n	102·20e
Chao Phraya, *river*	13·32n	100·36e
Chi, *river*	15·13n	104·45e
Chuŏr Phnum Krâvanh, *mountains*	12·30n	102·50e
Dom Noi, *river*	15·18n	105·30e
Dom Yai, *river*	15·19n	105·10e
Du Ngae, Khao, *mountain*	15·13n	98·44e
Ing, *river*	20·13n	100·27e
Han, Nong, *river*	17·10n	104·05e
Inthanon, Doi, *mountain*	18·38n	98·30e
Kamphaeng, Khao, *mountain*	14·40n	99·15e
Kha Khaeng, *river*	14·55n	99·10e
Khayung, *river*	15·14n	104·50e
Khieo, Khao, *mountain*	13·15n	101·04e
Khong, see Mekong		
Khorat Plateau	15·30n	102·50e
Khun Tan, Doi, *mountain*	18·30n	99·20e
Khwae Noi, *river*	14·00n	99·33e
Kok, *river*	20·14n	100·09e
Kra, Isthmus of	10·20n	99·00e
Krasieo, *river*	14·48n	100·07e
Kut, Ko, *island*	11·40n	102·35e
Laem, Khao, *mountain*	14·27n	101·34e
Lang Ka, Doi, *mountain*	19·00n	99·22e
Lanta Yai, Ko, *island*	7·35n	99·05e
Lao, *river*	19·59n	99·59e
Loei, *river*	17·50n	101·35e
Luang, *river*	18·00n	103·05e
Luang Chiang Dao, Doi, *mountain*	19·22n	98·52e
Luang, Khao, *mountain*	8·30n	99·45e
Luang Praban Range	19·00n	101·00e
Luang, Thale, *lagoon*	7·40n	100·15e
Mae Klong, *river*	13·21n	100·00e
Malay Peninsula	6·00n	102·00e
Maw-daung Pass	11·47n	99·39e
Mekong (Khong), *river*	15·15n	105·30e
Miang, Phu, *mountain*	17·45n	100·58e
Mokochu, Khao, *mountain*	15·57n	99·00e
Mun, *river*	15·19n	105·31e
Nan, *river*	15·42n	100·11e
Ngat, *river*	19·09n	99·01e
Pai, *river*	19·08n	97·33e
Pao, *river*	16·12n	103·40e
Pa Sak, *river*	14·11n	100·40e
Pattani, *river*	6·53n	101·16e
Phachi, *river*	13·55n	99·25e
Phangan, Ko, *island*	9·45n	100·04e

MAP INDEX (continued)

Phang Hoei, Khao, *mountain*	15·37n	101·20e
Phaniang, *river*	16·46n	102·25e
Phanom Bencha, Khao, *mountain*	8·15n	98·58e
Phanom Dongrak, Thiu Khao, *mountains*	14·25n	103·30e
Phet Buri, *river*	13·10n	99·57e
Phetchabun, Thiu Khao, *mountains*	16·20n	100·55e
Phong, *river*	16·27n	102·57e
Phra Rop, Khao, *mountain*	13·10n	99·30e
Phuket, Ko, *island*	8·00n	98·22e
Phum Duang, *river*	9·10n	99·20e
Ping, *river*	15·40n	100·09e
Plai Mat, *river*	15·23n	102·47e
Pran Buri, *river*	12·25n	100·00e
Sai Buri, *river*	6·42n	101·37e
Salween, *river*	17·45n	97·40e
Sam Chom, Khao, *mountain*	8·08n	99·28e
Samui, Ko, *island*	9·30n	100·04e
Sap, Thale, *lagoon*	7·30n	100·17e
Sap Songkhla, Thale, *lagoon*	7·13n	100·30e
Se Bai, *river*	15·19n	104·50e
Siam, Gulf of, see Thailand, Gulf of		
Si Yat, *river*	13·45n	101·14e
Sri Thep, *ruins*	16·25n	101·04e
Suphan Buri, *river*	13·29n	100·17e
Suthep, Doi, *mountain*	18·49n	98·54e
Taeng, *river*	19·05n	98·57e
Talibong, Ko, *island*	7·15n	99·25e
Talumphuk, Laem, *cape*	8·30n	100·10e
Tam Chuak, Laem, *cape*	8·35n	98·13e
Tao, Ko, *island*	10·05n	99·53e
Taphaen, *river*	14·07n	99·25e
Ta Pi, *river*	9·05n	99·12e
Thailand, Gulf of (Gulf of Siam)	10·00n	101·00e
Thap Than, *river*	15·20n	104·00e
Thaungyin, *river*	17·50n	97·42e
Three Pagodas Pass	15·18n	98·23e
Tun, *river*	17·20n	98·42e
Wa, *river*	18·35n	100·45e
Wang, *river*	17·07n	99·02e
Yai, Khao, *mountain*	15·27n	99·20e
Yai, *river*	7·10n	100·27e
Yang, *river*	15·44n	104·00e
Yao Yai, Ko, *island*	8·00n	98·36e
Yom, *river*	16·40n	100·14e
Yuam, *river*	17·47n	97·48e

The peninsula

The topography of the peninsula is rolling to mountainous, with little flat land. A gently sloping sandy coastline, the site of the famous beach Hua Hin, borders the Gulf of Thailand on the east. The Phet is one of several large rivers dammed for irrigation. Massive mountains on the west, reaching about 4,900 feet, contain difficult passes between Thailand and Myanmar. Toward Malaysia mountains look to the Andaman Sea to the west and to the South China Sea to the east. Off the rugged and much-indented west coast lie numerous major islands, including tin-rich Phuket and others rich in bird and fish life.

Drainage and soils. Between the ridges of Thailand's western border lie relatively flat basins drained by the four principal tributaries of the Chao Phraya, Thailand's major river. Important provincial centres, or *changwats*, such as Chiang Mai, Lampang, Phrae, and Nan, grew up along these four streams. The alluvial soils of these intermontane basins provide fertile soil for the cultivation of rice, vegetables, tobacco, and fruit trees.

Monsoon rains over the thin forest cover of the Khorat region produce rapid runoff; flooding occurs almost yearly at Ubon Ratchathani at the junction of the Mun and the Chi. These two major rivers have built up scattered alluvial lands that are used for rice growing.

Close to the Mekong River, which meanders across the north and east of the Khorat region and runs along most of the Laotian border, swampy land and lakes are common, in contrast to the aridity of much of the region. A high underground water table that can be tapped produces mostly unpalatable water, but the Thai government has done much to improve the freshwater situation in this part of the kingdom. The Mekong itself, from about 2,300 to 4,300 feet wide, is either studded with islands or broken up by impassable rapids. The governments of both Thailand and Laos are making efforts to use it in agriculture and industry.

A rolling plain in the north also is heavily dissected, principally by the three rivers—the combined Ping and Wang, the Yom, and the Nan—that merge to form the Chao Phraya and its delta. Like most deltas, that of the Chao Phraya, which stretches southward to the Gulf of Thailand, is braided into many small channels; it is joined by other rivers as it crosses the plains. The usual flooding of the flat delta in the wet season is an asset to the rice-growing activity, although higher ground on the extreme eastern and western edges of the delta requires irrigation. The entire delta was at one time covered by the Gulf of Thailand, but the waterborne sediments from the uplands to the north filled it in over many centuries. Such silting is a continuing obstruction to navigation channels of the Chao Phraya, but it also provides several feet of new land each year at the river's mouth—so much so that a temple once on an island now lies on the river's western bank.

Effects of the monsoons

Climate. The major influences on Thailand's climate are its location on the Indochinese Peninsula within the sphere of the tropical monsoons and certain topographic features that modify the effects of the rains. Beginning in May, warm, humid air masses flow northeastward over the region from the Indian Ocean, depositing great quantities of rain that reach a maximum in September. Between October and February, the wind currents are reversed, and cold, dry air masses are driven in from the northeast. Stagnant air in March and April is associated with the hot, dry season.

Topographic effects are most noticeable on the peninsula, where Ranong on the west receives 188 inches (4,772 millimetres) of rain annually, and Hua Hin on the east receives only 40 inches (1,007 millimetres). Songkhla has its rainy season during the winter months, the result of moisture picked up from the Gulf of Thailand by the cold northeast air masses. In this area a true tropical rain-forest climate prevails.

Nationwide, temperatures are relatively steady throughout the year, averaging between 75° and 86° F (24° and 30° C). The greatest fluctuations are in the north, where frost may occur in December on higher elevations; maritime influences moderate the climate in the south. The cold, dry winter air produces frequent morning fogs that generally dissipate by midday.

Forests

Plant and animal life. About one-half of Thailand is forested, one-fifth covered by grass, shrub, and swamp, and the remainder under settlement or cultivation. Forests consist largely of such hardwoods as teak and members of the dipterocarp family of timber- and resin-producing trees. As elsewhere in Southeast Asia, bamboo, palms, rattan, and many kinds of fern are common. The southeastern forest is dense with undergrowth and aerial roots, while orchids climb the trunks of many trees. Grasses and shrubs have sprung up across many cutover areas, and lotuses and water lilies dot most ponds and swamps throughout the country.

Elephants, buffalo, cattle, horses, and mules are among the important domesticated animals for agriculture and transportation. Some elephants are still found wild in the west and the northeast. Although forestry is now done by machines, elephants remain helpful in difficult terrain. Agricultural machines are rapidly replacing draft animals, and horses and mules increasingly are used only as pack animals along the mountain trails of the north. Wild animals are decreasing in quantity and in number of species due to game hunting. The rhinoceros and tapir are almost extinct. Deer and antelope can still be found, but the big cats are becoming fewer in number. The Siamese cat is popular among foreigners.

Lizards live around houses and prey on insects. Frogs and toads (some of them edible) are numerous, and crocodiles are also found; a crocodile farm in Bangkok raises them for commercial and scientific purposes. Snakes abound, including the king cobra, numerous vipers, and dangerous water snakes. A snake farm in Bangkok raises many kinds for extracting venom. The green turtle of the sandy coast lays hundreds of eggs at a time, which are highly regarded as culinary delicacies.

Both freshwater and marine fish are abundant, as are edible crustaceans, such as shrimp, prawn, and sea crabs.

The most useful insect is the silkworm. Many species are wild, but some are raised for the prosperous silk industry. Mosquitoes still carry malaria, although incidence of the malady is decreasing. White ants and moths are a scourge to clothing and books.

Settlement patterns. In the traditional or popular sense, the regions of Thailand are the Lanna Thai, or northern Thailand; the Isan, or northeastern Thailand; the central plain; the southeast; the Pak Tai, or south Thailand; and west and southwest Thailand.

The people of the mountainous Lanna Thai speak a dialect similar to that of neighbouring Myanmar. The women wear simple blouses and knee-length sarongs; the men wear loose trousers and cotton jackets called *mohom.* Most of the people live on glutinous rather than common rice. The people have a complexion that is fairer than that of other Thai. The fertile valley plains grow rice and such other crops as peanuts, beans, garlic, onions, corn, and tobacco on the irrigated lands. Tea, widely cultivated as a permanent crop, is consumed in the north as both a pickled and a drinking tea. Besides rice, the outstanding export of the region consists of teak and other commercial timbers. The mountains of the north are also the home of the mountain Thai, who migrate from Myanmar, Laos, and sometimes from southern China.

The Isan is situated on undulating terrain bordered on the west and south by rather hilly regions. The people dress much as do the northern Thai, except close to the Cambodian border, where the *pha chong kraben*—a side-rolled skirt for women, also adopted as a belted pantaloon-like garment for men (similar to the Hindu dress of India) is worn instead of trousers. The dialect is like that of Laos, although very similar to that of the central Thai. With some attention, however, the northerners and the northeasterners can understand one another. A glutinous rice is consumed in most of the northeast, except near the Cambodian border. Besides rice, various kinds of timbers, jute, and cattle constitute important exports. In general, the people depend on agriculture, but the scanty rainfall and sandy soil make the living in the region difficult. As a result, migration from the northeast to other parts of the country is common. The highway to the northeast and thence to Laos makes the region a growing crossroads. It is probably the only part of the country to which the government is devoting both attention and money in an effort to raise the standard of living to that of the other regions.

ntral
ain area

The central plain, lying mostly on the undulating to flat delta region, is the major rice-producing area of the country, much of whose surplus is exported. Because of its central position, population and manpower are heavily concentrated, especially in Bangkok, and industrial and commercial enterprises have grown faster there than elsewhere. The eastern, western, and northern peripheries of the central plain, however, remain to be developed. The soil of the northern half of the plain is inferior. People from the northeast and other places have migrated to the central area in hope of a better living.

The southeast, lying close to the sea on an undulating to hilly region, is well watered. Close to the sea fishing predominates; plantations and gem mines are found on the inner hill and mountain slopes. Thai of Chinese descent live here in great numbers. Tapioca and sugarcane are the chief crops, and fish supply the Bangkok and the upper-country markets and are also exported. Chanthaburi and Trat are famous for their sapphires. The beaches are popular and attract foreign tourists.

The people of the Pak Tai region of the peninsula speak a dialect that has a different intonation from that of the central Thai. The extreme south is inhabited by Malay-speaking Thai, the majority of whom are Muslim. The southern economy depends on exports of tin and rubber. Fishing provides income for the people living close to the seas on both sides of the peninsula. Phuket and Songkhla are well-known for their fine beaches, which are lined with beautiful coconut groves.

The west and southwest, consisting mostly of hilly to mountainous terrain adjoining the Myanmar border, is sparsely populated. The forest covers are chiefly mixed deciduous and bamboo. The soils are rich with minerals, such as tin, tungsten (wolfram), and fluorite. The Karen from Myanmar often migrate and live within the Thai border. They clear the land and cultivate the hill slopes in the same manner as the mountain Thai in northern Thailand. The region is also the chief supplier of the bamboo used in the curing of rubber.

Hill settlements depend much on shifting cultivation of upland crops. Such mountain Thai as the Karen, Meo, Yao, Lahu, Lisu, and the lowland Thai who have migrated usually settle on the ridges and the slopes in groups ranging from two or three to 100 or more houses, depending on the resources of the area. The Meo are opium cultivators, preferring to live on high slopes where opium grows well in the cool climate. The Karen live along the stream valleys and grow rice on well-tended, terraced fields. The Thai who have migrated earn a living from their tea and coffee plantations.

ral and
an
ailand

No true plains villages exist in Thailand, although in the northeast the villages are scattered on the higher grounds to escape the floods; the lower grounds are used for rice farming. In the north, where the villages are found on the alluvial basins of major rivers, increased population and transportation have tended to disperse the villages away from the river toward the main railroads and highways, reducing the amount of rice-growing land.

The delta is densely settled, but only on the high ground that is free from flooding. The irrigation canals modify the pattern of settlement; with increasing facility in transportation offered by small motorboats, the villages tend to become dispersed away from the rivers in an east–west direction. New highways also tend to modify the pattern of settlement, especially at the crossings of canals and rivers by highways, where new towns grow up rapidly.

The southern and southeastern plantations, especially the fruit and rubber plantations, are scattered along the fertile slopes, alternating with the low and narrow rice fields, and the villages are therefore arranged accordingly. Most of them are joined by good roads and highways. Alluvial deposits containing tin, no matter how isolated, are accessible to all kinds of land and sea communications. Settlement is almost continuous along both sides of the peninsula. Most of the people live by fishing, except in areas in which bird's-nest collecting (the jelly-like adhesive of which is the basis of a popular soup or sweet dessert) brings a good income. The coastal villages are connected both by land and sea highways.

Urban settlement in Thailand is found mostly on the plains and in the coastal areas rather than on the hills. Bangkok, for example, grew from a small settlement on the east bank of the Chao Phraya. In other large towns, such as Chiang Mai, the old, square city walls are still apparent, with numerous Buddhist temples scattered inside and outside the walls. Thus, urbanization in Thailand can be said to centre around the original sites of the palaces and the temples. Other interesting patterns of settlement can be observed in certain urbanized areas. Phitsanulok, for example, has a number of floating houses along both banks of the rivers. Chon Buri, a seaside town, has a trellis pattern of houses on islands that are accessible by footbridges. Such patterns of settlement reflect the strong demand for living space in the urban areas.

THE PEOPLE

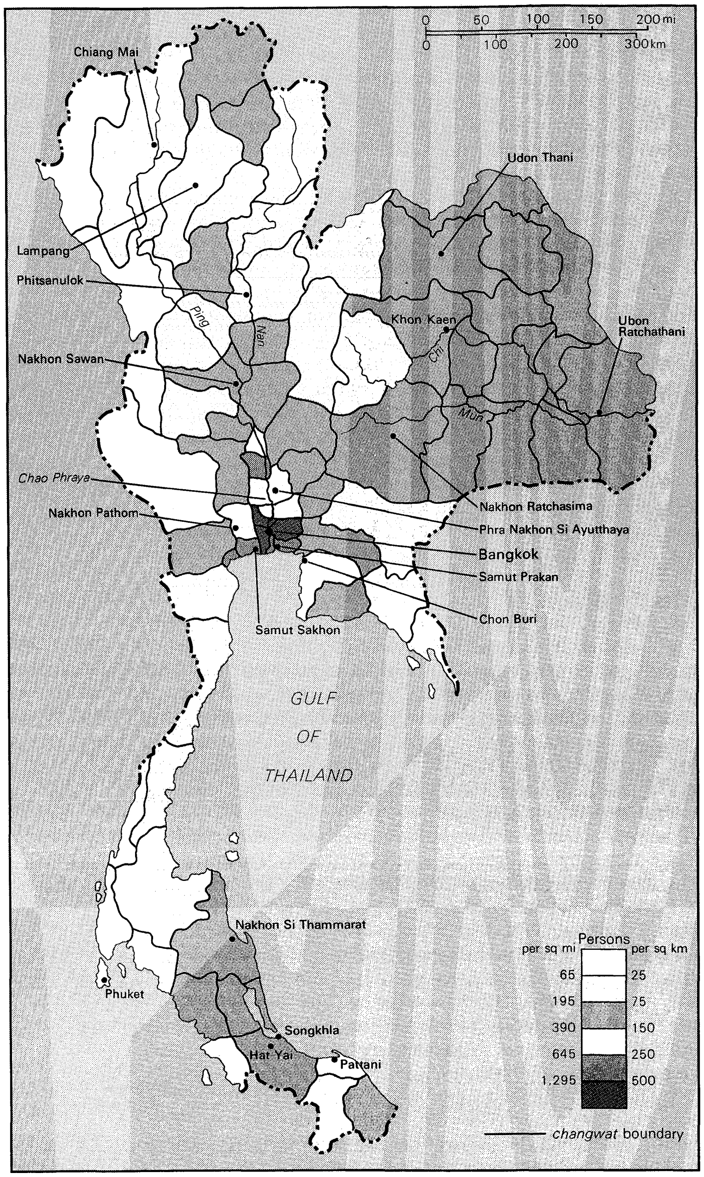
The diversity of ethnic, linguistic, and religious groups in Thailand is characteristic of most nations of Southeast Asia, where shifting political boundaries have done little to impede the centuries-long migrations of people.

Geographical
dispersion
of Thai
peoples

Ethnic distribution. Thai people are found not only in Thailand but also in Laos, Myanmar, India, Vietnam, and southern China. These Thai speak the Thai language, but with different accents and a few different words. Little difference exists between the Thai in Laos, Myanmar, China, and northern Thailand, but there is a noticeable difference between them and the Thai living in the central plain and close to Cambodia. The peninsular Thai are much influenced by the Malay.

Wars between Thailand and Myanmar in the past brought many refugees and prisoners of war into Thailand. The Mon, a Myanmar people, settled in many parts of the north, the centre, and the west; their distinctiveness is evident in their festivals and religious rites. Even the original Thai are not completely free from ethnic intermixture, and the ethnic origin of the modern Thai is even more complex when the Chinese and the Indian descendants are considered. Bangkok is the chief melting pot of the Thai race. Tracing true Thai blood in Bangkok should not be attempted, as a famous Thai scholar (himself of Chinese descent) has said, for what makes the Thai is the institution.

Except the Karen, who mixed rather easily with the northern Thai, the hill tribes, or the mountain Thai, prefer to keep themselves isolated. They occasionally come



Population density of Thailand.

By courtesy of the Royal Thai Survey Department

down to the markets to trade with the lowlanders. They are probably of purer stock than the modern Thai. Two small hill tribes are of special interest, the Lawa and the Semang. The Lawa are believed by some historians to be the original dwellers of the delta plain, driven into the hills of the northwest by the Thai who conquered the area. The Semang of the southern mountains live by hunting with blowpipes and spears. Another ethnic group that often escapes attention is that of the Chao Nam, or sea dwellers. Rarely settling permanently, they live by fishing along the western coast and the adjacent islands of the peninsula.

The mix of tongues and faiths

The Thai, who live in almost all areas, comprise the majority linguistic group. The greatest concentration is in the Chao Phraya Delta. Myanmar, Laotian, and Cambodian influences already have been noted. The people of Tak and Kanchanaburi, near the southwestern Myanmar border, speak Karen and Mon. From Chumphon to the south, the Thai speak a southern dialect. Most of the Thai have been in contact with one another, especially administrative officials and teachers, and the Thai living in remote areas can understand the central Thai.

Chinese is the second major language. In the commercial centres of Bangkok and other cities, Chinese or their descendants operate both large and small commercial enterprises. Those of Chinese descent also make a living as middlemen and storekeepers. Most of them speak Chinese, as do some of the northern hill tribes, such as the Meo and Ho.

Although no reliable statistics exist, English-speaking Thai probably make up the third major linguistic group. English is a required subject in secondary schools and the university, and frequent contact with American military personnel also encourages the speaking of English. The prevalence of various Indian dialects reflects the large number of Indian merchants and their descendants in the commercial centres. Other linguistic groups are found among the mountain Thai. Some of them can speak Chinese, although most of them understand the northern Thai dialect.

Brahmin religious influence

Religion. The religious groups in Thailand, in order of size, are the Buddhists, the Muslims, the Confucians, the Christians, and the Sikhs. Buddhism, professed by the vast majority of the population, is considered the national religion, and Buddhists are scattered throughout the nation. Muslims live mostly in the south, whereas about one-half of the Christians live in the central region. Hindus and the Sikhs are concentrated in the central region, chiefly around Bangkok. Although several of the hill tribes have converted to Buddhism or Christianity, most remain animists. An interesting religious group in Thailand, although totalling only a few thousand families, exerts much influence on the Thai religious life. These are the Brahmins. The royal and the official ceremonies are almost always directed or performed by the Brahmins, whose rites are blended harmoniously with those of the Buddhists. Brahmins are famous for their astrological experiences. The plowing ceremony, carried out in the presence of the king and queen by Brahmins and other officials, is believed to bring a good rice harvest. The Brahmins and the royal astrologers are also responsible for the preparation of the national calendar.

THE ECONOMY

Thailand's economy is still based on the production of basic agricultural, mineral, forest, and other raw materials. The Thai unit of currency is the baht.

Resources. There are numerous tin mines, most of them in the peninsula, making Thailand the world's third-largest producer of tin. Construction of a smelter has made possible domestic smelting of most of the ore. Iron-ore production has risen. Other important mining and quarrying operations produce gypsum, fluorite, tungsten, limestone, and marble.

Primary and fabricated products

Agriculture, forestry, and fisheries. Growing demands have encouraged a stronger and more diversified Thai agriculture. Rice is likely to remain as the major crop, unless radical dietary changes occur in Asia. Sweet corn, cassava, and plants yielding fibres are other major crops. Fine breeds of cattle and pigs have been introduced from the West. Hardwoods, such as teak and yang (a source of gurjun balsam), are major forest products; rubber trees—introduced into the country during the 19th century—are important. Fishery includes both marine species and freshwater fish caught in the rivers or in cultivation ponds.

Industry. So far, manufacturing does not play a large part in Thailand's economy. Most industries are involved in processing agricultural, forest, and mineral products. The main centre for large industry is the Bangkok-Thonburi metropolis, while numerous cottage industries in the north produce textiles teak, carvings, lacquer ware, and similar products. In the southeast, in addition to food processing, oil refining and gem cutting are carried on.

Thailand has several hydroelectric plants, but most electrical power is produced by generating plants using gas and solid fuels, such as lignite, which is mined in Thailand. Small amounts of petroleum are produced domestically, and the government has granted concessions for offshore explorations in the Gulf of Thailand.

Finance. The Bank of Thailand, established in 1942, issues currency, acts as central banker to the government and to the commercial banks, and serves as fiscal agent in dealing with international monetary organizations. Nearly one-half of the nation's retailing and other distribution businesses are located in Bangkok-Thonburi. Middlemen handle most farm commodities. Retail stores are small, except in Bangkok, which has several large department and cut-rate stores. In these, as well as in most food markets, prices are fixed; bargaining occurs mostly in the souvenir and gift shops of larger cities.

Trade. With agricultural and raw materials the basic exports, manufactured goods, such as machinery and transportation equipment, account for the highest value among imports. Thailand's major trading partner, both for exports and imports, is Japan; most other exports are sent to other Asian nations, whereas most other imports are from Europe and the United States.

overn-
ient and
conomic
anning

Administration of the economy. Aiming at diversification, the government has supported a greater number of small industries. To encourage exports, duties are low, except on rice, to which a premium is attached to prevent domestic shortages.

Unions are prohibited and strikes are not allowed unless management representatives fail to agree with employees and Labour Department mediators.

Transportation. Bangkok is the centre of Thailand's water, land, and air transport. The rivers of the delta have been used for transport since antiquity, and modern irrigation canals have added to the inland-waterway mileage. Because of the rains, it is difficult to keep some highways open all year, especially in the peninsula. Mountain trails are often the only means of travel in remote areas. Where roads are inadequate, airplane and helicopter services often compensate. Rail lines radiate from Bangkok in several directions, one linking up with the Cambodian rail system.

The port of Bangkok, at Khlong Toei, is the largest and busiest of the 22 in the kingdom, handling nearly all imports and exports. Don Muang Airport, north of Bangkok, is served by many international airlines, including Thai Airways International, a state-owned line. More than 20 smaller airports are located throughout the country.

ADMINISTRATIVE AND SOCIAL CONDITIONS

Government. Following the revolution of 1932, a provisional constitution was promulgated, stating that supreme power lay in the hands of the people. The monarch, the National Assembly, the Commissariat of the People (later, the State Council), and the law courts were to exercise power on their behalf. Since then, several constitutions have been created because of changes of government, but the provisions are similar.

onstitu-
onal
amework
d legal
stem

Under the present constitution, the king is head of state and of the armed forces. He is held to be sacred and inviolable, and in the name of the people he exercises legislative power, with the advice and consent of the Assembly. He also appoints the prime minister. Executive powers are vested in a Council of Ministers, judicial powers in the courts; both operate in the name of the king. The royal family is very much at the core of modern Thai society, being regarded as the symbol of national unity and the protector of national welfare and traditions.

In form, the Thai government resembles those of Western nations; various ministries are responsible for such matters as finances, agriculture, education, public health, communications, and justice.

The province, or *changwat,* of which there are 70, is the major unit of local government. Beneath these are districts, subdistricts, and communes and villages. The 120 municipalities in the kingdom are classified as cities, towns, or communes, according to their populations; they are run by an elected mayor and councillors.

Justice. Thai law has been influenced by the Hindu code of Manu, which probably was transmitted through the ancient Mon kingdom of central Thailand. Reform in the late 19th century introduced concepts of Western jurisprudence. All judges in the country's 110 courts are professionals, appointed without political consideration; they are bolstered by a system of judge trainees.

Armed forces. Under the king as commander in chief, the army, navy, and air forces are assisted by the Military Assistance Program, and by the Southeast Asia Treaty Organization (SEATO), of which Thailand is a member. Thai soldiers have fought in Korea and Vietnam, but lately they have been fully occupied in containing Communist infiltration from neighbouring countries.

Education. Children under 15 are required to complete seven years of elementary education. Secondary education generally lasts five years, but when it is vocational it lasts six. Only a small minority of students go on to secondary training, however. Ramkhamhaeng University is one of Thailand's newest universities. Eight other institutions, not including military academies, offer degrees in undergraduate and postgraduate fields.

Health and welfare. Thailand's health and welfare services remain far from adequate. The emigration of potential medical practitioners to more lucrative practices in the West has tended to undermine governmental attempts to upgrade services within the country. Mobile medical centres and helicopters attempt to alleviate the lack of widespread facilities. The doctor–patient ratio is quite low, and medical practice on the midwife level is common. Infant mortality and diseases of childbirth are leading causes of death, whereas malaria has been widely reduced through the use of DDT, which has hastened the devastation of forest areas as a side effect.

Only Bangkok-Thonburi and a few other large municipalities have housing shortages. The construction of government-financed housing cannot keep pace with demand, and slum areas have proliferated in some parts of the city.

CULTURAL LIFE

According to many historians, the Thai's original home was in China, perhaps as far north as Mongolia. The Thai brought with them many cultural institutions of the Chinese. They began settlement in the Indochinese Peninsula about 800 years ago, at which time some Indian colonies already were established. Indian culture has been continuously absorbed into Thai life. Modifications were affected by the cultures of the Mon people, the Javanese, and the Khmer and Burman people.

Traditional
arts and
customs

Thai arts are reflected in religion, architecture, porcelain and pottery, painting, music, drama, and literature. In religion artistic expression can be found in the sculpture of Buddha images. Thai architectural style is to be seen in the temples. Wood is usually the basic construction material. The ornamental parts are generally gilded and enriched with glass mosaic, gold leaf, porcelain, stucco, lacquer, and inlaid mother-of-pearl. The multiple-structured temple ground is a paradise for architects. Porcelain and pottery, although at first put to utilitarian uses, were later regarded as objects of art. Thai painting probably derived from India and Ceylon and is mostly religious; the artists are anonymous monks or dedicated laymen. The paintings are usually drawn on the temple walls, which are constructed with bricks and plaster.

Thai music is based on a unique system. It is not derived from the Chinese or Javanese systems, although the instruments used for playing may look the same.

The royal palace plays an important role in leading and preserving Thai culture through frequent royal functions and state ceremonies. Among these is the Kathin ceremony, a colourful pageant marking the end of the Buddhist Lent. It takes place with a procession of royal barges on the Chao Phraya, reconstructing a tradition dating from the earliest days of Buddhism. Thai temples hold ceremonies to mark the special events of the Buddha's life. These are often accompanied by fun fairs to attract large crowds to the temples.

The University of Fine Arts teaches all Thai fine arts, including drama and music. The university also designs architectural structures for the government as well as the religious institutions in a style that will preserve Thai forms. The Royal Institute of Thailand and the Siam Society are responsible for research and publication concerning the Thai way of life. The National Museum acts as an educational and information centre for the evolution of culture in the country.

The first type for printing Thai letters was devised by a British captain in 1828, and the first printing press was brought to Thailand by an American missionary in 1836. The Thai government made use of the printing press for the first time in 1839, when a royal proclamation banning opium smoking and trade was printed. (P.P.A.)

For statistical data on the land and people of Thailand, see the *Britannica World Data* section in the BRITANNICA WORLD DATA ANNUAL.

History

The present state of Thailand was officially known as Siam before 1939 and from 1945 to 1949.

The Thai peoples emerged out of historical obscurity about a thousand years ago when they began moving southward and westward from the China landmass to the northern parts of the Indochinese Peninsula. They moved into lands, from Assam in the west to northern Vietnam in the east, where most of the indigenous peoples were already under the influence of Indian culture.

EARLY THAI CULTURE

General characteristics

The Thai culture before this time cannot be precisely delineated, since there are no documents describing it. It was once assumed that the Thai peoples were the founders and rulers of the kingdom of Nanchao, an independent state that flourished from the 8th to the 13th century in Yunnan Province of China, but later studies have indicated that the rulers of this state were not Thai. The culture of the Thai, then, can be inferred only on the basis of elements that continue to exist among most peoples speaking Tai languages today. The Thai language itself, in various dialects, is spoken by a wide variety of people living today in Myanmar, Thailand, Laos, Vietnam, and southern China. The Thai are people of the lowlands; their homes are in river valleys, and their economy is dependent on river waters. They are subsistence farmers whose principal crop is rice. In the rice fields, their technology includes the use of the plow and the harrow for breaking the land and the hand sickle for harvest. Individual landholdings are typical, but cooperative efforts to build dikes or to harvest fields are common. Animals are domesticated, including the water buffalo and ox for labour and the pig and fowl for food.

Family relationships are important among the Thai, but they have no concept of the extended family or clan and no caste system. Women occupy a high position; they may inherit property, and widows are free to remarry.

The earliest Thai religious beliefs were animistic. Spirits were thought to inhabit various natural sites—large trees, high rocks or mounds, the living quarters of families, and the rice kernel, the habitat of the rice goddess. Other spirits were the ghosts of the dead. Spirits could be benevolent or malevolent; offerings were made to them and ceremonies performed to ensure their favour.

When the Thai settlers in Southeast Asia began to emerge into history, they possessed sufficient social cohesion and political order to establish a number of petty principalities. Thai chronicles, although written long after the events they describe, point to the emergence by the 13th century of several small Thai states in what is now northern Myanmar, Yunnan Province of China, and Assam in India, as well as in Thailand. In the 13th century several of these states rose to prominence, thanks partly to the rise of the Mongols, who sent armies to defeat Myanmar, Vietnam, and Champa but accepted Thai vassal status, and partly to the weakened condition of the Indianized states that had long been powerful in the area.

Mons and Khmers

In the centre of the Indochinese Peninsula two peoples, the Mons and the Khmers, bore the impress of the culture of India in religion, government, the arts, and literature. The Mons, in southern Myanmar and central Thailand, and the Khmers, in the Mekong Valley, had created distinctive and vigorous cultures and powerful political states. By the 13th century, the Mons had suffered from the depredations of the stronger Khmers, retaining political independence only in a relatively small state, Haripunjaya, in what is now northern Thailand. Mon culture, however, continued to be strong throughout central and northern Thailand. Particularly noteworthy was the devotion of the Mon peoples to Theravāda Buddhism. The Khmers, after an extraordinary burst of military and cultural expansionism under King Jayavarman VII (reigned 1181–*c.* 1219), began to decline in political power. The times were propitious for the emergence of the Thai.

SUKHOTHAI AND CHIANG MAI

Two principal Thai states emerged during the 13th century in the area that is now Thailand: Sukhothai in the Central Plain and Chiang Mai in the north. Sukhothai was founded around 1220 after a Thai revolt against Khmer rule at an outpost of the Khmer empire. Chiang Mai was founded by King Mangrai (reigned 1296–1317) in 1296, shortly after the Thai defeat of the Mon state of Haripunjaya.

From the outset, both Thai states showed that the Thai had already absorbed many elements of the cultures of the peoples around them. There are, in fact, reasons to assume that the political emergence of the Thai did not represent any drastic change in the ethnic character of Thailand. It seems likely that the Thai, who for centuries had been filtering into the area, at the time of their rise were a minority of the population but were able to rise to dominance because of superior organizational ability and military prowess. The subsequent expansion of the Thai may reflect their ability to retain social supremacy and to absorb large numbers of non-Thai peoples.

Among the cultural borrowings of the Thai, one of the most profound and important was their acceptance, from the Mon, of Theravāda Buddhism. Thai animistic beliefs did not disappear—they persist even today—but to them were added the humanistic and philosophic concepts of a popular and egalitarian world religion.

Sukhothai under King Ramkhamhaeng

Sukhothai, regarded by the Thai today as their first historical kingdom, achieved a considerable expansion by the end of the 13th century. Little is known of Sukhothai during its first two reigns, but its third king, Ramkhamhaeng (reigned 1279?–?1317), has left a remarkable stone inscription that is the earliest example of writing in the Thai language. The Ramkhamhaeng inscription, dated 1292, pictures the Thai state as rich and prosperous, active in trade, and benevolently governed by a paternal monarch. According to the inscription, the state taxed its citizens modestly; treated all citizens, including non-Thai, alike; and provided justice for all. And "all its citizenry, all chiefs and nobles, everyone everywhere, all men and women, all groups, believe in Buddhism." The kingdom is described as extensive, covering much of central Thailand down to Nakhon Si Thammarat on the peninsula, but excluding the plain of the Chao Phraya River below Nakhon Sawan, a region once dominated by the Khmer but apparently during Ramkhamhaeng's reign governed as an independent state called Lavo.

Allowing for some royal exaggeration of the virtues of the state, much of Ramkhamhaeng's description rings true. It seems clear that the Thai rulers of Sukhothai were proud of the differences between their social order and that of their Khmer predecessors and adversaries. The emphasis on their human and humane king, on their low tax burden, on their popular Buddhist faith was undoubtedly meant to point out the contrast with the very different "god-king" status of the Khmer monarch, the heavy exactions in taxes and labour of the Khmer state, and the high cost and remoteness of the aristocratically oriented Khmer Hindu cults.

Although much of Sukhothai "high" culture was a result of cultural borrowing—Buddhism from the Mon, the system of writing from the Cambodians, artistic techniques and styles from various sources—many of these cultural borrowings were transmuted into Thai forms. Particularly outstanding were Thai bronze sculptures of the Buddha that were cast with great skill in distinctive and original forms. Thai ceramics, inspired apparently by Chinese Sung ware, were produced in great quantity at kilns in Sawankhalok. The high quality of these ceramics made them prized in trade.

King Ramkhamhaeng combined the abilities of effective ruler, warrior, and patron of Buddhism and the arts. His successors did not have his range. His son, Loe Thai (reigned 1317?–?47), the fourth king of the dynasty, poured his energies into religion, and a weakening of the state was the eventual outcome. Sukhothai apparently regained some power under King Lu Thai (reigned 1347?–74), but it fell under the domination of a new Thai state to the south in 1378.

THE AYUTTHAYA PERIOD (1350–1767)

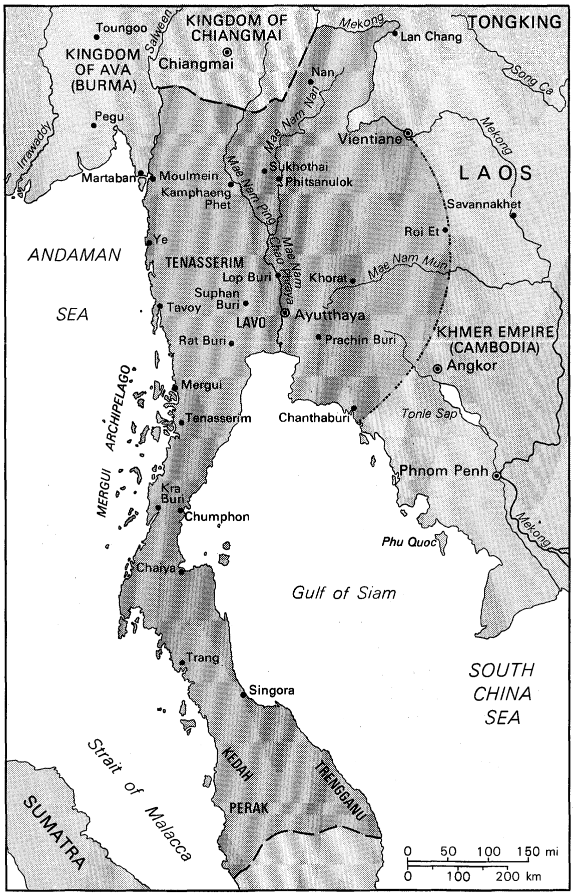
The successor state to Sukhothai was the Thai kingdom of Ayutthaya that had emerged around 1350. According to tradition, Ayutthaya's first king, Ramathibodi I, inherited the states of Lavo and U Thong (their capitals correspond to the towns of Lop Buri and Suphan Buri in Thailand today) and combined them into a single strong kingdom. Ayutthaya was located in the centre of the rich rice plains of the Chao Phraya Basin, a region strongly subject to the influences of the Khmer, the Mon, and the Burmese. From the outset it appears to have competed successfully for power with Sukhothai and with the Khmer. By the early 15th century it had absorbed Sukhothai and devastated the eroding Cambodian kingdom.

During the first century after the foundation of Ayutthaya, many changes reflecting the influence of Cambodia occurred in Thai society, particularly on the upper levels. Under King Trailok (reigned 1448–88) further changes were made and former changes were codified; the character of the Thai state was permanently transformed.

yut-
hayan
overn-
ment

The most significant of the changes concerned the nature and organization of government. The self-image of Sukhothai kings had been that of a father bringing riches, justice, and the benefits of Buddhism to his people. In Ayutthaya these ideals were expanded and overlaid with the Khmer view of the ruler as god-king. Court ceremonials and court Brahmins were introduced to bring to kingship the magical attributes of divine intermediary. The court calendar swelled with ceremonies for the manipulation of cosmic forces to promote the success and prosperity of the state. (These ceremonies are described in a voluminous [718-page] 19th-century work in Thai by King Chulalongkorn, "The Royal Ceremonies of the Twelve Months.") In a practical way, new techniques for centralizing power, including a bureaucratic system of government, were introduced. Members of royalty who had been serving as provincial officials were replaced by appointed nobles. A hierarchy of princes and nobles was established whereby the relative status of each was formalized. A system of patrons and clients extended throughout the population so that, theoretically, every man fulfilled obligations—in loyalty and goods and services—to someone above him and received, in return, protection from a patron. Commoners paid their military service, compulsory labour services, and taxes to local masters who, in turn, were subject to higher masters above them. The apex of the pyramid of patronage was the king himself.

From *Atlas of South-East Asia*



Ayutthaya kingdom *c.* 1400.

The formal hierarchical system awarded each man *sakdi na,* or "dignity marks," in which the position of each man was equated with an acreage of land. Ordinary freemen were actually allowed to take possession of twenty-five *rai* of land, equal to about ten acres, to represent their *sakdi na.* The rank of officials of the government and members of the royal family was also expressed in the acreage terms of *sakdi na;* the acreage indicated could number in the hundreds or thousands. But for nobles and princes the term stood for power and position, for "dignity," and not for actual landed estates. High royal and noble officials were, in fact, required to live and work in the capital.

Many of the principles of government were enacted into codes of law, and a large body of law is reputed to date back to the reign of King Ramathibodi I. The basis of law was Hindu, and the codes were probably transmitted to the Thai from Mon sources. These codes were expanded a century later. They included laws on evidence, on slavery, on robbery, on divorce. A Palace Law spelled out the duties of members of the royal family, rules of succession, and court procedures. The institution of the rank of *uparat* also dates back to the mid-15th century. Often inappropriately termed "Second King" by Western visitors to Thailand, this rank was Indian in origin and was, in fact, the highest princely rank awarded to a member of the royal family. The *uparat* functioned as general of the armies and had some claim to being the heir apparent.

By and large, in the context of the times and in comparison with neighbouring states, Thailand in Ayutthaya was a strong and well-governed state. There were, however, periodic internal disorders, products, it would seem, of contests for power between and within the royal and noble elites. In order to retain power a king had to preserve a balance of strength between princes and nobles. The prime ingredient for power was population, for land was plentiful and people were scarce. If a king allowed too many clients to fall under the control of his relatives, the princes, or even of his own bureaucratic officials, the nobles, political conflict would likely ensue. Many such conflicts are recorded—one of the bloodiest resulting from the usurpation of the throne by a high noble early in the 17th century. This noble, whose position was essentially that of minister of the Southern Provinces, paved his way to the throne by putting two kings to death, and after his accession as King Prasatthong, some 3,000 persons, members of the former ruling dynasty and other potential rivals, were executed. One prince who had followed the established practice of seeking sanctuary by becoming a Buddhist monk was first lured out of his monastery and then put to death.

Ayutthayan society and culture

The mass of the people in Ayutthayan times, the peasant farmers, were divided into two main classes: freemen and slaves. Freemen had the right to till an area of land equivalent to ten acres. The obligations of the freemen to government in taxes and labour were paid to local patron-officials. This link was all-important. The way to improve one's status was to gain special favour through a patron or, if he were unable to help, to find a new patron who might give special favour. Exactions on freemen, it would seem, were seldom excessive. The fact of underpopulation meant that freemen could escape onerous obligations or an overly demanding patron by moving and clearing new land; the existence of numerous decrees against this practice testifies to its existence and prevalence. Another "escape" for freemen was slavery. There were many so-called slaves in Ayutthayan times; the term slave, however, is misleading—the great majority of slaves could better be called bondsmen, for the "slave" was usually also a peasant-farmer who could own land and could redeem his status by repayment of his debt.

Theravāda Buddhism in Ayutthayan times continued to supply an element of social cohesion in the Thai state. From king to peasant, all Thai regarded themselves as Buddhists despite the introduction of court Brahmanism on the top levels of society and the continued persistence of animism on all levels. Buddhism, indeed, was thoroughly acculturated in Ayutthayan times. On the popular level, Buddhist amulets and reliance on monks with magic power interpenetrated the more sublime ideas of the doctrine. At court, one king, Trailok, in 1458 cast 500 images of the Buddha in order to prevent repetition of the great famine that had occurred in 1457. The Buddhist monastic establishment itself played an important social role, providing moral and practical education for boys and giving young men who spent some time as monks the opportunity to "gain merit" for themselves and their parents. For men who chose to remain monks, the monastic order constituted a society parallel to secular society that allowed for considerable social advancement.

In its relationships with outside states, Ayutthayan Thailand was comparatively strong. Warfare was endemic, and the sway, or empire, of Ayutthaya fluctuated considerably during its 400-year history. But Thai dominance of the central Chao Phraya Plain remained unchallenged. Most of the wars were petty campaigns during the dry season, when the terrain permitted the movement of men and the agricultural cycle made their absence from the fields possible. Most of the wars were contests for sovereignty over border principalities. Some were essentially raids to acquire population—*i.e.,* agricultural manpower and skilled artisans—and transportable wealth.

Ayutthaya's emergence in the 14th century in an area once ruled by Cambodia brought on a series of wars with Cambodia. By the early 15th century Ayutthaya had reduced Cambodia to vassal status, and, although Cambodia was frequently able to throw off its dependency, the balance of power between the two states thereafter continued to favour the Thai. Ayutthayan dominance over Sukhothai, achieved early in the 15th century, brought the kingdom into confrontation with the northern Thai states of Chiang Mai and Laos. During the reign of King Trailok these wars reached a peak of intensity, and Ayutthaya went to war with Chiang Mai seven times. In order to improve his military position with respect to Chiang Mai, King Trailok in 1463 moved his capital to the northern town of Phitsanulok, where it remained throughout the rest of the reign. Ayutthaya was unable to retain power over Chiang Mai, and Chiang Mai had periods of independence and periods of dependency on Myanmar throughout the Ayutthayan period; it was not until the late 18th century that Thailand finally succeeded in gaining permanent control over Chiang Mai.

Ayutthayan–Myanmar rivalry

The most powerful rival of the Thai in the Indochinese Peninsula was Myanmar. On two occasions when the ethnically diverse area of Myanmar was welded together into a political whole, Myanmar imperial expansion extended into Thailand. One series of campaigns started around 1549. In the first engagement of this series, according to the Thai chronicles, the Thai king Maha Chakkraphat (reigned 1548–69) found himself in difficulties and was rescued by his wife, Queen Suriyothai, who lost her life in the incident. The campaigns ended in 1569 in a Thai defeat, and the country was overrun by Myanmar forces. Myanmar troops were garrisoned in Thailand, and the Myanmar placed on the throne a Thai king who accepted their suzerainty. Thailand remained a Myanmar vassal-state for 15 years. In 1584–87, a Thai prince, Naresuan, since regarded by the Thai as a national hero, led a successful revolt against Myanmar and reestablished Thailand's independence. In another period of Myanmar unity and expansionism in the 18th century, they succeeded again in 1767 in capturing the Thai capital at Ayutthaya. The Myanmar armies did not remain for long; they withdrew after sacking the Thai capital and thus ending the Ayutthaya period of Thai history.

Ayutthaya's relations with most other Asian states were limited to trade and the exchange of occasional embassies. The pursuit of friendly relations with China, started in Sukhothai times, was carried on throughout the Ayutthayan period. These relations were typical of those existing between China and outlying states: the Thai acknowledged the "elder brother" position of China and sent tributary missions triennially to the Chinese court. This formal suzerain–vassal relationship had little political significance but was willingly maintained by the Thai for the very real benefits it brought in the way of trade. Thailand, as a registered vassal, was permitted to send ships to China to purchase silks, pottery, and other luxury goods that were much in demand among the Thai upper classes. Thai trading vessels, and Chinese ships emanating from Thai ports, also conducted a lively trade in such commercial centres as Malacca, Vietnam, Java, and India.

Relations of a different sort were maintained with Ceylon; they were religious contacts. Some of the principal religious leaders in Sukhothai had come from Ceylon. Ceylon was looked to in Sukhothai and Ayutthayan times as the font of Theravāda Buddhist knowledge, and several religious missions were sent to Ceylon to obtain religious instruction, religious texts, and religious relics. During the 18th century, Buddhism in Ceylon underwent a serious decline, and a reversal in the religious relationship occurred. Ceylon then turned to Thailand for aid in purifying its ordination and other religious practices. One Thai mission that went to Ceylon in 1753 stayed for three years and performed some 700 ordinations for monks and 3,000 for novices. The Thai missions of this period established a Thai school of Buddhism in Ceylon, called Syāmvaṃsa, that still exists.

European contacts

In long-range terms the most significant foreign contacts of Thailand—and the rest of Asia—were those with Europeans, who began arriving in the 16th century. By any standard, however, the European impact on Thai society in Ayutthayan times was slight. Portuguese diplomats, traders, and missionaries began arriving at the court of Ayutthaya shortly after the Portuguese capture of Malacca in the Malay Peninsula in 1511. Portuguese trade and missionary activities never became very important; the largest group of Portuguese in Thailand were adventurers who assumed roles as mercenaries in the Thai armies and who introduced some Western military techniques to Thai warfare. In the 17th century the Dutch and English began to arrive and to establish trading centres near the capital and in peninsular Thailand. The growth of Dutch trading interests throughout Eastern Asia was reflected in the emergence of the Dutch as the leading Western traders in Thailand. By 1664 the Dutch, by threatening the use of force, had obtained a treaty giving them monopolies over certain portions of Thailand's external trade. The Thai, seeking to offset Dutch power, welcomed the arrival of French missionaries in the 1660s, and in 1684 initiated a series of diplomatic exchanges between the courts of Ayutthaya and Versailles. The principal French spokesmen for promoting expansion of French interests in Thailand were clerics who, misunderstanding the characteristic Thai attitude of religious toleration, became convinced that the Thai king Narai (reigned 1657–88) could be converted to Catholicism. The arrival in 1687 of a large French mission that included some 600 soldiers aroused Thai fears, and in 1688 a coup by anti-Western Thai leaders led to the expulsion of the French and the beheading of their chief advocate at the Thai court, Constantine Phaulkon, a colourful Greek adventurer who had become a Thai noble and a leading counsellor of the king. The Thai coup of 1688 inaugurated a new Thai policy of minimizing contacts with Westerners—a policy that was to continue for 150 years.

THE THON BURI AND EARLY BANGKOK PERIODS (1767–1868)

The Ayutthayan Period in Thai history came to an end as a result of the successful Myanmar invasion of 1767. A new period began with the rise of a provincial noble, Phraya Taksin, who succeeded in ousting the Myanmar and eliminating Thai rivals to political power. Taksin established a new capital for the kingdom at Thon Buri, on the Chao Phraya River some 40 miles south of the ruined city of Ayutthaya. The new capital was somewhat less accessible to Myanmar armies and more accessible to foreign merchants—principally Chinese—who brought desired products and whose trade could be taxed to provide significant revenues for the government. King Taksin followed a policy of political reunification, including the reclaiming of vassal states that had broken off from Thai suzerainty at the time of the Myanmar invasion. A campaign in Cambodia led by Thailand's leading general, Chao Phraya Chakkri, was under way in 1781 when a palace coup took place in Thon Buri. Taksin, who had apparently become mad and imagined himself an incarnate Buddha, was deposed. Chakkri hastily returned to the capital, where a group of officials invited him to assume the throne. He did so in 1782, thus establishing a new dynasty, the Chakkri dynasty, that continues to head the Thai state. One of the first acts of the new king was to move the capital across the Chao Phraya River to Bangkok.

he
ırly
hakkri
ıngs

There is much continuity in the history of Thailand from Ayutthaya to Thon Buri-Bangkok. The ethos of the state, the customs and livelihood of the people, underwent no marked change. Taksin and the first Chakkri kings were primarily engaged in a policy of restoration both internally and externally. Law codes, religious works, and literary texts were rewritten. New temples and palaces were built using the patterns, and even the very bricks, of old Ayutthaya. The first three Chakkri kings were all restorers. Rama I (reigned 1782–1809) continued the policy of military reunification and brought Cambodia and other lost territories back into vassal status. He also gathered whatever evidence survived from Ayutthaya to re-establish court rituals and to set down comprehensive law codes and authoritative texts for the Buddhist canon. Rama II (reigned 1809–24) paid much attention to arts and letters. The King himself was an outstanding poet and, together with a team of court poets, wrote definitive Thai versions of the Hindu *Rāmāyaṇa* and other classical works. Rama III (1824–51), during whose reign the British defeat of the Myanmar kingdom (1826) ended the Thai–Myanmar rivalry, was able to expand the Thai empire in the south and in the north. In the south he strengthened Thai control over vassal states in the Malay Peninsula; in the north he was able to absorb most of the territory of the Lao kingdom of Wiangchan (Vientiane) after he put down a revolt by the leader of that vassal state, Chao Anu. Rama III's armies also fought several campaigns in Cambodia against the Vietnamese, who were extending their power into Cambodian lands during the early 19th century. These campaigns ended in 1845, and negotiations that followed led to an agreement that Cambodia should be a tributary state to both Thailand and Vietnam.

Although the restoration of Ayutthayan society was a primary goal during the early Bangkok Period, there were differences between Ayutthaya and Bangkok. For one thing, restoration is conscious traditionalism, which is not quite the same as tradition itself. For another, the policy of restoration necessarily led to choices, choices that the Thai may have presumed were made on the basis of authority but in fact were often made on the basis of utility. The reform of the law code in 1805, for example, began after the king, Rama I, discovered in a contested divorce case that the law was not just. In the traditional Thai view, law is of divine origin; it is immutable, and the role of government is not to enact law but to administer it. An unjust law, then, could only be explained as a false law. The reform of 1805 was therefore seen as a work of removing falsifications, although in fact it was a work of reformation.

Some basic social changes in the early Bangkok Period seem also to have been introduced in order to prevent the recurrence of the internal weaknesses and political factionalism that had contributed to Ayutthaya's downfall. Labour-service requirements on freemen were reduced, and payment of obligations in money rather than services was encouraged. These moves can be seen as an attempt by the monarchy to focus greater strength in the royal house, to bring more of the population under the patronage of the king rather than that of the princes.

Other changes in the early Bangkok Period came from the outside and reflect general changes in the Asian scene: the growth of commerce, the great increase in Chinese migration, and the revival of European interest in eastern Asia. In the late 18th and early 19th centuries, trade, primarily with China, seems to have undergone rapid expansion, thanks in part to the growth in numbers of Chinese immigrants acting as merchants and entrepreneurs. The principal Thai export item in the early 19th century was sugar, a new industry that began because of Chinese enterprise.

Chinese immigration assumed the proportions of a flood. By the mid-19th century, approximately half of Bangkok's estimated population of 400,000 were Chinese; these newcomers greatly increased commercial activity in Thailand and came to control much of the country's internal trade.

Revival of Western interest in eastern Asia, which at the start was primarily British interest in China, brought Western traders and diplomats to "rediscover" Thailand in the early 19th century. The Westerners were politely received, and the Thai agreed to sign limited treaties (with Britain in 1826 and the United States in 1833).

King
Mongkut's
treaties
with the
West

The acceleration of Western demands for free trade and diplomatic representation in Thailand—emphasized by the British conquests in Myanmar, expansion in Malaya, and forced opening of China—led to a change in Thailand's foreign policy under King Mongkut (Rama IV; reigned 1851–68). Mongkut signed with Britain (1855), the United States (1856), France (1856), and other Western states treaties that held Thai taxation on imports and exports to a low level and permitted the establishment of foreign consulates with extraterritorial powers. They led to greatly expanded trade with the West, to development of commercial rice farming in the Central Plain, and to expansion of foreign influence.

King Mongkut is noted particularly for his realistic evaluation of Western power and his decision to treat peaceably to gain the best terms possible. Concessions in trade were followed by concessions in territory: Thailand relinquished its rights in Cambodia to the French in 1867. Western influence on internal affairs, however, was limited to the acceptance of some innovations introduced by Western missionaries, including printing and vaccination, and the hiring of some Western technicians and teachers; the latter included Anna Leonowens, whose romanticized report of Mongkut's reign became the basis of a musical comedy, *The King and I.* Perhaps more indicative of the monarch's attitude toward Western culture was his sponsorship of a reformed sect of Thai Buddhism, the Thammayut, that aimed at purification of Buddhist practice. This reform reflected the desire to make Buddhism less vulnerable to the criticisms of Western missionaries. In fact, the Thai were (and are) little attracted to Christianity.

KING CHULALONGKORN AND INTERNAL REFORM

King Mongkut was succeeded in 1868 by his 15-year-old son Chulalongkorn (Rama V; reigned 1868–1910), who, after a five-year period of regency, assumed full powers in 1873. Chulalongkorn continued his father's policy of concessions toward the West, and territory was surrendered to the French in Laos and Cambodia and to the British in Malaya. The large concession of all Lao vassal territories east of the Mekong River to France in 1893 was made with extreme reluctance, and the Thai attempted without success to enlist British support against the French. The arrival of French gunboats in the Chao Phraya River with guns trained on the palace gave the Thai little choice. In 1896 the French and British imperial powers agreed to recognize the central part of Thailand as a free neutral buffer zone.

Internal
reform

To the policy of accommodation to the West, Chulalongkorn added the first effective program of internal reform. There were enormous political difficulties in bringing these internal changes about, for many of them undercut the bases of power of influential men at court. The young King proceeded gradually and was aided by a few gifted and far-sighted leaders of "Young Siam," most notably the King's half-brother Prince Damrong. The internal reforms accomplished during Chulalongkorn's reign included the creation of a highly centralized bureaucracy, the strengthening of government revenues, the drawing of vassals and outlying dependencies into the regular provincial administration, the abolition of slavery and labour-service requirements, the end of the formalized system of patron-client relationships, the beginning of new state services such as public education, and the introduction of technical improvements such as railway and telegraph lines. The first tentative move toward internal reform, undertaken in 1873, was aimed at reducing inefficiencies in revenue collection so that a larger share of government revenues would come directly to the King. In 1892 the government was completely reorganized into ministries with functional responsibilities. The model for these reforms was the West; their purpose was to provide Thailand with the internal strength to meet the challenge of the West. This policy was successful: Thailand retained its independence, the country grew stronger, and the monarchy retained its leadership.

Chulalongkorn's policies were continued by his sons Vajiravudh (Rama VI; reigned 1910–25) and Prajadhipok (Rama VII; reigned 1925–35). King Vajiravudh opened (1917) Thailand's first university, which was named after his father; in 1921 he made primary education compulsory. Vajiravudh strengthened the Thai army and navy and created a paramilitary organization, the Wild Tiger Corps, for civil servants. In 1917 he took Thailand into World War I on the side of the Allies, and at the end of the war the Thai won Western assent to new treaties eliminating unequal provisions. King Vajiravudh's major work, however, was the promotion of Thai nationalism; he is best known in Thailand today for a great volume of literary works, much of which stresses the need for the Thai people to be united, to be committed to their nation, religion, and king.

The most notable event during the reign of King Prajadhipok was a coup d'etat that brought an end to absolute monarchy in Thailand in 1932. Prajadhipok did not have the inclination to rule that his elder brother and father had had, and he turned many affairs of state over to his royal relatives, who, believing that government expenses were excessive, embarked on a policy of economy. This policy, enlarged during the worldwide depression of the 1930s, adversely affected many government officials, civil and military, and contributed to the disaffection of younger members of the ruling elite.

THAILAND SINCE 1932

The coup of 1932

The coup d'etat of 1932 was the work of several discontented elements of the educated ruling class who acted from a variety of motives: unhappiness with Prajadhipok's retrenchment policies, with the domination of government by members of the royal family, with disappointment of personal hopes for position and power, with a monarchic institution that they considered anachronistic, and with the lack, as they saw it, of social and economic progress in Thailand. The intellectuals of the coup, led by a young lawyer, Pridi Phanomyong, made plans for a democratic regime that would enact far-reaching social and economic reforms.

On June 24, 1932, the rebels (the People's Party), without bloodshed, seized control of the army, imprisoned the royal officials, and won the King's acceptance of a constitutional regime.

The disparate elements that composed the coup party soon broke up into factions in the new government: conservative military leaders, on the one hand, and more radical civilians, on the other. When it appeared that the newly appointed parliamentary Assembly, composed of coup members, would accept a socialist economic plan presented by the civilian leader Pridi Phanomyong, the King dissolved it. Fearing that the King would regain control of the government, the military leaders forced the reconstitution of the Assembly. This was followed by an attempted royalist countercoup. By the end of 1933, the government was clearly under the control of the military leaders, and for the most part it has remained so ever since. In 1939 a declaration was issued naming the country Thailand, not Siam.

In 1939 Field Marshal Pibul Songgram became premier and embarked on a strongly nationalistic policy that was chauvinist at home, irredentist and pro-Japanese abroad. Thailand's leaders, whose policy was always to promote Thai independence, appreciated the fact that the country was small and defenseless against a big power. In the 19th and early 20th centuries, when the major power was Britain, the Thai government had been pro-British; in 1941, with Japanese armies at the Cambodian border, the Thai government saw no alternative but to become an ally of Japan.

In 1945 civilian governments headed or supported by Pridi Phanomyong came to power, favoured by the Allied Powers. Problems created by the war, corruption, and the death by gunshot in 1946 of the young king Ananda Mahidol (Rama VIII), who had succeeded Prajadhipok after his abdication in 1935, faced the government, and late in 1947 the army, sensing that Allied interest in Thai internal affairs had declined, removed the civilian government, accusing its leaders of regicide. Pibul Songgram again became premier in 1948.

Sarit as premier

During the 1950s and '60s, military dominance remained unchanged. For nine years Pibul played the role of arbiter between competing factions, but by 1957 the army faction led by Sarit Thanarat had gained predominance of power and seized control of the government. Sarit became premier in 1958, and, to emphasize his intention of ruling without interference, he abolished the constitution and dispensed with the formality of elections. After Sarit's death in 1963, his successor as commander in chief, Thanom Kittikachorn, assumed the premiership. In 1968 a new constitution was inaugurated, and in 1969 a partly elected National Assembly was reconstituted. Both constitution and assembly were suspended in late 1971.

The autocratic military-dominated regime faced mounting problems in the early 1970s. Internally, while the economy continued to expand, social and regional imbalances grew more marked; externally, American withdrawal from Vietnam stimulated criticism of a long-standing policy of reliance upon the United States. Discontent exploded in late 1973, led by an organized student protest movement. On October 14, in a clash with military units, some 100 student marchers were killed. The discredited government lost army support. King Bhumibol Adulyadej (younger brother of King Ananda and king from 1946) dismissed the government leaders, three of whom fled the country, and launched a process that led to the inauguration of an elected civilian government.

Democratic government lasted three years. During those years censorship of communications media was lifted, trade unionism was expanded, and the government attempted some programs of economic amelioration. But no national consensus emerged. Polarization of society was evidenced in a plenitude of competing political parties, protest movements, strikes, and virulent leftist and rightist propaganda.

The return of former premier Thanom late in 1976 set off new student protests and clashes and opened the way on October 6 for the military forces to suppress the students, dismiss the civilian government, and resume power. In November 1977 Supreme Commander Kriangsak Chamanand became premier, promising peace and order, internal reforms, and regularization of relations with Communist neighbours. Parliamentary elections took place in April 1979 under a constitution providing for limited democracy. (W.F.V.)

Throughout the 1970s Thailand was a haven for refugees from the rest of war-torn Southeast Asia. As a result of the Vietnamese invasion of Kampuchea in 1979, Thailand was again deluged with refugees. In response, the Thai government, along with international relief agencies, set up huge refugee camps in the areas bordering Kampuchea.

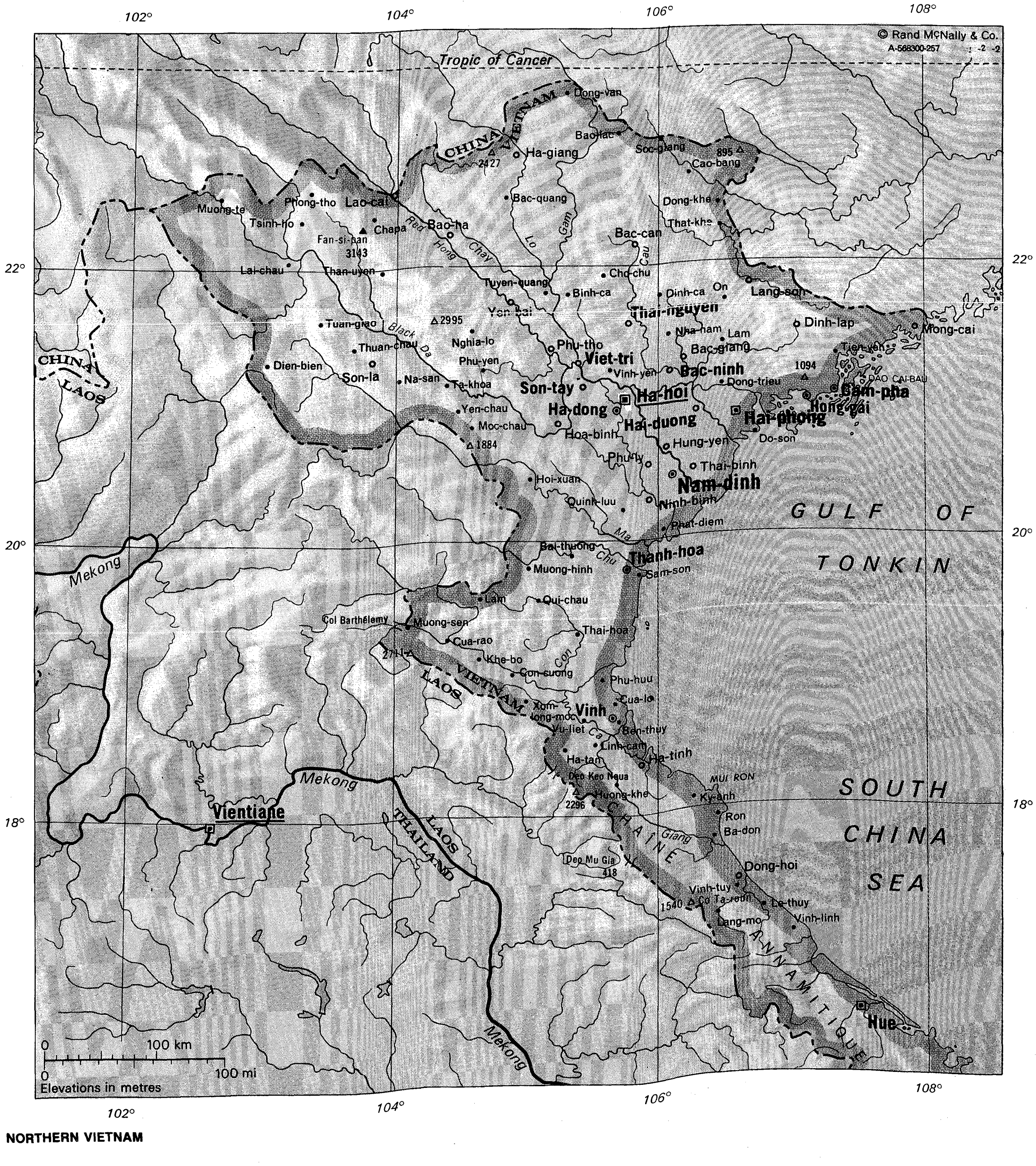
Relations with Vietnam were increasingly strained, and the situation worsened after several Vietnamese incursions into Thai territory. In 1980 the new prime minister of Thailand, Prem Tinsulanond, led the move to deny UN recognition to the Vietnamese-backed Kampuchean regime. Prem Tinsulanond, personally popular and supported by the monarchy, survived abortive coup attempts in 1981 and 1985. (Ed.)

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

VIETNAM

The Socialist Republic of Vietnam (Cong Hoa Xa Hoi Chu Nghia Viet Nam) is a densely populated nation in the southeastern part of the Indochinese Peninsula. It has an area of 127,200 square miles (329,600 square kilometres). From north to south it has a length of about 1,025 miles (1,650 kilometres), and at its narrowest part it is about 30 miles wide. Vietnam is bordered by the People's Republic of China on the north, Laos and Kampuchea on the west, and the Gulf of Thailand and the South China Sea on the south and east. The capital, Hanoi, is located in the north. Ho Chi Minh City (formerly Saigon) is in the south.

On September 2, 1945, following the Japanese surrender that ended World War II, the Democratic Republic of Vietnam was proclaimed in Hanoi. The return of the French, however, precipitated the First Indochina War, which came to a close with the Geneva Accords of 1954. This agreement provided for the temporary military partition of Vietnam at the 17th parallel and promised nationwide elections in July 1956 (the elections were never held). The two political entities that emerged were the Democratic Republic of Vietnam, better known as North Vietnam, and the Republic of Vietnam, usually referred to as South Vietnam. By the late 1950s, insurgency supported by North Vietnam began to spread in South Vietnam, and it intensified considerably following the toppling of the government and the assassination of President Ngo Dinh Diem in 1963. In March 1965, U.S. Marines landed at Da Nang, marking the beginning of direct U.S. military involvement in the conflict; at about the same time, North Vietnamese troops began infiltrating the south. A cease-fire agreement signed in Paris in January 1973 technically ended the war and provided for the withdrawal of the last U.S. forces. The cease-fire was not honoured, however, and an offensive by North Vietnamese armed forces in South Vietnam early in 1975 brought about the collapse of the Saigon government on April 30. The Provisional Revolutionary Government of South Vietnam, which was created by the Communists in 1969, emerged in May 1975 as the official ruling body of the south for a year. Negotiations between delegations from the north and south



NORTHERN VIETNAM

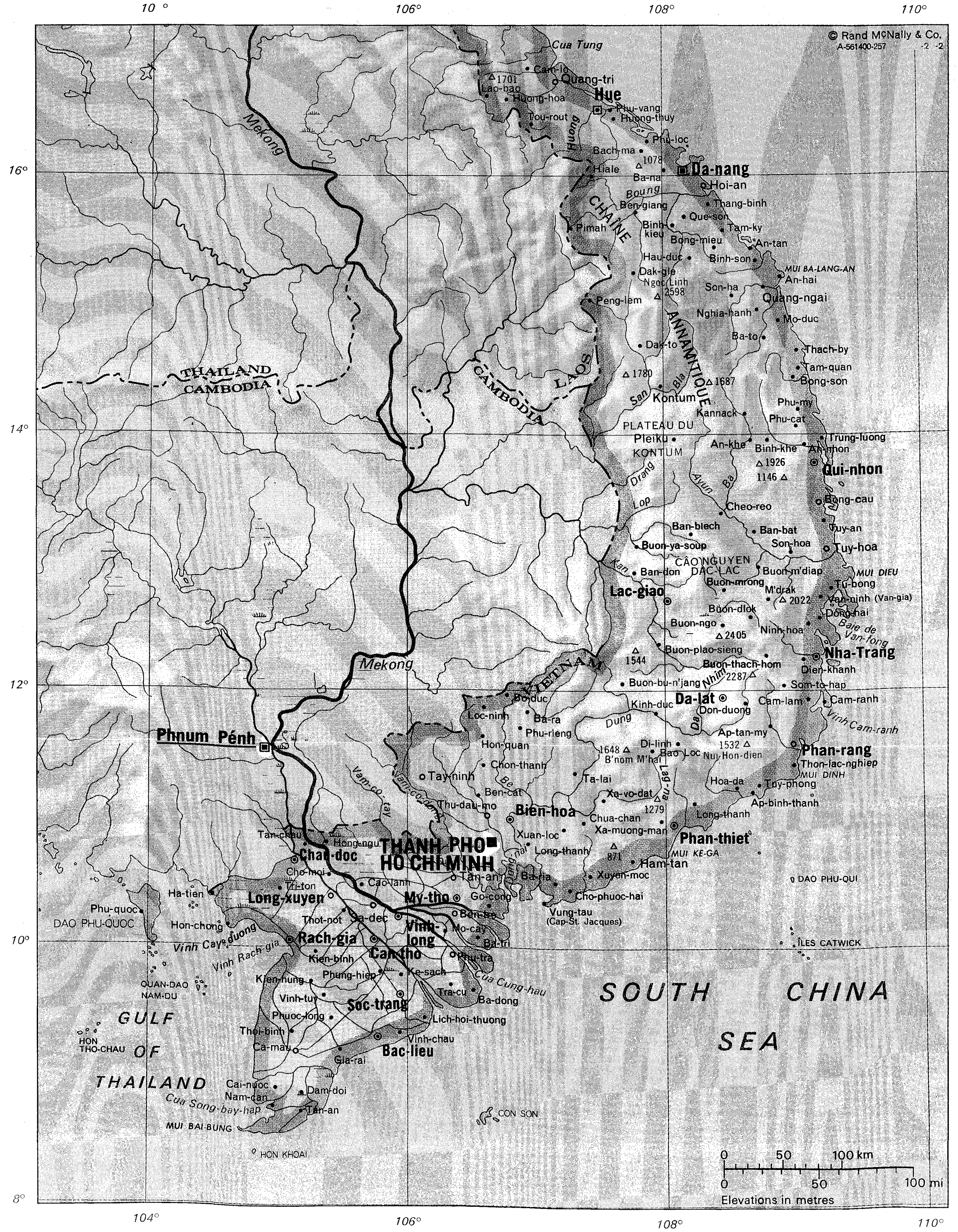
resulted in elections on April 25, 1976, of a new National Assembly for the whole country. This National Assembly convened for the first time in Hanoi on June 24; on July 2 it proclaimed the reunification of the country.

Physical and human geography

THE LAND

Relief. Vietnam encompasses 127,200 square miles (329,600 square kilometres). Its principal physiographic features are the Chaîne Annamitique (Annamite Chain or Cordillera; Truong Son), extending from north to south in central Vietnam and dominating the interior, and two extensive alluvial deltas formed by the Red River (Song Hong) in the north and the Mekong River (Song Cuu Long) in the south. Between these two deltas is a long, relatively narrow coastal plain.

From north to south, the uplands of northern Vietnam can be divided into two distinct regions—the region north of the Red River and the massif that extends south of the Red River into neighbouring Laos. The Red River forms a deep, relatively wide valley that runs in a straight northwest–southeast direction for much of its 130 miles (210 kilometres) from the border with the People's Republic of China to the edge of the delta. North of the Red River, the relief is moderate with two essential regions. One region has heights characterized by a northwest–southeast trend such as are found between the Red River and the Song Lo (Clear River); appended to this is a marked depression from Cao Bang to the sea. The second region is charac-



SOUTHERN VIETNAM

terized by mountains in its eastern portion that form arcs convex to the east and southeast. This region also contains a sub-area of wide terraces, extensive alluvial plains, and low hills; it extends from the Red River delta along the Lo, Chay, Chu, Thuong, Luc Nam, and Ky Cung rivers.

Compared with the area north of the Red River, the vast massif extending southwest across Laos to the Mekong River is of considerably higher elevation and is overlain with poor soils. Among its outstanding topographical features is the Fan Si Pan chain, the highest peaks of which are Mount Fan Si Pan (10,312 feet [3,143 metres]) and Pou Luong (9,796 feet [2,986 metres]). It is a crystalline formation with few plains. South of the Black River are the Ta P'ing, Son La, and Moc Chau plateaus, which are separated by deep valleys.

In central Vietnam, the Chaîne Annamitique runs parallel to the coast, with several peaks rising more than 6,000 feet (1,830 metres). Several spurs jut into the South China Sea, forming compartments isolated from one another. One such spur dissects the coastal plain at Mui Ron (previously called the Gate of Annam) and another rises at the Hai Van pass (the Col des Nuages), just north of Da Nang. Communication across the main chain in central Vietnam is difficult. Located at Lao Bao inland from Quang Tri

MAP INDEX

Cities and towns

An-hai	15·13n	108·55e
An-khe	13·57n	108·39e
An-nhon	13·53n	109·06e
An-tan	15·25n	108·41e
Ap-binh-thanh	11·11n	108·42e
Ap-tan-my	11·43n	108·48e
Bac-can	22·08n	105·49e
Bac-giang	21·16n	106·11e
Bach-ma	16·12n	107·52e
Bac-lieu	9·17n	105·43e
Bac-ninh	21·11n	106·03e
Bac-quang	22·29n	104·52e
Ba-don	17·45n	106·27e
Ba-dong	9·40n	106·33e
Bai-thuong	19·54n	105·23e
Ba-na	16·00n	107·59e
Ban-bat	13·13n	108·39e
Ban-blech	13·12n	108·13e
Ban-don	12·53n	107·48e
Ban-me-thuot, see Lac-giao		
Bao-ha	22·11n	104·21e
Bao-lac	22·57n	105·40e
Bao Loc	11·30n	107·54e
Ba-ra	11·50n	106·58e
Ba-ria	10·30n	107·10e
Ba-to	14·46n	108·46e
Ba-tri	10·02n	106·35e
Ben-cat	11·09n	106·36e
Ben-giang	15·41n	107·47e
Ben-thuy	18·39n	105·42e
Ben-tre	10·14n	106·22e
Bien-hoa	10·59n	106·49e
Binh-ca	21·48n	105·18e
Binh-khe	13·57n	108·49e
Binh-kieu	15·35n	108·06e
Binh-son	15·18n	108·45e
Bo-duc	11·58n	106·48e
Bong-cau	13·27n	109·13e
Bong-mieu	15·25n	108·24e
Bong-son	14·26n	109·01e
Buon-bu-n'jang	12·05n	107·40e
Buon-dlok	12·34n	108·42e
Buon-m'diap	12·56n	108·43e
Buon-mrong	12·48n	108·28e
Buon-ngo	12·30n	108·28e
Buon-plao-sieng	12·21n	107·56e
Buon-thach-hom	12·17n	108·49e
Buon-ya-soup	13·05n	107·52e
Cai-nuoc	8·57n	105·03e
Ca-mau	9·11n	105·08e
Cam-lam	11·54n	109·08e
Cam-lo	16·46n	107·00e
Cam-pha	21·01n	107·19e
Cam-ranh	11·54n	109·12e
Can-tho	10·02n	105·47e
Cao-bang	22·40n	106·15e
Cao-lanh	10·27n	105·38e
Cap-Saint-Jacques, see Vung-tau		
Chapa	22·21n	103·50e
Chau-doc	10·42n	105·06e
Cheo-reo	13·24n	108·27e
Cho-Chu	21·54n	105·39e
Cho-moi	10·51n	106·38e
Chon-thanh	11·24n	106·36e
Cho-phuoc-hai	10·26n	107·18e
Chua-chan	10·57n	107·25e
Con-cuong	19·02n	104·54e
Cua-lo	18·49n	105·43e
Cua-rao	19·16n	104·27e
Dak-gle	15·12n	107·48e
Dak-to	14·42n	107·51e
Da-lat	11·56n	108·25e
Dam-doi	8·50n	105·15e
Da-nang	16·04n	108·13e
Dien-bien	21·23n	103·01e
Dien-khanh	12·15n	109·05e
Di-linh	11·35n	108·04e
Dinh-ca	21·45n	106·03e
Dinh-lap	21·33n	107·06e
Don-duong	11·51n	108·35e
Dong-hai	12·34n	109·13e
Dong-hoi	17·29n	106·36e
Dong-khe	22·26n	106·27e
Dong-trieu	21·05n	106·31e
Dong-van	23·16n	105·22e
Do-son	20·42n	106·47e
Gia-rai	9·14n	105·27e
Go-cong	10·22n	106·40e
Ha-dong	20·58n	105·46e
Ha-giang	22·50n	104·59e
Hai-duong	20·56n	106·19e
Hai-phong	20·52n	106·41e
Ham-tan	10·40n	107·45e
Ha-noi	21·02n	105·51e
Ha-tien	10·23n	104·29e
Ha-tan	18·30n	105·20e
Ha-tinh	18·20n	105·54e
Hau-duc	15·20n	108·13e
Hiale	15·55n	107·33e
Hoa-binh	20·50n	105·20e
Hoa-da	11·11n	108·33e
Ho Chi Minh, Thanh Pho	10·45n	106·40e
Hoi-an	15·52n	108·19e
Hoi-xuan	20·22n	105·07e
Hon-chong	10·09n	104·36e
Hong-gai	20·57n	107·05e
Hong-ngu	10·48n	105·20e
Hon-quan	11·39n	106·36e
Hue	16·28n	107·36e
Hung-yen	20·39n	106·04e
Huong-hoa	16·35n	106·48e
Huong-khe	18·13n	105·41e
Huong-thuy	16·26n	107·41e
Kannack	14·07n	108·36e
Ke-sach	9·46n	105·58e
Khe-bo	19·08n	104·41e
Kien-binh	9·55n	105·18e
Kien-hung	9·44n	105·16e
Kinh-duc	11·49n	107·57e
Kontum	14·21n	108·00e
Ky-anh	18·05n	106·18e
Lac-giao (Ban-me-thuot)	12·40n	108·03e
Lai-chau	22·02n	103·10e
Lam	21·21n	106·31e
Lang-mo	17·14n	106·27e
Lang-son	21·50n	106·44e
Lao-bao	16·38n	106·46e
Lao-cai	22·30n	103·57e
Le-thuy	17·14n	106·49e
Lich-hoi-thuong	9·26n	106·09e
Linh-cam	18·31n	105·34e
Loc-ninh	11·50n	106·35e
Long-thanh	10·47n	106·57e
Long-xuyen	10·23n	105·25e
Mo-cay	10·08n	106·20e
Moc-chau	20·50n	104·38e
Mo-duc	14·53n	108·54e
Mong-cai	21·32n	107·58e
Muong-hinh	19·49n	105·03e
Muong-sen	19·23n	104·08e
Muong-te	22·28n	102·37e
My-tho	10·21n	106·21e
Nam-can	8·46n	104·59e
Nam-dinh	20·25n	106·10e
Na-san	21·12n	104·02e
Nghia-hanh	14·58n	108·46e
Nghia-lo	21·36n	104·31e
Nha-nam	21·27n	106·06e
Nha-Trang	12·15n	109·11e
Ninh-binh	20·15n	105·59e
Ninh-hoa	12·30n	109·07e
On	21·40n	106·35e
Peng-lem	15·01n	107·26e
Phan-rang	11·34n	108·59e
Phan-thiet	10·56n	108·06e
Phat-diem	20·06n	106·06e
Phong-tho	22·32n	103·21e
Phu-cat	14·01n	109·03e
Phu-huu	18·58n	105·31e
Phu-loc	16·16n	107·53e
Phu-ly	20·32n	105·56e
Phu-my	14·11n	109·02e
Phung-hiep	9·49n	105·49e
Phuoc-long	9·26n	105·26e
Phu-quoc	10·12n	103·57e
Phu-rieng	11·41n	106·53e
Phu-tho	21·24n	105·13e
Phu-tra	9·56n	106·20e
Phu-vang	16·31n	107·37e
Pimah	15·35n	107·25e
Pleiku	13·58n	108·01e
Quang-ngai	15·07n	108·48e
Quang-tri	16·45n	107·13e
Que-son	15·40n	108·13e
Qui-chau	19·33n	105·06e
Quinh-luu	20·13n	105·50e
Qui-nhon	13·46n	109·14e
Rach-gia	10·00n	105·04e
Ron	17·53n	106·27e
Sa-dec	10·18n	105·45e
Saigon, see Ho Chi Minh, Thanh Pho		
Sam-son	19·44n	105·54e
Soc-giang	22·54n	106·01e
Soc-trang	9·36n	105·58e
Som-to-hap	12·10n	108·43e
Son-ha	15·03n	108·33e
Son-hoa	13·02n	108·58e
Son-la	21·19n	103·54e
Son-tay	21·08n	105·30e
Ta-khoa	21·13n	104·18e
Ta-lai	11·24n	107·23e
Tam-ky	15·34n	108·29e
Tam-quan	14·33n	109·02e
Tan-an	8·46n	105·11e
Tan-an	10·32n	106·25e
Tan-chau	10·48n	105·15e
Tay-ninh	11·18n	106·05e
Thach-by	14·40n	109·03e
Thai-binh	20·35n	106·01e
Thai-hoa	19·19n	105·25e
Thai-nguyen	21·36n	105·50e
Thang-binh	15·43n	108·21e
Thanh-hoa	19·48n	105·46e
Than-uyen	22·00n	103·54e
That-khe	22·16n	106·28e
Thoi-binh	9·21n	105·05e
Thon-lac-nghiep	11·20n	108·53e
Thot-not	10·16n	105·31e
Thuan-chau	21·26n	103·41e
Thu-dau-mo	10·58n	106·39e
Tien-yen	21·20n	107·24e
Tou-rout	16·24n	107·01e
Tra-cu	9·42n	106·15e
Tri-ton	10·25n	105·00e
Trung-luong	13·57n	109·14e
Tsinh-ho	22·22n	103·14e
Tuan-giao	21·35n	103·25e
Tu-bong	12·47n	109·18e
Tuy-an	13·21n	109·13e
Tuyen-quang	21·49n	105·13e
Tuy-hoa	13·05n	109·17e
Tuy-phong	11·13n	108·44e
Van-ninh (Van-gia)	12·42n	109·13e
Viet-tri	21·18n	105·26e
Vinh	18·40n	105·40e
Vinh-chau	9·20n	105·58e
Vinh-linh	17·04n	107·02e
Vinh-long	10·15n	105·58e
Vinh-tuy	9·37n	105·21e
Vinh-tuy	17·24n	106·36e
Vinh-yen	21·18n	105·35e
Vu-liet	18·43n	105·23e
Vung-tau (Cap-Saint-Jacques)	10·21n	107·04e
Xa-muong-man	10·58n	108·00e
Xa-vo-dat	11·09n	107·29e
Xom-long-moc	18·51n	105·01e
Xuan-loc	10·56n	107·14e
Xuyen-moc	10·34n	107·25e
Yen-bai	21·42n	104·52e
Yen-chau	21·03n	104·18e

Physical features and points of interest

Annamitique, Chaîne, *mountains*	17·20n	106·30e
Ayun, *river*	13·24n	108·28e
Ba, *river*	13·05n	109·18e
Bai-bung, Mui, *point*	8·36n	104·43e
Ba-lang-an, Mui, *cape*	15·14n	108·56e
Barthélemy, Col, *pass*	19·26n	104·06e
Be, *river*	11·06n	106·57e
Bla, *river*	14·20n	107·55e
Black, *river*	21·15n	105·20e
Boung, *river*	15·51n	107·51e
Ca, *river*	18·46n	105·47e
Cai-Bau, Dao, *island*	21·10n	107·27e
Cam-ranh, Vinh, *bay*	11·53n	109·10e
Catwick, Îles, *islands*	10·00n	109·00e
Cau, *river*	21·07n	106·18e
Cay-duong, Vinh, *bay*	10·05n	104·45e
Chay, *river*	21·39n	105·12e
Chu, *river*	19·53n	105·45e
Clear, see Lo		
Con, *river*	19·04n	105·00e
Con Son, *islands*	8·43n	106·36e
Cung-hau, Cua, *river mouth*	9·50n	106·30e
Dac-lac, Cao Nguyen, *plateau*	12·50n	108·05e
Dieu, Mui, *cape*	10·42n	107·59e
Dieu, Mui, *point*	12·53n	109·28e
Dinh, Mui, *cape*	11·22n	109·01e
Dong-nai *river*	10·45n	106·48e
Drang, *river*	13·35n	107·32e
Dung, *river*	11·12n	107·15e
Fan-si-pan, *mountains*	22·15n	103·46e
Gam, *river*	21·55n	105·12e
Giang, *river*	17·40n	106·30e
Hon-dien, Nui, *mountain*	11·35n	108·40e
Huong, *river*	16·33n	107·38e
Khoai, Hon, *island*	8·26n	104·48e
Kan, *river*	13·02n	107·30e
Keo Neua, Deo, *pass*	18·23n	105·09e
Kontum, Plateau du, *plateau*	13·55n	108·05e
La-gna, *river*	11·12n	107·16e
Linh, Ngoc, *mountain*	15·05n	108·00e
Lo, *river*	21·18n	105·25e
Lop, *river*	13·17n	107·41e
Ma, *river*	19·47n	105·56e
Mekong, *river*	10·33n	105·24e
M'hai, B'nom, *mountain*	11·33n	107·45e
Mu Gia, Deo, *pass*	17·40n	105·47e
Nam-du, Quan-dao, *islands*	9·42n	104·22e
Nhim, Da, *river*	11·42n	108·15e
Phu-qui, Dao, *islands*	10·33n	108·57e
Phu-quoc, Dao, *island*	10·13n	104·00e
Rach-gia, Vinh, *bay*	10·00n	105·00e
Red, *river*	20·17n	106·34e
Ron, Mui, *point*	18·07n	106·22e
San, *river*	13·50n	107·25e
Song-bay-hap, Cua, *bay*	8·46n	104·52e
South China Sea	9·00n	108·20e
Ta-roun, Co, *mountain*	17·23n	106·17e
Thailand, Gulf of	9·00n	104·00e
Tho-chau, Hon, *island*	9·20n	103·28e
Tonkin, Gulf of	20·00n	108·00e
Tung, Cua, *river mouth*	17·01n	107·07e
Vam-co-dong, *river*	10·30n	106·33e
Vam-co-tay, *river*	10·30n	106·33e
Van-fong, Baie de, *bay*	12·35n	109·20e

is the chief pass, the Ai Lao, which provided an ancient invasion route between the coastal plain and the interior. The southern portion of the Chaîne Annamitique has two identifiable regions. One consists of mountain plateaus of approximately 1,700 feet (518 metres) in elevation that have experienced little erosion, as in the vicinity of Lac Giao (former Ban Me Thuot). The second region is characterized by heavily eroded plateaus; in the vicinity of Pleiku, the plateau is about 2,500 feet (760 metres) above sea level and in the Da Lat area, about 4,900 feet (1,500 metres) in elevation.

The Red River delta

Drainage. Below the northern uplands is the Red River delta. Roughly triangular in shape, the 4,250-square mile (11,000-square kilometre) delta extends 150 miles (240 kilometres) inland and measures 75 miles (120 kilometres) along the Gulf of Tonkin. The Red River delta can be divided into four sub-regions. The northwest section has the highest and most broken terrain, and its extensive natural levees invite settlement despite frequent flooding. The low-lying eastern portion of the delta has bench marks of less than seven feet (two metres) elevation in the vicinity of Bac Ninh. Rivers running through this area form small valleys only slightly lower than the general surface level, so they are subject to flooding by unusually high tides. The third and fourth sub-regions are the poorly drained lowlands to the west and the coastal area, which is marked by the remains of former beach ridges left by continuous deltaic aggradation.

The Mekong River delta

South of the mountain range there is an identifiable terrace region that gives way to the Mekong River delta. The terrace region includes the alluvial plains along the Saigon and Dong Nai rivers. The lowlands of southern Vietnam are dominated by alluvial plains, the most extensive of which is the Mekong River delta, covering an area of 15,600 square miles (40,400 square kilometres). Smaller deltaic plains also occur along the south-central coast of the South China Sea.

Soils. In northern Vietnam, the heavy seasonal rains wash away rich humus from the highlands, leaving slow-dissolving alumina and iron oxides that give the soil its characteristic reddish colour. The soils of the Red River delta vary; some are fertile and suitable to intense cultivation, while others lack soluble bases. Nonetheless, the delta soils are in good condition, without any lateritic tendencies, and are easily worked. The diking of the Red River to prevent flooding deprives the delta's paddy fields of enriching silts, necessitating the use of chemical fertilizers.

There are at least 25 soil associations, but certain soil types predominate. Among these are red and yellow podzolic soils (soils that are heavily leached in their upper layers, with a resulting accumulation of materials in the lower layers), which occupy nearly half of the land area, and latosols (reddish brown leached tropical soils), which comprise about one-tenth. These soil types dominate the central highlands. Alluvial soils account for about one-fourth of the land in the south and are concentrated in the Mekong River delta, as are peat and muck soils. Gray podzolic soils are found in parts of the central highlands and in old terraces along the Mekong River, while regurs (rich black loams) and latosols occur in both the central highlands and the terrace zone. Along the coast of central Vietnam are regosols (soft, undeveloped soils) and non-calcic brown soils.

Climate. The northern part of Vietnam is on the edge of the tropical world. During January, the coldest month of the year, Hanoi has a mean temperature of 62° F (17° C). The annual average temperature is 74° F (23° C). Farther south, the average annual temperature in Hue is 77° F (25° C) and in Saigon is 81° F (27° C); in the highland city of Da Lat (4,920 feet [1,500 metres] in elevation), it is 70° F (21° C). The winter season in northern Vietnam lasts from November to April; from early February to the end of March there is a persistent drizzling rain that the Vietnamese call "dust rain" or "flying rain." The summer in northern Vietnam lasts from April or May to October and is characterized by heat, heavy rain, and occasional typhoons. In central and southern Vietnam, the southwest monsoon winds between June and November bring rains and occasional typhoons to the eastern slopes of the mountains and the lowland plains. Between December and April there is a drier period that is characterized by northeasterly winds and, in the south, by high temperatures.

Plant and animal life. The vegetation of Vietnam is rich and diversified because of the great range of climate, topography, and soils and the varying effects of human habitation. The forests of Vietnam can be divided into two broad categories: evergreen forests, which include the conifers, and deciduous forests. There are more than 1,500 species of woody plants in Vietnam, varying in size from small shrubs to large trees and ranging from hardwoods such as ebony and teak to palms, mangroves, and bamboos. There are also numerous species of woody vines, or lianas, and herbaceous plants. In the aggregate, the dense and open forests, savannas, brushland, and bamboo cover approximately half of the total area of Vietnam.

In most areas the forests are mixed, containing a large number of species within an area of a square mile. Rain forests are relatively limited and pure stands are few. The nearest to pure forest types are the pines—the three-needled *Pinus khasya* and two-needled *P. merkusii* found in the uplands—and the mangrove forests of the coastal swamps. In the mountainous regions there are subtropical species such as varieties of *Quercus, Castanopsis, Pinus,* and *Podocarpus.* Brushwood, bamboo, weeds, and tall grasses invade cutover forests and grow around settlements and along arterial highways and railroads. Between the cutover areas and the upland forests are other mixtures of forest types.

A large part of the forest is of the closed type and rich in broad-leaved evergreens and semi-evergreens, some of which produce valuable timbers. About one-fifth of this forest type is composed of undisturbed (primary) forests. Other types of forests in this region include secondary forests; open forests, which usually have trees of the Dipterocarpaceae family and *Lagerstroemia* (crape myrtle) species; mangrove forests; and barren lands of sand dunes with eucalyptus and small, thorny deciduous trees and *Casuarina* species of flowering plants. Cogon grass (*Imperata cylindrica*) is commonly found in the open forests, and grass savannas occupy large areas formerly covered by forests. Grass and sedge swamps are characteristic of the Dong Thap Muoi (Plain of Reeds), a depression in the Mekong River delta.

Effects of defoliation

During the Vietnam War, herbicides were used by the U.S. Army to defoliate forests in South Vietnam. About one-fifth of the total acreage of South Vietnam was sprayed with defoliant chemicals, which affected all types of vegetation, especially coastal mangrove forests and inland forests.

Permanent cultivation covers large areas of the Vietnam lowlands and smaller portions of the highlands. Much of this acreage is in wet rice, the staple of the Vietnamese diet. Plantations of banana, other fruit trees, coconut, and sugarcane are found in the Mekong River delta, while tea and coffee are grown in the central highlands. Extensive rubber estates occur in the central highlands and the terrace region. Fields, groves, and kitchen gardens throughout Vietnam contain a wide variety of fruit trees (banana, orange, mango, jackfruit, and coconut) and vegetables. Kapok trees are found in many villages, and the Vietnamese cultivate areca palms and betel peppers for their nuts and leaves.

The most common domesticated animals in Vietnam are water buffalo, cattle, dogs, cats, pigs, goats, ducks, and chickens. Wild game in the central highlands includes elephants and tapirs; of the varieties of rhinoceroses (*Rhinoceros sondaicus* and *Didermocerus sumatrensis*), none has been seen since the early 1940s. Also found in the forests are big cats, including tigers, leopards, ounces (*Leo uncia*), and clouded leopards (*Leo nebulosa*); several kinds of wild oxen, including gaurs and koupreys; and various types of bears, among them the black bear and honey bear. Deer are plentiful; they include the small musk deer and barking deer. Other common wild animals are wild boars, porcupines, jackals, otters, mongooses, hares, skunks, and squirrels, including flying squirrels. In the highland villages, traps are placed near granaries to catch rats, which are roasted and eaten.

There are many different kinds of small wild cats and three types of civets—Malagasy civets, binturongs, and palm civets. Primates such as the langur, macaque, gibbon, and rhesus monkey live in the forests. Crocodiles are found on the edges of some lakes and along river banks; other reptiles include several kinds of lizards, pythons, and cobras. Of the wide variety of land and water birds, 586 species have been identified in southern Vietnam alone.

Settlement patterns. Diverse cultural traditions, geographical variations, and the events of history have created distinct traditional regions. The general topographical dichotomy of highland and lowland regions also has ethnolinguistic significance; the lowlands generally have been occupied by ethnic Vietnamese, while the highlands have been the home of numerous smaller ethnic groups that differ culturally and linguistically from the Vietnamese. The highland peoples can be divided into the northern ethnic groups, with affinities to peoples in southern China, and the southern highland populations, with ties to the Mon-Khmer and Austronesian peoples of Cambodia, Indonesia, and elsewhere in Southeast Asia. A north–south variation also evolved among the ethnic Vietnamese as they expanded from the Red River delta along the coastal plain into the Mekong River delta. After the mid-19th century, Vietnam was divided by the French into Tonkin in the north, Annam in the centre, and Cochinchina in the south. The Vietnamese themselves have long made a distinction among the northern region (*bac ky*), with Hanoi as its cultural centre; the central region (*trung ky*), with the traditional royal capital of Hue; and the southern region (*nam ky*), with Saigon (now Ho Chi Minh City) as its urban centre.

ighland
nd
owland
ichotomy

Topography has strongly influenced human settlement in Vietnam. In the Red River delta a number of distinct settlement patterns can be identified, the most common of which is the "relief" village associated with topographical features such as levees, hillsides, or beach ridges. Levee settlements, usually located along river banks, often abut each other and form one elongated settlement, while clusters of farmsteads at the base of a hill may cover the hill if it is low. Most villages are enclosed by a bamboo hedge or earthen wall.

Lowland Vietnamese villages on the coastal plain in central Vietnam are characteristically close-knit, small clusters of farmsteads near a watercourse. Fishing villages normally are situated in a sheltered inlet. Vietnamese villages in the Mekong River delta are less varied than those in the Red River delta. Settlements are strung out along the rivers, canals, and streams and major roads. On the deltaic plain, most settlements are loose-knit clusters of farmsteads, although some farmsteads are scattered about the paddy fields.

In some Vietnamese settlements, particularly those in northern and central Vietnam, geomantic principles determine the internal orientation of houses and village buildings such as the *dinh,* the communal meeting hall and sanctuary for the cult of the guardian spirit, and the village Buddhist pagoda. In central Vietnam, houses and the *dinh* traditionally face the sea.

The physical pattern of settlements of the Cham and Khmer minorities closely resembles that of the Vietnamese settlements. Among the highland ethnic groups, settlement patterns vary considerably. Permanent and shifting (slash-and-burn) fields are within walking distance of the village, kitchen gardens are invariably found close to the farmstead, and good water sources are nearby. Internal arrangement of the settlement may not have any particular orientation or it may be dictated by practical considerations or ideational principles. Most groups in the northern highlands have no particular patterns for their settlements. In the central highlands, however, the Sedang prefer not to have the morning sun strike the main entrance of a house, and the Mnong Chil build their dwellings facing away from the prevailing winds. In the same region, many Jarai longhouses have a north–south orientation because of wind and rain patterns; the main entrance always faces south so that the sun may dry the rice, cotton, and other things they place on the veranda. The Katu orient their settlements around a stake to which sacrificial animals are tied and build their communal men's houses (where rituals and gatherings of the leaders take place) facing away from the rising sun.

For the most part, highland people prefer to build their houses on pilings. Exceptions occur among the Hmong (Miao, Meo), Mien (Yao), Stieng, and some Mnong groups. Floor plans for houses vary from group to group depending on whether or not a house will be occupied by an extended family and whether there are cultural prescriptions for certain parts of the house. In the central highlands, for example, the Rhadé trace descent through the female line and husbands live with their wives' families after marriage; their longhouses, therefore, are arranged with compartments for the women of the family and their husbands and children.

Vietnam's traditional major cities are Hanoi, Hue, and Saigon (Ho Chi Minh City). Archaeological research in the vicinity of Hanoi has revealed traces of the ancient citadel of Co Loa, which dates to the 3rd century BC. Throughout Vietnamese history the Hanoi area has been a place of importance and the site of several early capitals. Hanoi served as the French colonial capital from 1902 until 1954, and the city retains the charming architecture that is part of that heritage. Hanoi also has long been a centre of trade and industry for northern Vietnam; the city's outport of Haiphong was developed by the French in the late 19th century as a trade and banking centre.

The city
of Hue

The imperial past is rooted in Hue. The Nguyen family controlled central and southern Vietnam from Hue after 1687, and Emperor Gia Long established his capital there in 1802. Located on the Song Huong (Perfume River), Hue was intended to be the physical and administrative centre of the kingdom and was laid out in the early 19th century according to geomantic principles. The heart of the city is the imperial citadel of Dai Noi ("Great Within"), modelled on the Manchu palace in Peking. Hue was the intellectual and religious heart of Vietnam, and its economic functions were ancillary. Many of the traditions of the mandarins (senior public officials) have been preserved in Hue, and until vast destruction occurred during the Tet Offensive of 1968, traditional Vietnamese architecture was much in evidence.

Saigon, renamed Ho Chi Minh City in 1976, was built largely by the French on a plan devised by Théodore Lebrun, an adviser to Gia Long. As the administrative capital and principal port of the French province of Cochinchina, its riverfront was dominated by the port and commercial establishments. A French-provincial style of life prevailed in the city and its architecture recalled towns and cities in southern France. With the establishment of Hanoi as the capital of the Socialist Republic of Vietnam in 1976, Saigon lost its administrative functions as the capital of South Vietnam. The city of Cholon, adjoining Saigon, was a major centre for the ethnic Chinese until 1978, when large private enterprises were abolished and vast numbers of Chinese left the country.

THE PEOPLE

Ethnic distribution. Vietnam has one of the most complex ethnolinguistic patterns in Asia. The great tradition of China is represented by the ethnic Vietnamese, who were Sinicized during the 1,000 years of Chinese rule. Since the 10th century the Vietnamese have carried this tradition from the Red River delta southward along the coastal plain to the Mekong River delta. Chinese influence permeates all levels of the Vietnamese language, one of the Mon-Khmer languages of the Austro-Asiatic language family.

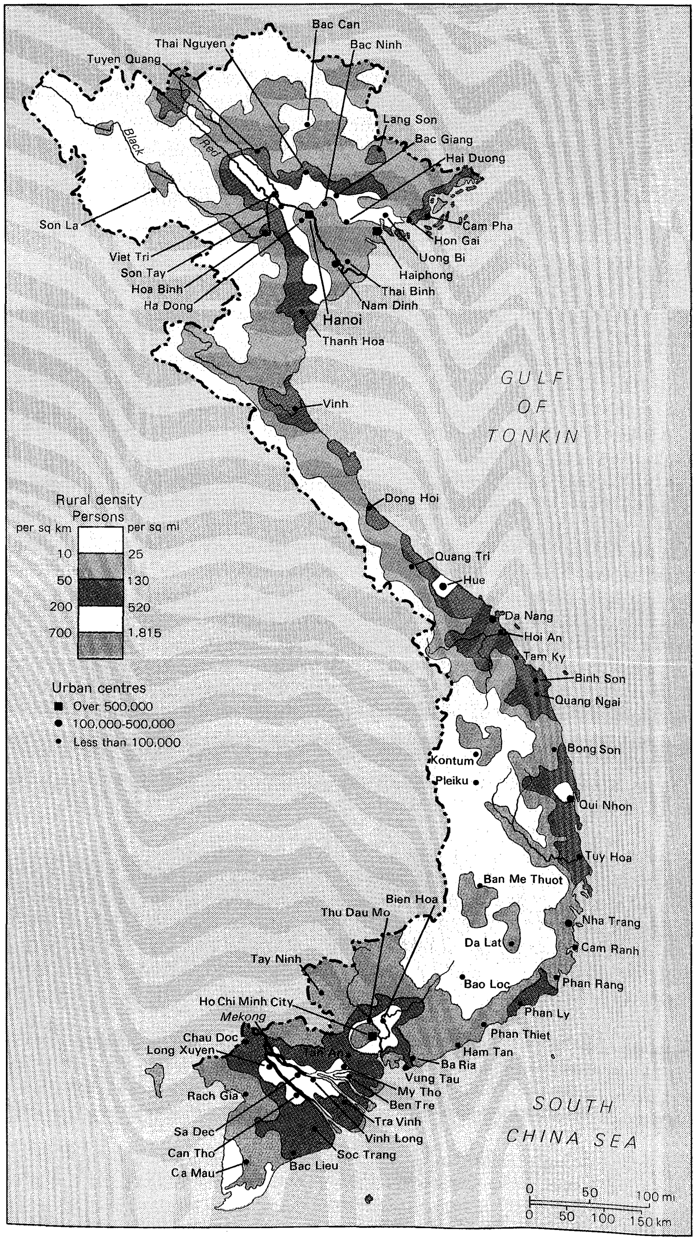
Cham and
Khmer
minorities

Indian influence is found among the Cham and Khmer minorities. The Cham, whose language belongs to the Austronesian language family, are remnants of a once large population that formed the majority in the Indianized kingdom of Champa, which existed in what is now central Vietnam from the 2nd century AD until its absorption by the Vietnamese between the 15th and 17th centuries. Cham populations remain in the south central coastal plain, where they cling to many customs of the Brahmins (the highest Hindu caste), and in a small pocket of the Mekong River delta close to the Cambodian border, where they have adopted Islām. The Khmer, whose language is one of the Mon-Khmer languages, are scattered throughout the Mekong River delta. They are descendants of a population that was once part of the Khmer Empire, the most spectacular traces of which are to be seen in the ruins of Angkor in neighbouring Cambodia.

The Montagnards

Other minorities include the ethnic groups that inhabit the highlands. In the central highlands are people less advanced than the lowlanders and, while their cultures vary considerably, they still share many characteristics derived from a way of life oriented around kin groups and small communities. Known collectively by the French as the Montagnards ("Highlanders"), they have affinities with other Southeast Asians. Those who speak languages of Austronesian stock include the Rhadé, Jarai, Chru, and Roglai; they have linguistic ties with the Cham, Malay, and Indonesians. Other Montagnards—such as the Bru, Pacoh, Katu, Cua, Hre, Rengao, Sedang, Bahnar, Mnong Maa, and Stieng—speak languages of Mon-Khmer stock, affiliating them with the Cambodians (Khmer). Unlike the lowlanders, the Montagnards historically remained aloof from Chinese and Indian influences, but between the end of the 19th century and 1954 they were exposed to French influence. French missionaries and administrators provided roman script for some of the Montagnard languages and additional orthographies have been devised since. A desire to preserve their own cultural identities has given rise to intense ethnonationalism among the Montagnards. This is the principal ideology that has sustained the United Struggle Front for the Oppressed Races (known by its French acronym, FULRO), which has been active since the mid-1960s.

By courtesy of the Central Intelligence Agency



Population density of Vietnam.

The tribal Tais

The various groups in northern Vietnam have ethnolinguistic affiliations with peoples in southern China. They include the tribal Tai (Thai) groups—the Tai Dam (Black Tai), Tai Khao (White Tai), Nung, and Tay (Tho)—who are generally found in the upland valleys. Representing the largest of the highland ethnolinguistic groups, they speak Tai languages. At higher elevations are scattered the Hmong and Mien, whose languages are of the Sino-Tibetan language family. With the establishment of the Socialist Republic of Vietnam in 1976, a program was begun to move ethnic Vietnamese from the lowlands into the mountain region.

Religion. Beliefs and values of Taoism and Confucianism shape the Vietnamese view of the universal order, the cult of the ancestors is ubiquitous, and the Mahāyāna (Greater Vehicle) tradition of Buddhism is widespread. Animist beliefs are held by many tribal peoples. During the 1920s, the syncretic religion of Cao Dai appeared, and in 1939 the Hoa Hao sect, which claims to be a "pure Buddhism," spread through parts of the Mekong River delta.

Roman Catholicism and Protestantism

Roman Catholicism was first introduced among the Vietnamese in the 17th century, and it spread widely following the French conquest in the mid-19th century. The heaviest concentrations of Roman Catholics in Vietnam were initially in the north; the majority of them, however, fled to the south after the partition of the country in 1954.

In 1959 all foreign clergy were expelled from North Vietnam, leaving only native priests. Though the 1960 constitution guaranteed freedom of religion, the North Vietnamese government tried to supplant organized religion with its own "scientific materialism" and to gain control over the existing faiths by sponsoring patriotic religious organizations. These included the Unified Buddhist Church; the Patriotic Catholic Church, whose clergy and members have renounced allegiance to the pope; the Viet Nam Cao Dai Union; and the Viet Nam Protestant Church.

Protestantism came to Vietnam with a mission in the Mekong River delta in 1911, and it spread primarily among small segments of the urban population in the central and southern regions. Congregations grew in Saigon, in the Mekong River delta towns, and in the highland centres of Da Lat, Ban Me Thuot, and Pleiku. With the conquest of South Vietnam by North Vietnam, all foreign Christian clergy were expelled.

THE ECONOMY

Vietnam has a definite economic potential. Its population is literate and energetic, and it has a number of natural resources. Its geographic position puts it within easy reach of all the other countries of the region. Vietnam's varied topography lends itself to a wide range of agricultural and industrial exploitation, and its long coastline provides excellent harbours and access to marine resources. In addition to vast areas of cultivable land, there are good soils supporting valuable forests and there are deposits of minerals (including coal) and possibly of crude oil and natural gas. A potential for the generation of hydroelectric power exists. Economic development, however, is seriously constrained by a number of social, political, economic, and military factors.

Economic integration

The reunification of Vietnam required the integration of the society and economy of the south into the socialist framework determined for the new nation by the north. During the period of separation between 1954 and 1975, there were three layers to the Vietnamese economy: the bottom layer was based on the cultivation of rice; the middle layer was dominated by mining in the north and rubber estates in the south; and the third layer was a wartime creation with large-scale Soviet and Chinese aid in the north and substantial U.S. aid in the south. The socialist economy of North Vietnam developed one of the most impressive industrial systems in Southeast Asia, although living standards remained low because of basic inefficiencies, the decision during the war years to disperse industry in the face of U.S. bombing, and the need to maintain a large military complex. The economy of South Vietnam was based largely on free enterprise, with less industry but a more important agricultural sector and a higher standard of living than in the north.

Since reunification in 1976, the task of developing the economy has been rendered difficult by the disruptions and destruction of 30 years of war, the need to integrate the two divergent economies, and other factors. In the north, collective ownership accounted for the vast majority of the means of production. State ownership in the south, however, affected only a relatively small part of the economy, covering most of the total industrial output but leaving the greater part of agriculture, trade, and transport in private hands. The stated intention of the unified government was to extend full socialization to the south on the pattern already established in the north.

'he
outhern
armers

Originally it was announced that the two economies would be gradually integrated after unification. It was decided, however, to accelerate the pace of economic integration. But the program of agricultural reform ran into serious barriers. In the rich Mekong River delta, the government had to deal not with large landlords, but with their former tenants, who had become individual proprietors. The majority of the southern farmers in the late 1970s were middle-income peasants who owned enough land and implements to farm successfully without relying on outside labour. With their skill and experience in farm management, they were the central figures in the rural economy of the south. The poor constituted a minority of the peasantry in the Mekong River delta, whereas they had been in the majority in the north.

The government's move to socialize the commercial sector by abolishing major private enterprises in the south was more successful than the attempt to collectivize agriculture. The government's moves against the private commercial sector, which mainly affected the ethnic Chinese, and deteriorating political relations with China precipitated the flight of ethnic Chinese from Vietnam in 1978. This serious loss of economic talent was compounded by the regime's policy of placing anyone associated with the former South Vietnamese government or the U.S. mission in jails or re-education centres. The resettlement program, which affected vast numbers of people, including highly trained and educated individuals, also resulted in a loss of human resources. The large military organization maintained by the Vietnamese further strains the national budget and diverts large numbers of personnel from the reconstruction of war-damaged roads, railway lines, dikes, irrigation projects, and bridges.

Resources. Rich mineral deposits are a major source of wealth. In the north, there are large reserves of anthracite coal. Raw phosphates and high-grade chromite come from Vietnam. There are also significant deposits of tin, antimony, bauxite, gold, iron ore, lead, tungsten, zinc, and lime.

Vietnam has signed an agreement with a West German firm to explore offshore oil deposits. Under the agreement, an area off the Mekong delta coast has been given to the company for exploration. Similar agreements have been signed with Italian and Canadian companies, and Norwegian and French government aid projects involve seismic surveys and drilling off Vung Tau, a beach resort near Saigon.

Forests are potentially a major natural resource. Fish and shellfish from Vietnam's coastal waters and the plains of the Mekong and Bassac rivers are a major resource and the second most important food staple after rice.

Agriculture, forestry, and fisheries. Agriculture is by far the most important economic sector in Vietnam. It is estimated that nearly three-fourths of the population earns its income from farming. In addition, agriculture is the main source of raw materials for the processing industries and a major contributor to exports. The primary agricultural areas are the two large river deltas and the southern terrace region, the Red River delta and the Mekong River delta. The strip of coastal land between the deltas is a region of low productivity; the central highlands area also is of low productivity, but has a significant agricultural potential.

Rice
production

Rice is the most important crop, grown on about three-fourths of the cropped land, principally in the Red River and Mekong River deltas. Rice accounts for more than half of the total agricultural output. Drying and storage facilities remain inadequate, and there are additional losses due to industrial purposes (such as the production of alcohol) and seed requirements. Other food crops are corn (maize), sweet potatoes, and manioc (cassava).

Agriculture is basically labour intensive in Vietnam, but whether the resettlement of millions of people will affect rice production is uncertain. The greater part of the plowing is still done by water buffalo.

Industrial
crops

Industrial crops include rubber, tea, coffee, fruit trees, sugarcane, tobacco, and jute. Located in the south, the rubber industry was seriously disrupted by the war. Coffee production is mainly of Arabica, Robusta, and Chari varieties. Citrus fruits (oranges, limes, and grapefruit) are important crops, as are bananas and pineapples.

The export of seafoods such as lobsters, prawns, shrimp, and crabs is an excellent potential source of foreign exchange, but production methods are inefficient. Government officials have reported that refugees who left the country by sea took more than 5,000 fishing boats, seriously hampering plans to increase production. The most important fisheries are located on the plains of the Mekong and Bassac rivers. The principal fish farmed are species of tilapia, carp, catfish, and snakehead.

Forestry is a major industry. There is little modernization of forest industries. Charcoal is produced in mangrove areas, and there are a number of furniture and pulp and paper factories, as well as wooden handicraft centres. In the south, most timber has been used for reconstruction.

Industry. At the end of the Vietnam War, industries in the southern and northern parts of the country faced different difficulties. Wartime destruction of the industrial plant had not been great in the south, but the supply of imported raw materials and spare parts had been seriously disrupted. Industries that relied on local raw materials resumed operations quickly, but they experienced problems of unreliable supplies and poor logistics. The task of adequate maintenance, repair, and replacement of industrial machinery was complicated by the heterogeneity of equipment that had been purchased from different countries.

Northern
industrial
development

In the north, there was a period of intense reconstruction of industry from 1973 to 1975. Despite the reconstruction progress, there has been serious underutilization of industrial capacity, as well as unsatisfactory levels of production and a deterioration in the quality of output. Industry has been plagued by old equipment that is subject to frequent breakdowns, inadequate protection of equipment against the weather, insufficient maintenance of tools and materials, and inadequate standards and guidelines for the use of equipment.

Finance and trade. Vietnam's major trading partners are the Soviet Union and Japan; the government joined the Council for Mutual Economic Assistance (Comecon) as a full partner in 1978.

Major exports are coal, rubber, light industrial products, handicrafts, fish, and processed agricultural products. The total value of these exports, and especially of coal, was adversely affected by the refugee exodus and the war in Cambodia and political conflict with China in the late 1970s.

Imports are directed toward food, agriculture, and industry; power; and railways. The volume of imports of raw materials, fuels, fertilizers, and insecticides is expected to grow at the same rate as the projected expansion of related economic activities. The government so far is restricting imports of consumer goods apart from food grains and other essential foodstuffs.

Vietnam is a member of the World Bank and has received aid from the International Monetary Fund. Loans from Denmark, The Netherlands, Kuwait, and the Organization of Petroleum Exporting Countries (OPEC) have been received. After protracted negotiations, Hanoi has agreed to pay the debt owed to Japan by the former Saigon government, and in return it has received a promise of a loan for the purchase of nonmilitary goods. Vietnam has also received aid from Comecon, the Asian Development Bank, and the United Nations Development Program.

Trade unions

Administration of the economy. The state has established strong economic controls. The Vietnam General Confederation of Trade Unions, an instrument of the Communist Party, is the only legal labour organization. North Vietnam had adopted the Soviet-style "single manager system" in industry, giving factory administrators a wide margin of latitude in factory management and placing a heavy burden on labour unions to exercise political leadership among the workers.

Much of the money supply has been eliminated by the change to a single Vietnamese currency because the amount of new currency issued is limited and the use of commercial credit has been virtually abolished.

Economic controls

Income is more evenly distributed because wages have been established by the state. Transportation to and from work, education, medicine, and health services are free. Although housing is relatively poor, it is provided to all workers in state-owned enterprises at a charge of one percent of salary. Price controls and rationing are also used throughout Vietnam to influence the level and distribution of personal income. Food is rationed, and food prices on the free market are much higher than the level of controlled prices. Many industrial consumer goods, such as cloth, are also rationed.

Transportation. The geography of Vietnam has rendered transportation between the north and south difficult. Except for air and sea communications, traffic has been limited to the narrow coastal corridor. The two large deltas, where most of the population is concentrated, have good internal transportation systems based on vast networks of navigable inland waterways, roads, and cart trails. Air travel is being further developed.

The war years brought about the major development of roads in the north and the south. In the south, the emphasis was on the construction and improvement of major arterial highways, while in the north priority was given to the construction of secondary roads.

Port facilities

Vietnam's rail system suffered heavy damage during the war years. After unification, a comprehensive reconstruction and modernization program was launched. There is an extensive network of navigable rivers and canals in the Red River delta (3,700 miles, or 6,000 kilometres) and the Mekong River delta (3,000 miles, or 4,800 kilometres). Maintenance work to avoid silting has been hampered by wartime damage to dredging equipment and by general deterioration. Most coastal and ocean shipping is centred in the northern port of Haiphong and the southern port of Saigon. The ports of Hon Gai and Cam Pha are used mainly as export stations for coal, while Vinh and Da Nang serve as transit ports for Laos. The war brought considerable improvements to port facilities at Cam Ranh, Nha Trang, and Qui Nhon.

Noi Bai airport at Hanoi was opened to international traffic in 1978. Soviet, East German, Chinese, and Laotian aircraft have begun using the new facility. Vietnam agreed to open the air route across the central part of the country to international airlines, with some exceptions, notably those based in the United States.

ADMINISTRATIVE AND SOCIAL CONDITIONS

Government. The elections of April 25, 1976, produced a new unicameral National Assembly of 492 members (243 from the south and 249 from the north), the major task of which was to establish a single national government for the unified nation. The principle of universal suffrage for those 18 years of age or older determined the electorate. On June 24 the assembly met for the first time, and on July 2 it exercised its first prerogatives, creating the Socialist Republic of Vietnam and electing the nation's president, two vice presidents, and cabinet, and the members of the Standing Committee of the National Assembly. The primary task of the Standing Committee was to draft a new constitution, which was published in August 1979.

The National Assembly is the supreme organ of the government. The Council of Ministers is comprised of the prime minister and deputy prime ministers, all of whom are named by the National Assembly, and the heads of government ministries and various state organizations. The Council coordinates and directs the activities of the ministries and the various state organizations at the level of the central government and supervises the activities of administrative committees at the level of local government.

Each ministry is headed by a minister, who is assisted by several vice ministers. Responsibilities of the ministries usually are defined along narrow functional lines; there are, for example, numerous economic ministries concerned with subsectors such as grains and food products, marine products, afforestation, water conservancy, light industry, engineering and metals, power and coal, and construction. Larger ministries tend to be relatively self-sufficient, with their own colleges, training institutions, and health, social, and cultural facilities. There are also a number of commissions under the Council of Ministers, including the State Planning Commission, headed by a deputy prime minister; the State Bank, headed by a director general with ministerial rank; and the Central Agriculture Commission. Under the prime minister's office are a number of general departments that have a lower status in the administration than ministries. The more important departments are headed by ministers. Committees under the prime minister's office are also formed to supervise projects, such as the Da River hydroelectric project, that involve more than one ministry.

The traditional administrative units of local government in Vietnam have been the provinces, districts, and villages. Following reunification in 1976, local administration was reorganized into 35 provinces and 3 municipalities (Hanoi, Haiphong, and Saigon) on the provincial level. There are about 500 districts. At either the provincial or district level, the highest government authority is an elected People's Council, the actual work of which is carried out by administrative committees appointed by the councils. Village administration is represented by village people's councils.

The Vietnamese Communist Party

The most important political institution in Vietnam is the Vietnamese Communist Party (Dang Cong san Viet Nam), founded in 1976 from the Vietnamese Workers' Party (Lao Dong) of North Vietnam. Any citizen of the age of 18 or over who "has engaged in labour and not been an exploiter" is eligible to join.

Effective authority within the Vietnamese Communist Party lies with the Politburo's 14 voting members and 3 alternates and the Central Committee's 101 full members and 32 alternates. Of the two bodies, however, the Politburo has more political power. It is not known precisely how decisions are made within the Politburo, but it appears that they result from a process of compromise among its senior members. The secretary general of the Vietnamese Communist Party presides over the party's Secretariat, which is elected by the Central Committee. The Secretariat directs the tasks of the party organization and the coordination of party and state bureaucracies in carrying out the resolutions of the Central Committee and the Politburo; it also has responsibility for implementing party resolutions.

Popular associations for youth, women, and farmers disseminate party policies and serve as training grounds for potential party members. The most important of these organizations is the Vietnamese Women's Union, which reflects the critical role played by women in Vietnam's revolution. The Ho Chi Minh Communist Youth Union is largely responsible for the Youths' Federation, while local party units and agricultural cooperative organizations assume leadership over the Farmers' Federation. The Vietnam Federation of Trade Unions has the responsibility of safeguarding workers' welfare, but does not function as a Western-style bargaining unit.

The National United Front

Non-party political organizations are grouped under the Fatherland Front. In 1955, the Fatherland Front grew out of the Lien Viet Association, an amalgamation of non-communist political groupings opposed to the French during the Indochina War. In 1977, the National United Front was organized with the merger of the Fatherland Front, the National Liberation Front (NLF) of South Vietnam, and the Vietnam Alliance of National Democratic and Peace Forces (which had been formed in 1968); the last two had presented themselves as independent southern movements.

Justice. The judicial system consists of the courts and the People's Organs of Control. The National Assembly supervises the work of the Supreme People's Court, which is the highest court of appeal and the court of first instance for special cases (such as those involving treason). This court in turn supervises the judicial work of local People's Courts, which are responsible to their corresponding People's Councils. The People's Courts function at all levels of government except the village level, where the village administrative committee functions as a primary court.

ople's rgans of ontrol

The People's Organs of Control act as watchdogs for the state; they monitor the performance of government agencies, maintain vast powers of surveillance, and act as prosecutors before the People's Courts. The People's Supreme Organ of Control is responsible only to the Standing Committee of the National Assembly.

Armed forces. Military forces include the army, paramilitary regional and provincial forces, the militia, and the reserves. There are separate military commands in Hanoi, Haiphong, and Saigon.

Education. Because of their Confucianist traditions, education has always been important to the Vietnamese. Rural education in the south was badly disrupted during the war years, and all religious and private schools were nationalized after 1975. The government pursued a policy of education reform. Twelve years of schooling are provided free to everyone. The educational system is divided into three levels; after the first nine years of schooling, students are selected to continue into higher education or to attend technical or vocational training institutions.

The government claims that illiteracy has been abolished in the south, as it already had been in the north. Emphasis is being placed on training in science and technology, although a lack of equipment hinders the program, and several thousand students are sent to study languages and technology abroad. While most of the students sent to other countries go to the Soviet Union and eastern Europe, some study in France, Great Britain, and Australia.

Health and welfare. Before unification, health services were underdeveloped in the rural areas of South Vietnam, but were well developed in North Vietnam. After reunification there was a general increase in health facilities and personnel (physicians, assistant physicians, nurses, and midwives) throughout Vietnam. Health facilities include hospitals, health centres, leprosy centres, sanatoriums, and village health and maternity centres.

CULTURAL LIFE

npor- nce of etry

Chinese influence permeates all aspects of traditional Vietnamese culture, and it is strongly manifest in language, art, architecture, music, theatre, literature, and poetry. Deeply rooted in the Vietnamese oral tradition, poetry was highly regarded by the educated class centred in the royal courts, and it was expressed in Chinese form and style. By the 14th century, however, a demotic script called *chu nom* ("southern characters") came into use among the Vietnamese literati, permitting the emergence of an explicitly Vietnamese "tale in the southern script" (*truyen nom*) that, during the 17th and 18th centuries, evolved into the form of a long narrative poem. This art form reached its culmination in the best known of Vietnamese poems, *Kim Van Kieu* ("The Tale of Kieu") by Nguyen Du (1765–1820). Further evolution of this literary tradition was arrested by the French conquest and the imposition of French culture on the Vietnamese elite. Despite this and later influences from the outside, poetry has retained its importance in the Vietnamese artistic tradition. Vietnamese poetry is now written in a romanized script based on the first Latin script for the tonal language devised by Father Alexandre de Rhodes, a Jesuit priest from Avignon.

Traditional Chinese opera, called *hat tuong* in the north and *hat boi* in the south, is popular among the Vietnamese, as is the more indigenous *cai luong*, a satirical musical comedy. The theatre is strictly controlled and all actors and other performers are unionized employees of the state. Painting has failed to flourish among the Vietnamese; it has been bound for centuries first by rigid traditional Chinese forms, then by a style imitative of French Impressionism, and now by the tenets of Socialist Realism. High quality lacquer ware, however, continues to be produced.

Folk traditions flourish among the peoples of the central highlands, who continue to produce most of the things they need. Precious hardwood is used for carving crossbows and figures, most of which are destined for tombs. Among the Rhadé, a mountain tribal group living mainly in Dac Lac province, hardwood is used in the construction of longhouses built on pilings. Among the highlanders, women weave blankets, blouses, skirts, and loincloths, while the men weave baskets and mats. They have a variety of musical instruments, but gongs are the most common. The Cham and Khmer minorities retain some folk arts, but as they become assimilated into Vietnamese culture, their traditions are fading. (G.C.H.)

For statistical data on the land and people of Vietnam, see the *Britannica World Data* section in the BRITANNICA WORLD DATA ANNUAL.

History

In spite of a growing number of ethnographic and archaeological studies covering the territory historically occupied by Vietnam, all knowledge about the origin of the Vietnamese people remains hypothetical. A long-held notion that identified the Vietnamese with one tribe of the Viets of southern China (Yüeh in Chinese) has been abandoned. Though the Viets—who about 334 BC were driven by the expanding Chinese from their habitat south of the Yangtze River, moved farther south, and may have entered the Red River Delta—gave the Vietnamese people its name, the theory that regards them as direct ancestors conflicts with ethnographic and biological evidence, all of which points to strong social and cultural affinities between the Vietnamese and peoples of the Tai and Indonesian families. The dominant view is that the Vietnamese people originated in the Red River Delta and represent a racial and cultural fusion, the various elements of which are still not fully determined.

ETHNIC ORIGINS AND CHARACTERISTICS

As is true of all of the peoples of Southeast Asia, diversity characterizes the evolution of the peoples who inhabited the regions of Indochina now occupied by the Vietnamese. There is little controversy concerning the earliest known inhabitants of Southeast Asia: they were Negroid Pygmies who may have arrived in Indochina between 25,000 and 50,000 years ago and of whom remnants are still found in Malaya and the Philippines. They were followed by Australoid and Melanesian Negritos of the Australian and Papuan species. Skulls found along the coast of the South China Sea confirm that Negritos were the earliest known inhabitants of Vietnam. Finds at Hoa Binh and Bac Son belonging to the Paleolithic Period show a preponderance of Melanesian and other Indonesian types, as do, to an even larger degree, Neolithic finds of the 2nd millennium BC.

Early inhabitants of Vietnam

Although the Negritos have disappeared from Vietnam, many racial characteristics among the mountain tribes of South Vietnam appear to support the view that some were absorbed by later arrivals in Indochina. These were the Austronesians (Indonesians), the main groups of whom penetrated from southern China into the eastern regions of the Indochinese peninsula probably between 2500 and 1500 BC, some possibly even earlier.

Soon afterward there arrived the ancestors of two Austroasiatic families. These newcomers, who may have started a movement of the Austronesians toward the island world of Indonesia, never left the mainland. These were linguistically classified groups of Mon-Khmer and Malayo-Polynesian peoples. Descendants of the Mon-Khmer still inhabit Cambodia and the lowlands of Thailand and

Myanmar, and it was a Malayo-Polynesian people that formed the Kingdom of Champa, situated along the east coast of Indochina. Racial and linguistic characteristics indicate that these Austroasiatics have contributed to the formation of the Vietnamese people. Both linguistic groups are still represented among the mountain tribes of South Vietnam and Laos.

Formation of the Vietnamese people

The formation of the Vietnamese people as a distinct ethnic group, however, took place long after the arrival of the Austronesians and Austroasiatics in Indochina. In the Bronze Age, which probably began in Vietnam some time between 600 and 400 BC, the inhabitants of the Red River Delta still were essentially of Indonesian stock. These people, who brought bronze from southern China, were probably the last Indonesian immigrants to Indochina. They were soon followed by immigrants of the Viet and Tai families—who added a Mongoloid strain to the mixed Indonesian stock in the Red River Delta—and, after 111 BC, by Chinese soldiers, administrators, and later Chinese refugee scholars. During these centuries—between *c.* 200 BC and AD 200—the Vietnamese were formed as a separate ethnic entity through a fusion of all these elements.

Language research, which offers a more reliable way of distinguishing the various ethnic groups of Southeast Asia, confirms the mixed racial and cultural origin of the Vietnamese people. Although the Vietnamese language is a distinct entity, it can nevertheless also be described as a fusion of Mon-Khmer, Tai, and Chinese elements. From the monotonic Mon-Khmer language family, Vietnamese derived many of its basic words; from the Tai languages, it took tonality and a number of grammatical elements; and from the Chinese, in a more advanced state of civilization, it took not only a script but also most of its governmental, literary, philosophical, and technical vocabulary.

Ethnography also reveals the extent to which the ancient Vietnamese culture was a composite of elements found among most peoples of the mainland and islands of Southeast Asia. Totemism, animism, tattooing, the chewing of betel nuts, the blackening of teeth, and many marriage rituals and seasonal festivals indicate the relationship between the Vietnamese and peoples of Melano-Indonesian, Khmer, and Tai stock; customs, religion, historical legends, and social organization all underline this relationship. To be sure, Chinese civilization later became the main force in shaping Vietnamese culture, and Indian civilization, via the states of Champa and Cambodia, was a secondary influence. But the failure of the Chinese to absorb the Vietnamese people underscores the fact that strong elements of an authentic local culture must have been developed in the Red River Valley long before China established its 1,000 years of rule over Vietnam.

LEGENDARY AND EARLIEST KNOWN HISTORY OF VIETNAM

According to their most authoritative legends, the history of the Vietnamese people begins with King De Minh, a descendant of a divine Chinese ruler who was also the legendary father of Chinese agriculture. De Minh and an immortal fairy of the mountains produced Kinh Duong Vuong, ruler of the Land of Red Demons, who married the daughter of the Dragon Lord of the Sea. Their son, Lac Long Quan, Dragon Lord of the Lac, is regarded as the first authentic Vietnamese king (Lac was the first designation for the Vietnamese). To make peace with the Chinese, Lac Long Quan married Au Co, a Chinese immortal, who bore him 100 eggs, from which sprang 100 sons. Later, the King and Queen separated; Au Co moved with 50 of her sons into the mountains, and Lac Long Quan kept the other 50 sons and continued to rule over the lowlands. Lac Long Quan's oldest son succeeded him as Hung Vuong (Vuong means king). Hung Vuong is regarded as the real founder of the Vietnamese nation and of the first Vietnamese dynasty—the Hong Bang.

Lac Long Quan

This legend and other related legends, most of which received their literary form only after AD 1200, describe in mythical terms the fusion, conflicts, and separation of peoples from the north and south and of peoples from the mountains and the lowlands near the sea. The legends show the immortals as mountain dwellers, while the people along the sea are represented by the Dragon Lords—a division found in many legends among the Indonesians all over Southeast Asia. The retreat of Au Co and 50 of her sons into the mountains may well be a mythical record of a separation among the proto-Vietnamese in the Red River Delta: those who left the lowlands could be the ancestors of the Muong, who still live in the hills surrounding the Red River deltas and who are the only ethnic minority of Vietnam closely related in language and customs to the Vietnamese.

The Hong Bang dynasty, legend continues, had 18 kings, each of which ruled for about 150 years. Their country, called Van Lang ("Land of the Tattooed Men"), is said to have extended not only over the Red River Delta but also over much of southern China. The last of the Hong Bang kings was overthrown by the ruler of neighbouring Thuc, who invaded and conquered Van Lang, united it with Thuc, and called the new state Au Lac, which he ruled under the name Au Duong Vuong. Au Lac existed only until 208 BC, when it was incorporated by the former Chinese general Trieu Da (Chao T'o in Chinese) into the kingdom of Nam Viet (Nan Yüeh in Chinese).

Nam Viet covered much of southern China and was ruled by Trieu Da from his capital near the present site of Canton. Its population consisted chiefly of the Viets who had earlier been driven by the Chinese from their kingdoms south of the Yangtze River. Trieu Da, after throwing off Chinese sovereignty and killing all officials loyal to the Chinese emperor, adopted the customs of the Viets and made himself the ruler of a vast non-Chinese empire. After it had incorporated Au Lac, Nam Viet included not only the Red River Delta but also the country as far south as modern Da Nang. The end of Au Lac in 207 BC marks the end of legendary history and the beginning of Vietnamese history as recorded in Chinese historical annals.

After almost 100 years of diplomatic and military duels between the Chinese empire and Trieu Da and his successors, Nam Viet was conquered (111 BC) by the Chinese under the Han emperor Wu Ti. Thus did the territories occupied by the ancestors of the Vietnamese fall under Chinese rule. Nam Viet became the Chinese province Giao Chi (later Giao Chau), which was divided into nine military districts. The three southernmost of these covered the northern half of what is now Vietnam.

Early Vietnamese history

When China extended her rule over Vietnam the people of the Red River Delta were about to enter from the Bronze into the Iron Age, although some stone implements were also still in use. These ancestors of the Vietnamese were already experienced in the cultivation of wet rice. They had learned how to irrigate their rice fields by using the tides that backed up the rivers. Ploughs and water buffalo were still unknown (the land was prepared for cultivation with polished stone hoes), but the proto-Vietnamese are thought to have been able to produce two rice crops annually. They also fished and hunted. Their weapons were mainly bows and arrows; the bronze heads of their arrows were often dipped in poison to facilitate the killing of such larger animals as elephants, whose tusks were traded for iron from China.

The social organization of the early Vietnamese, before Chinese rule, was hierarchical, forming a kind of feudal society that until recently still existed among the Tai and Muong minority populations of North Vietnam. Power was held by tribal chiefs at the head of one or several communities. These chiefs were civil, religious, and military leaders, and their power was hereditary; they were large landowners who kept the mass of the people in virtual serfdom. At the head of this aristocracy stood the king, probably the most powerful among the tribal chiefs.

Religion was characterized by animistic beliefs in supernatural beings and spirits common among primitive agricultural and hunting peoples. Some of these were the spirits of dangerous animals, others of deceased important persons who, like spirits generally, needed to be propitiated. A great religious festival, almost a carnival, was held at the beginning of spring and was marked by abandon and promiscuity.

In all these respects, the inhabitants of the Red River Delta, prior to their subjugation by the Chinese Empire, showed numerous affinities with most of the people of mainland and island Southeast Asia—the Tai, the Mon-Khmer, and the Melano-Indonesians. It was not until several centuries after the imposition of Chinese rule that the Vietnamese developed the more distinct ethnic characteristics of a people destined to become a nation of their own.

FUNAN AND CHAMPA

pread
f Indian
ıfluence

The formation of the Vietnamese people proceeded under Chinese political rule, and two states arose in the regions of what is now South Vietnam. The political institutions, religion, philosophy, and art of the people were shaped in part by the spread of Indian influence over the Indochinese peninsula. One of these states, Funan, was founded in the 1st century AD, extending over the Mekong Delta and the entire region of what is now Cambodia. The other state, Champa—founded, according to Chinese annals, in AD 192—occupied the east coast of Indochina from the confines of the Mekong Delta to as far north as the mountain ridge of Hoanh Son ("Horizontal Mountain") near the 18th parallel. A period of Chinese imperial decline explains the ability of the founders of Champa to include in their kingdom the southernmost and strongly Indianized province of Chinese-occupied Giao Chau. The formation of these two states was not due to Indian military conquests. The Austroasiatic peoples of these regions of Indochina became Indianized through lively commercial relations with India and through the penetration of Indian priests and literati before the rise of both kingdoms.

Although Indianized Funan dominated much of Indochina for several centuries and occupied a region that later became a vital part of Vietnam, no direct contacts existed between the people of Funan and the Vietnamese. The kingdom of Funan disappeared during the 6th century, almost 1,000 years before the Vietnamese reached the Mekong Delta. Funan was conquered by a people from the north, the Mon-Khmer, who in the next few centuries founded the powerful Cambodian Empire. Whatever influence the Indianized culture of Funan had on the evolution of Vietnamese culture was exercised indirectly via Champa and later via Cambodia.

Champa was still a strong neighbour of Vietnam after the Vietnamese had made themselves independent of China in the 10th century. Like Funan before it, Champa had a powerful fleet, which the Chams, who lacked enough land for agriculture along the mountainous coast, employed for commerce and in plundering campaigns against their neighbours. For many centuries Champa was a semipiratical country that, from hidden bays, conducted attacks on trading ships along the entire east coast of Indochina and waged border wars even against China.

Cham civilization also was thoroughly Indianized. Sanskrit was used as a sacred language, and Indian influence dominated art and agriculture, as is evidenced by the collection of Cham art in the Museum of Da Nang and by the ruins of a Cham city in the province of Quang Nam.

Champa's aggressiveness and Vietnam's need for territorial expansion explain the centuries of clashes between the two countries, which, toward the end of the 15th century, led to the virtual destruction of Champa. For this reason Champa's history can be viewed to advantage only in connection with the history of Vietnam, both while it was under Chinese rule and after it had become independent in the 10th century.

VIETNAM UNDER CHINESE RULE (111 BC–AD 939)

The history of the Vietnamese people under more than a thousand years of Chinese rule reveals an evolution toward national identity, apparently inevitable when seen under three aspects. The first of these aspects is the introduction into the Red River Delta of the more advanced civilization of China, including technical and administrative innovations and the higher degree of Chinese learning, which made the Vietnamese the most advanced people of the Indochinese peninsula. The second is the efforts of the Chinese governors to achieve complete Sinicization through the imposition of Chinese culture and customs and the employment of political power. The third and most significant aspect of Vietnamese history during this period is the resistance of the Vietnamese people to total assimilation and the use they made of the benefits derived from Chinese civilization in their struggle against Chinese political rule.

ıtroduc-
on of
hinese
viliza-
on

Soon after extending their domination over what is now North Vietnam, the Chinese constructed roads, waterways, and harbours to facilitate communications with, and to ensure their administrative and military control over, these newly acquired provinces. They improved local agriculture with better methods of irrigation and the introduction of the metal plough and draft animals. They brought with them new tools and weapons, advanced the art of pottery, and used new mining techniques. But for more than 100 years after annexing Nam Viet in 111 BC, they abstained from interfering with the local administration. In the Chinese province of Giao Chau, the hereditary lords exercised control over the peasant population just as they had done while Giao Chau was a province of Nam Viet. Thus, although divided into military districts headed by Chinese governors, Vietnam remained, in fact, a leniently governed Chinese protectorate.

This form of government changed only at the beginning of the Christian Era, when an energetic governor realized that the sway of the local Viet lords over the population was an obstacle to Sinicization. The desire to exploit the fertile Red River Delta and its mountainous backcountry was certainly one reason the expansionist Han dynasty wanted to hold on to Vietnam: there were great forests and precious metals in the mountains and pearls in the sea, there was ivory, and there was a peasantry to be taxed and recruited for forced labour. China's main interest in holding the Red River Delta, however, was its value as an important stopover for ships engaged in China's nascent maritime trade with the Spice Islands, India, and even the Near East. Vessels from many countries with which China developed commercial relations docked at the harbours of the Vietnamese coast, not only bringing new goods but also establishing contacts with a wider world and thus promoting the development of the country. In this process, which began early in the 1st century AD, economic, cultural, and political functions developed that the hereditary local lords were unfit to discharge—another reason why direct Chinese rule through the importation of more and more Chinese officials became necessary.

Attempts
at Sinicization

As in all regions conquered by the Chinese Han dynasty (206 BC–AD 221, with a brief interruption between AD 8 and AD 23), efforts to set up direct Chinese rule were accompanied by a variety of attempts to transform the people of the Red River Delta into Chinese. Local customs were suppressed, and Chinese customs, rites, and institutions were imposed by force. Taoist and Confucian learning was pressed, together with the Chinese language; and even Chinese clothing and hairdress became obligatory. Many of these Chinese innovations were beneficial to the Vietnamese and were readily integrated into the indigenous local culture, but Sinicization never succeeded in reconciling the Vietnamese people and especially their leaders with Chinese political domination. Not only the masses of the people but even the educated Vietnamese who knew Chinese and wrote only in Chinese held on to the local spoken language.

Events in China accelerated the efforts at Sinicization in the Red River Delta. When a usurper, Wang Mang, took power from the Han dynasty between AD 8 and 23, thousands of officials, soldiers, and scholars fled China southward into the province of Giao Chau, the governor of which had refused to recognize the new imperial government. He thus received the manpower needed to push ahead with his policy of replacing local lords with Chinese officials and of spreading Chinese ideas and ways of life. Only then did the hereditary chiefs, after having submitted to Chinese overlordship for more than 100 years, begin to offer resistance.

Vietnamese
resistance

The first major rebellion against Chinese rule broke out in 39, led by the noble lady Trung Trac, whose husband, a tribal lord, had been executed by the Chinese. She and her sister Trung Nhi gathered the tribal chiefs and their armed followers, attacked, overwhelmed the Chinese strongholds, and had themselves proclaimed queens of an independent

Vietnamese kingdom. Three years later a strong army sent by the Han emperor re-established Chinese rule; the local aristocracy was deprived of all power; Vietnam was given a centralized Chinese administration; and Sinicization was resumed with increased intensity.

Although resistance to total Sinicization continued, more than 200 years passed before the Vietnamese were able to stage the next uprising. It too was led by a woman, Trieu Au, but it was crushed, after only six months. Like the Trung sisters before her, Trieu Au is said to have ended her life by throwing herself into a river.

This second uprising, which took place in 248, was preceded by far-reaching changes both in China and in the Chinese occupied Vietnamese lands. The decline of the Han dynasty before its fall in 220 had weakened Chinese power at the periphery of the empire. The rising kingdom of Champa not only took from Chinese control the southernmost regions of Giao Chau but soon began to attack the Chinese possessions farther north and even the coast of China itself. In the Red River Delta the Chinese governor Che Sie (ruled 187–226) had made Giao Chau virtually independent of China, which, after the fall of the Han dynasty, was split first into three contending kingdoms and later, up to 589, into six warring dynasties. During these centuries the number of Chinese immigrants, many of them scholarly officials, in Giao Chau increased; there was continuous intermarriage and a systematic dissemination of Chinese learning among the Vietnamese. A Sino-Vietnamese elite, whose members, although a product of Chinese civilization, nevertheless felt and acted more and more like nationalist Vietnamese, developed; and a representative of this class, Ly Bon, himself of Chinese ancestry, *c.* 542 led another rebellion that broke Chinese rule and for several years established another independent Vietnamese kingdom. Though he was defeated after only three years in power, his followers controlled distant mountainous regions of Vietnam until 603.

Chinese rule, although challenged several more times, remained secure as long as China was effectively ruled by the T'ang dynasty (618–907). But after the fall of the T'ang, the struggle to free Vietnam led to an almost uninterrupted series of uprisings that culminated in a disastrous defeat of a Chinese army in 939, the year from which the Vietnamese date the beginning of their independence.

THE FIRST PERIOD OF INDEPENDENCE (939–1407)

Ngo Quyen, the Vietnamese commander who deafeated the Chinese in 939, became the first head of the new state of Vietnam. To underline the continuity of Vietnam's national existence, he set up his capital at the site from which the country had been ruled before it was conquered by the Chinese. But for more than half a century independence brought to Vietnam neither peace nor political stability. When the reign of Ngo Quyen ended after only six years, Vietnam was torn apart by a dozen local military chiefs who, during the next 22 years, fought each other for power. A measure of stability was created by Dinh Bo Linh, who defeated the local lords, proclaimed himself emperor in 968, and pacified the new Sung dynasty (960–1279) of China with tributes. Still more progress was made under Le Hoan, the founder of the Earlier Le dynasty, who ruled Vietnam from 980 to 1005. But already in 981 he had to fight off the first Chinese attempt to reconquer Vietnam, and soon afterward he conducted several campaigns against the aggressive Chams to the south.

The Ly dynasty. Internal stability and economic progress, through the consolidation of the monarchy at the head of a centralized administration, came to Vietnam only under the first of her three great dynasties—the Ly, who ruled from 1009 until 1225. They called their country Dai Viet, but the Chinese continued to call it Annam. The Ly built many new dykes and canals and were famous as promoters of agriculture. As early as 1044 they began to build roads between the main cities, and they also founded the first postal service. They moved the capital back to Hanoi, then called Dong Kinh, which later as Tonkin became a European designation for northern Vietnam. The Ly gradually replaced the divisive local lords with a hierarchy of state officials trained in a civil service institute set up in 1076. Buddhist learning, widespread in Vietnam long before independence, was vigorously promoted, as were art and literature, the latter chiefly through the dissemination of classical Chinese texts. But although Dai Viet under the Ly made great political, economic, and cultural progress, its prosperity was constantly threatened by external wars. In 1057 the Chinese advanced once more against the Red River Delta, but they were met by the Vietnamese and defeated on Chinese soil in a war that lasted four years. Champa and also Cambodia, which had by then become the greatest power in Indochina, invaded the southern provinces of Vietnam while the Vietnamese were fighting off the Chinese. Cambodia attacked Vietnam again in 1128 and 1132, together with Champa, and staged five more invasions of Vietnam between 1138 and 1216.

The Tran dynasty. By that period, the Ly dynasty was already in decline. It was followed, after many years of civil strife, by the second great Vietnamese dynasty—the Tran, which lasted from 1225 to 1400. For most of their rule, the Tran successfully pursued the same 'olicies that had made the country strong under the Ly; the Tran too were harassed by Champa, but they managed to arrange several periods of peaceful coexistence. In the eyes of the Vietnamese, the greatest achievement of the Tran dynasty was the defeat of the Mongolian armies of Kublai Khan, who was about to make himself master of the Chinese Empire, replacing the Sung dynasty in China in 1279. Three Mongol invasions in 1257, 1284, and 1287 of armies said to have numbered between 300,000 and 500,000 men succumbed to Vietnamese resistance. The Tran general who defeated the Mongolian invaders, Tran Hung Dao, is still venerated as one of the great heroes of Vietnamese history.

Defeat of the Mongols

Vietnam had hardly recovered from these invasions when the war with Champa was resumed. For ten years after 1312 Champa was a feudatory state of Vietnam, but it was freed under a new king, Che Bong Nga, who invaded Vietnam several times and in 1371 actually stormed and pillaged Hanoi.

The drain of these wars on Vietnam's resources, together with the declining vigour of the Tran rulers, led to a deep economic and social crisis, which a high official, Ho Qui Ly, exploited to usurp the throne in 1400. Ho Qui Ly was a capable and bold reformer, and the country began to recover under his reign, but he was not to be given the time needed to consolidate his power. The deposed Tran appealed to the Chinese to help them regain the throne. China, then ruled by the Ming dynasty, which had ousted the Mongols in 1368 but had not broken with China's policy of imperialist expansion, readily complied with the Tran's request and invaded Vietnam in 1407. Vietnam succumbed, largely because the country was politically divided. Instead of re-establishing Tran rule, the Ming set up a direct Chinese administration; the country was ruthlessly exploited, and the old policy of assimilation resumed with such severity that the survival of the Vietnamese people as a separate nation depended once more on their ability to free themselves from Chinese rule.

EXPANSION, DIVISION, AND REUNIFICATION (1400–1802)

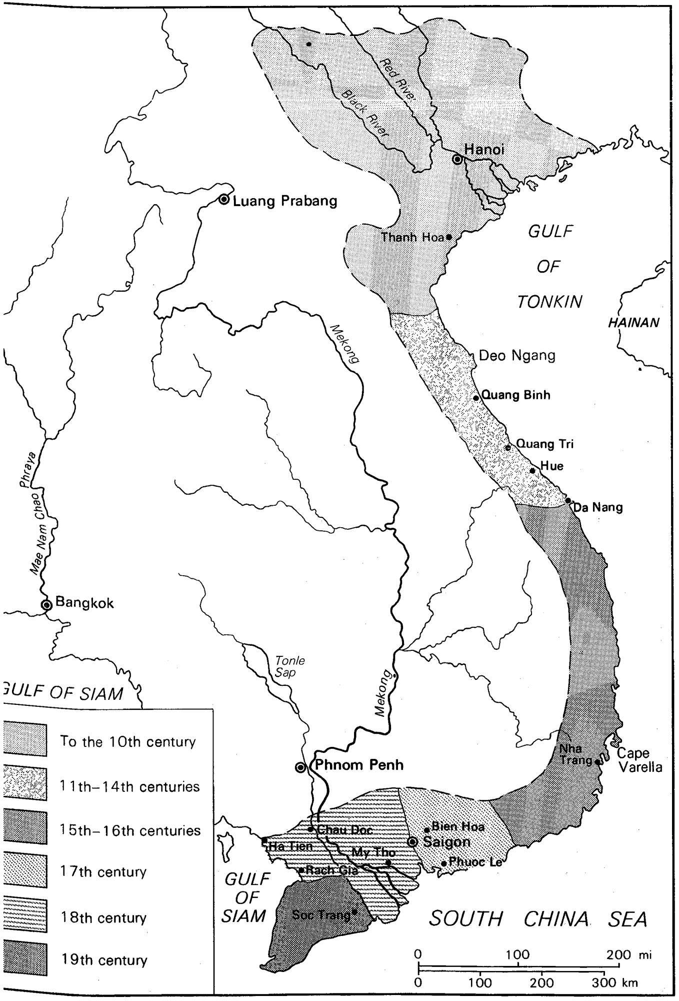
By the beginning of the 15th century, the evolution of the Vietnamese people had reached a point at which the attempt to make them Chinese could only strengthen their nationalist sentiments and arouse their determination to throw off the Chinese yoke. Under the leadership of Le Loi, a rich landowner in the province of Thanh Hoa south of the Red River Delta, a movement of national resistance started in 1418 and forced the Chinese to evacuate after ten years of struggle. Le Loi, who ascended the throne shortly afterward under the name of Le Thai To, became the founder of the third great Vietnamese dynasty, the Later Le (sometimes simply referred to as the Le), which, although not actually in power after 1600, headed the state until 1787.

The Later Le dynasty. Like the better rulers of the Ly and Tran dynasties, Le Thai To and some of his successors introduced many reforms. They gave Vietnam the most advanced legal code in Southeast Asia; promoted art, literature, and education; advanced agriculture; protected communal lands against the greed of large landowners; and even enforced a general repartition of land among the entire population at the expense of the large landowners. But the problem of the landless remained acute, due to population increase and the limits of available land in the north. The lack of land was one of the reasons the Le dynasty pursued a policy of territorial expansion and was a chief motive behind their efforts to drive the Chams from the small but fertile deltas to the south. Most of Champa was in fact conquered in 1471 under the leadership of Le Thanh Tong (reigned 1460–97). Masses of landless people were settled during the next decades in soldier-peasant villages all the way south from the vicinity of Da Nang to the neighbourhood of Nha Trang, the first great push in the Vietnamese march to the South. The elimination of Champa was followed by incursions into the Cambodian territory of the Mekong Delta, which the declining Khmer Empire was no longer able to defend. Saigon became Vietnamese shortly before 1700, and the rest of the south followed during the next 60 years. With the exception of the province of Soc Trang, which was not annexed until 1840, Vietnam had reached its present size by 1757.

This extension of Vietnam to a length of 1,000 miles altered the historical evolution of the state, the chief characteristic of which had been the existence of a strong central power at the head of a unified administration. Vietnam was divided twice, and its partitioned governments were at war with each other for countless years.

Two divisions of Vietnam. The first and shorter division of the country occurred soon after the elimination of Champa. The governor of Hanoi, Mac Dang Dung, made himself master of Vietnam in 1527. The deposed Le rulers and the generals loyal to them regained control of the lands south of the Red River Delta in 1545, but only after almost 50 years of civil war could they reconquer Hanoi and the north.

Adapted from Andre Masson, *Histoire du Vietnam* (1972), in series *Que Sais-je?*; Presses Universitaires de France



Expansion of precolonial Vietnam.

Of much longer duration and greater historical significance was the second division of Vietnam, which occurred about 1620, when the noble family of the Nguyen, who had governed the country's growing southern provinces from Hue since 1558, refused obedience to Hanoi. In Hanoi the Le monarchs were rulers in name only after the country was reunited following its first division; all real power was in the hands of the Trinh family, who had made themselves hereditary princes in charge of government. For 50 years the Trinh rulers tried in vain to regain control of the southern half of the country by military means. The failure of their last campaign in 1673 was followed by a 100-year truce, during which both the Nguyen and the Trinh paid lip service to Vietnamese unity under the Le dynasty but maintained separate governments over the two halves of the country.

National unity was re-established only after a period of revolution, political chaos, and civil war, which lasted from 1772 until 1802. Although the revolution started in the south, it was directed against the ruling houses of both south and north. It was led by three brothers, whose name in history—Tay Son—was that of their native village. The Tay Sons overthrew the southern regime in 1777 and killed the ruling family. While the Tay Sons waged war against the north, one member of the southern royal family—Nguyen Anh, who had escaped the massacre—regained control of Saigon and the deep south in 1778, but he was driven out again by the Tay Sons in 1783. When the Tay Sons also defeated the Trinh in 1786 and occupied Hanoi, Vietnam was briefly reunited under their rule. In 1788 the Chinese tried to exploit the crisis of Vietnam, but the Tay Son rulers, who had abolished the Le dynasty, were able to defeat the Chinese invaders. During the same year, however, Nguyen Anh succeeded, with French military assistance, in occupying Saigon and the Mekong Delta. In a series of campaigns that lasted 14 years, Nguyen Anh defeated the Tay Sons and gained control of the entire country. When Hue and Hanoi fell to his armies in 1802, he proclaimed himself emperor of reunited Vietnam under the name of Gia Long.

STATE AND SOCIETY OF PRECOLONIAL VIETNAM

Organization of the state

The rule of Gia Long and his successors up to the conquest of Vietnam by France in the late 19th century brought no innovations in the organization of the state, the basic character of which had already been firmly established by the Ly emperors during the 11th century. The Ly had successfully fought the revival of a local feudalism, which was rooted in the powers exercised by the tribal chiefs before the coming of the Chinese. From the 11th century on, Vietnam remained a centralized state headed by a monarch whose absolute powers were said to derive from a mandate from Heaven—one aspect of the thoroughly Confucian character of the Vietnamese state. Following the Chinese model, the Ly established a fixed hierarchy of state officials with nine degrees of civil and military Mandarins, appointed by the emperor and responsible only to him. All Mandarins—those at the very top at the imperial court as well as the lowest ranks in the provincial and local administration—were recruited in only one way: through civil service examinations taken after years of study. As a rule, only the wealthy could spend the time required for these studies. Nevertheless, except in periods of dynastic decline when offices were sometimes for sale, the road to positions of power was scholarship, not wealth.

The concept of a division of powers was alien to the precolonial rulers. The emperor, with the help of high-court Mandarins, was not only the supreme lawmaker and head of all civil and military institutions but also the dispenser of justice in both criminal and civil cases, and he delegated his powers to the hierarchy of Mandarins in the provinces and villages. Even public functions of a religious character were the sole prerogative of the emperor and his representatives on the lower levels of the administration. No military caste ever exercised control over the state, no religious hierarchy existed outside the Mandarins, and no aristocracy with political influence was allowed to arise. Titles of nobility, bestowed as honours, were not hereditary.

The economic policies of the great Vietnamese dynasties also favoured the maintenance of imperial and Mandarin power. Through the 900 years of independence between Chinese domination and French colonial rule, the country's economy remained almost exclusively agricultural. Villages of artisans and fishermen existed, and there was some mining; but the mass of people were engaged in the cultivation of rice, and neither national nor international trade was systematically promoted. No property-owning middle class of merchants ever threatened the authority of the scholar Mandarins, and the periodically rising power of great landowners was diminished from time to time through the redistribution of land. Gia Long and his successor, Minh Mang, actually abolished all huge landholdings during the first half of the 19th century. Theoretically the emperor owned all land, and it was by imperial decree that the settlers on newly conquered territories received their plots in the villages that sprang up all the way south from the Red River Delta to the Mekong.

Vietnam's rigid absolutism was somewhat limited because of a Confucian concept, according to which the family was the basic unit of civilized society; submission to the authority of the family head thus was the foremost moral obligation of every citizen. The absolutism was also eased slightly by the limited authority granted to the village administration, whose purely local affairs were dealt with by a council of notables elected as a rule from the more prosperous or otherwise prominent citizens; among these notables' duties were the enforcement of law, the levying of army and forced labour recruits, and the assessment of taxes. Next to devotion to family, loyalty to the village was traditionally the first duty of every Vietnamese.

Apart from periods of dynastic decline and civil war, Vietnam was competently ruled under the Ly, Tran, and Le dynasties. Its administration was probably well in advance of any other native administration in Southeast Asia. But Mandarin rule also had its drawbacks: the static nature of the economy and a firm opposition to any innovation, technical, social, or cultural. The Mandarins and the scholarly rulers of Vietnam looked for enlightenment only to the past. Their hostility to change hardened in the early 17th century, when innovation appeared to them more and more to be a Western means to penetrate and undermine traditional Vietnamese society.

WESTERN PENETRATION INTO VIETNAM

Early Western contacts

As early as 1516 a number of Portuguese adventurers had inaugurated the modern era of Western penetration into Vietnam. They were followed in 1527 by visiting Dominican missionaries, and in 1535 by the first military man, Captain Antonio da Faria, who established a Portuguese trading centre at Faifo (Hoi An), some 15 miles south of present-day Da Nang. More Portuguese missionaries arrived during that century, and a mission was set up at Faifo in 1596, but it became permanent only after 1615, when Jesuits expelled from Japan were allowed to enter Vietnam. The Portuguese, Spanish, and Italian missionaries were joined in the same year by a French Jesuit, the famous Alexandre de Rhodes, who completed a transcription of the Vietnamese language into Latin characters and became the first French author on Vietnam.

From the very beginning, French Catholic missionary work developed hand in hand with French efforts to establish trade relations with Vietnam. Rhodes propagated the advantages of trade as fervently as he preached the need for greater missionary efforts, but to establish profitable trade with Vietnam was a difficult task, as the Portuguese, Dutch, and English had already learned. The Dutch had established a trading centre in Hanoi in 1637, the English in 1672, and both did poorly. Based on information supplied by the missionaries, the French East India Company and the Society of Foreign Missions, both founded in 1664, nevertheless pursued their efforts untiringly. A French trading centre was established in Hanoi in 1680, but it failed as dismally as did those of the English and the Dutch. By 1680 the war between the north and south had come to an end, and with it ended the interest of the two Vietnamese governments in the acquisition of modern weapons and other goods from the West. After 1700 only the Portuguese at the port of Faifo were able to maintain some trade with Vietnam.

The continued decline of Portuguese power in the Far East and the withdrawal of the Dutch and English from Vietnam encouraged renewed French efforts during the 18th century to establish trade with both halves of the country. They failed for two reasons: the first was the rising hostility of the Mandarins toward missionaries and Westerners generally; the second was the poverty of the people, as one persistent Frenchman, Pierre Poivre, who had visited Vietnam several times, pointed out in 1750. The Mandarins, realizing that their own moral authority was being undermined by the teachings of Christianity, began to persecute the missionaries who had converted about a tenth of the Vietnamese. The rulers of both Vietnams became more and more anti-Western, an attitude that subsequently convinced all Frenchmen who entered Vietnam that only through military intervention could Vietnam be made a base for French colonial expansion in the Far East. But France, unable after 1760 to defend its Indian possessions against the English, lacked the resources for military action in Vietnam.

Growth of French influence

The fortunes of the French took a turn for the better only toward the end of the 18th century, when another missionary—Pigneau de Béhaine, titular bishop of Adran—succeeded in making himself political and military adviser to Nguyen Anh, who, after 1778, was engaged in the struggle to wrest the country from the Tay Sons. The Bishop, after failing to persuade Paris to send an army in support of Nguyen Anh on the eve of the French Revolution, privately recruited sailors, soldiers, officers, and engineers from French overseas possessions, and their military expertise contributed decisively to the victory of Nguyen Anh over the Tay Sons. As Emperor Gia Long, Nguyen Anh remained grateful to the French for their help and kept many as advisers at his court (Pigneau de Béhaine died in 1799). But Gia Long did not favour the Christian religion, and under his strongly anti-Western successor Minh Mang (reigned 1820–1841), all French advisers were dismissed. Subsequent efforts by Paris to establish trade and diplomatic relations with Vietnam were politely rejected, and the persecution of the missionaries and their followers was intensified. Seven missionaries and an unknown number of Vietnamese Christians were executed between 1833 and 1838—some because they were involved in open rebellion against the rule of Minh Mang. After these events, France abandoned its efforts to get into Vietnam by diplomatic means. From 1840 on, French Catholic propaganda openly asked for military intervention in favour of the persecuted missionaries.

Under Minh Mang's successor, Thieu Tri, who reigned from 1841 to 1847, most missionaries were expelled; some were sentenced to death after returning clandestinely, but none was executed. Nevertheless, the French took reprisals against Vietnam by bombing the harbour of Tourane (Da Nang) in April 1847. But only 10 years later, when France under Napoleon III was ready to participate in the Western policy of colonial expansion, missionary propaganda, strongly supported by naval officers operating in the Far East, succeeded in persuading the government to prepare a military expedition against Vietnam.

THE CONQUEST OF VIETNAM BY FRANCE (1858–83)

The decision to invade Vietnam was made by Napoleon III in July 1857. It was the result not only of missionary propaganda but, after 1850, also of the upsurge of French capitalism, which generated the colonialist concept of a need for overseas markets and the collateral request for a larger French share in Asian territories conquered by the West. The naval commander in the Far East, Adm. Rigault de Genouilly, long an advocate of French military action against Vietnam, was ordered to attack the harbour and city of Tourane and to turn it into a French military base. The execution of his order, however, was delayed until the summer of 1858 by the Anglo-French operation of 1857–58 against China. Genouilly arrived at Tourane on August 31, 1858, with 14 vessels and 2,500 men; the French

French invasion o Vietnam

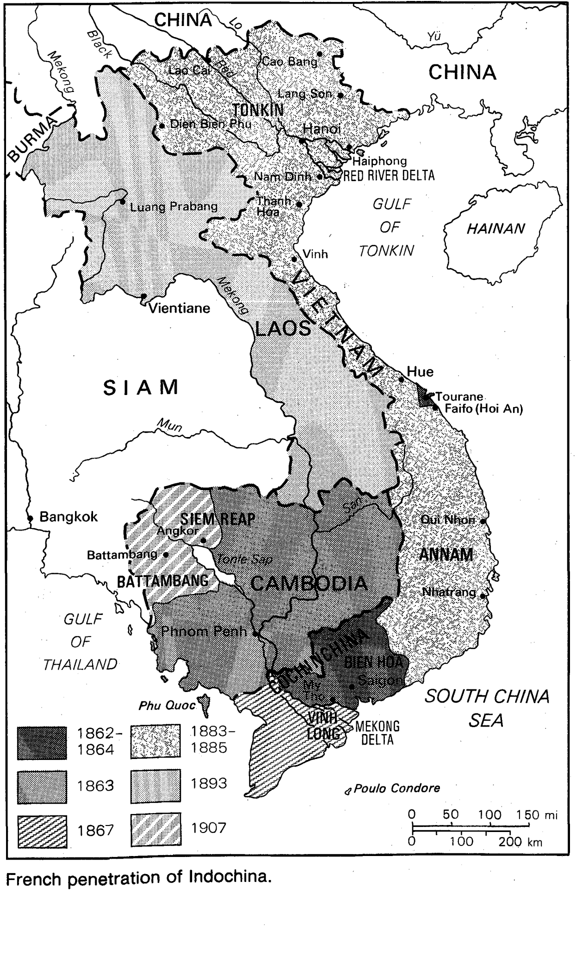
stormed the harbour defenses on September 1 and occupied the town a day later. But Genouilly made no progress on land beyond Tourane, and because he lacked the necessary shallow-draft vessels to go on the Perfume River from the ocean up to Hue, he was unable to threaten the Vietnamese capital. The collaboration of the Vietnamese Catholics with the invaders, which missionary propaganda had promised, failed to come about. Furthermore, French casualties from tropical diseases soon exceeded their battle dead, and when the rainy season started in October, Genouilly's army was completely immobilized.

Genouilly recognized that he could make no further progress around Tourane and decided to attack Saigon. Leaving a small garrison behind to hold Tourane, he sailed southward on February 2, 1859. Saigon was taken on February 17, but strong Vietnamese resistance prevented Genouilly from going beyond Saigon. There too the Catholic uprisings, which had been predicted by Genouilly's missionary advisers, failed to materialize. To save his troops threatened by Vietnamese attacks at Tourane, Genouilly returned there in April 1859, leaving a Franco-Spanish garrison at the fortress of Saigon. (The Spanish, whose Dominican missionaries were also persecuted in Vietnam, took part during the first phase of the invasion with troops from the Philippines, but they withdrew from it in 1862.) Again completely frustrated at Tourane, Genouilly resigned his command; his replacement, Admiral Page, arrived in October 1859 with orders to evacuate Tourane. The last French soldier left the city and harbour on March 22, 1860.

The second phase of the Anglo-French war against China once more tied down the Far Eastern fleet of the French during 1860, and they made no progress at all in their attempted conquest of Vietnam. Only in February 1861 was the garrison at the fortress of Saigon relieved by a new expedition under the command of Adm. Léonard Charner. By the end of June 1861, after receiving new reinforcements from the French operating in China, Charner had conquered Saigon and three adjacent provinces. The Mandarin administrators, though they refused to collaborate with the French, were unable to offer effective military resistance against an invader equipped with modern weapons. The Court of Hue, ruled since 1847 by Tu Duc, the last emperor of precolonial Vietnam, was forced by Charner's successor, Adm. Louis-Adolphe Bonard, to sign away Saigon and the three lost southern provinces in a peace treaty concluded with France in June 1862. In addition to Saigon and its environs, it conceded to the French the island of Poulo Condore (Con Son), three open ports for trade with Vietnam, full freedom of action for the missionaries, and a heavy financial indemnity. Tu Duc finally ratified this treaty in April 1863, in part because the French, with new reinforcements from Africa and Algeria, had suppressed the first Vietnamese uprising but also because Tu Duc needed his troops to fight a local rebellion against his rule, which had broken out in the north. Tu Duc was also helpless four years later, when the next French governor at Saigon, Adm. Pierre-Paul-Marie de la Grandière, under the pretext of forestalling an anti-French uprising, annexed the Vietnamese provinces west of the Mekong River and thus extended French rule over the whole of Cochinchina. (Cochinchina had been the name given to Vietnam by the first Westerners to visit the country, but it was used by the French only for the south. The north became known in the west as Tonkin, while Annam, under French rule, was the name used both for the centre of Vietnam and for the country as a whole. The people of Vietnam, generally referred to as Annamese or Annamites, never used these foreign designations. They called the north Bac Bo, the centre Trung Bo, and the south Nam Bo.)

French military and diplomatic successes

Adapted from Andre Masson, *Histoire du Vietnam* (1972), in series *Que Sais-je?*; Presses Universitaires de France



French penetration of Indochina.

The French had taken more than eight years to make themselves masters of Cochinchina (a protectorate had been imposed on Cambodia in 1863), and it would take them 16 more years to extend their control over Tonkin and Annam and complete their conquest of Vietnam. They made a first attempt to enter the Red River Delta in 1873 after a young explorer and adventurer named Francis Garnier had shown, in a hazardous expedition, that the Mekong River could not serve as a trade route for the French into southwestern China. Garnier, with the reluctant consent of the Court of Hue, was sent to Hanoi to arbitrate a conflict between the local authorities and a weapons-smuggling French trader commanding a band of mercenaries. He arrived there in November 1873 at the head of a small force supplied by Saigon; he sided with the French trader and, after receiving reinforcements from Saigon, stormed the citadel of Hanoi. In a short time Garnier extended his control, through the reckless use of artillery, over all of the important cities of the north. But when Garnier was killed in December 1873 in a battle with Chinese pirates hired by the Vietnamese authorities, the first attempt to conquer the north collapsed. France, scarcely recovered from a defeat suffered at the hands of Prussia in 1871, was in no mood to supply Saigon with the money and manpower needed to extend and maintain French control over Tonkin. The anticolonial faction in the Paris government contended that it was costly enough to control and develop Cochinchina and more important for France to rebuild her strength in Europe than to undertake doubtful colonial adventures.

Only after ten more years of rapid economic progress was France again ready to join the race of the Western powers for colonial expansion. In April 1882 the administration at Saigon, with the blessing of Paris, sent a force of 250 men to Hanoi under Capt. Henri Rivière; and when Rivière, in May 1883, met the same fate as Garnier, the French Chamber of Deputies immediately voted the necessary credits for the imposition by force of French control over Tonkin. A strong expeditionary corps moved into the Red River Delta in August 1883, while the French fleet bombarded Hue, where, unknown to the French, the emperor Tu Duc had died a few weeks earlier. With Tonkin occupied and Hue at the mercy of the French, the court Mandarins decided to submit. On August 25, 1883, they signed a treaty that made Tonkin and Annam protectorates of France, an action that turned out to be only another step toward the establishment of direct French rule over the whole of Vietnam.

Ten years later the French annexed Laos also. In the case of Cambodia, the French had laid claim to the rights that Vietnam had exercised over its weakened neighbour; Laos, although no longer controlled by Vietnam but by Siam (Thailand), was taken under the same pretext. It was added to the so-called Indochinese Union, which was created by the French in 1887 and consisted, after 1893, of Vietnam, Cambodia, and Laos.

COLONIAL VIETNAM (1883–1939)

The lack of a consistent policy, which characterized the period of conquest, continued to plague the French in Vietnam during the next decade and a half. Though it is true that Vietnamese armed resistance and Mandarin sabotage, particularly in Annam and Tonkin, impeded constructive economic and social measures, the chief reason for the lack of progress in any field was the absence of agreed-upon purposes and of a firm direction in running the colony. For years the French were unable to decide how to administer their new possessions and how to develop them economically.

Paul Doumer. All this changed quickly and drastically after the arrival in 1897 of Governor General Paul Doumer, who completed the pacification of the country, destroyed the old organization of the Vietnamese state, and established direct French rule on all levels of the administration. His autocratic and centralized regime left the Court of Hue and its Mandarins without a trace of real power. Emperors, as well as Mandarins who refused to be mere figureheads, were replaced by men willing to serve the French. The administration was staffed from top to bottom with officials imported from France; even as late as the 1930s, after several periods of reforms and concessions to local nationalist sentiment, Vietnamese officials were employed only in minor positions and at very low salaries, and the country was still administered along the lines laid down by Doumer.

Establishment of direct French rule

Doumer's economic and social policies also determined for the entire period of French rule the development of French Indochina, as the colony became known in the 20th century. The railroads, highways, harbours, bridges, canals, and other public works built by the French were almost all started under Doumer, whose aim was a rapid and systematic exploitation for the benefit of France of Indochina's potential wealth; Vietnam was to become a source of valuable raw materials and a market for tariff-protected goods of French industries. The exploitation of natural resources for direct export was the chief purpose of all French investments, with rice, coal, rare minerals, and later also rubber as the main products. Doumer and his successors up to the eve of World War II were not interested in promoting industry, the development of which was limited to the production of goods for immediate local consumption. Among these industries—located chiefly at Saigon, Hanoi, and Haiphong—were breweries, distilleries, small sugar refineries, rice and paper mills, and glass and cement factories; the greatest industrial establishment was a textile factory at Nam Dinh, which employed more than 5,000 workers. The total number of workers employed by all industries and mines in Vietnam was 100,000 in 1930. Because the aim of all investments was not a systematic economic development of the colony but the attainment of immediate high returns, only a small fraction of profits was reinvested.

Effects of French colonial rule. Whatever economic progress Vietnam made under the French after 1900 benefitted only the French and the small class of rich Vietnamese created by the colonial regime. The masses of the Vietnamese people were deprived of such benefits by the social policies inaugurated by Doumer and maintained even by his more liberal successors, such as Paul Beau (1902–07), Albert Sarraut (1911–14, 1917–19), and Alexandre Varenne (1925–28). Through the construction of irrigation works, chiefly in the Mekong Delta, the surface of rice land was quadrupled between 1880 and 1930; but the individual peasant's rice consumption decreased during the same period without compensation by other foods. The new lands were not distributed among the landless and the peasants, who had too little acreage; they were sold to the highest bidder or given away at nominal prices to Vietnamese collaborators and French speculators. These policies created a new class of Vietnamese landlords, together with a class of landless tenants who worked the fields of the landlords for rents of up to 60 percent of the crop, which was sold by the landlords at the Saigon export market. The mounting export figures for rice were due not only to the increase in cultivable land but no less to the mounting degree of peasant exploitation.

The peasants who owned their land were rarely better off than the landless tenants. The peasants' share of the price of rice at the Saigon export market was less than 25 percent. Peasants continually lost their land to the large owners because they were unable to repay loans given them by the landlords and other money lenders at exorbitant interest rates. As a result the large landowners of Cochinchina—a mere 2.5 percent of the total number of landowners—owned 45 percent of the land, while the small peasants—who accounted for 71 percent of the owners—owned no more than 15 percent of the land. The number of landless families in Vietnam before World War II was estimated at 50 percent of the population.

The peasants' share of the crop after the landlords, the money lenders, and the middlemen (mostly Chinese) between producer and exporter had taken their share, was still more drastically reduced by the direct and indirect taxes Doumer had imposed to finance his ambitious program of public works. The most hated of these taxes was that on salt, the production and sale of which became a state monopoly and the price of which rose 3,000 percent over its precolonial rate. Other ways of making the Vietnamese pay for the projects undertaken for the benefit of the French were the recruitment of forced labour for public works and the absence of any protection against exploitation in the mines and rubber plantations, although the scandalous working conditions, the low salaries, and the lack of medical care were frequently attacked in the French Chamber of Deputies. The mild social legislation decreed in the late 1920s was never adequately enforced.

These policies had a negative effect also on France. Although France accounted for almost 60 percent of all Indochinese imports between 1920 and 1939, Indochina never became an important market for French industry. The absence of a really developmental economic policy prevented any significant increase in the wealth of Vietnam, and the social policies of the colonial regime perpetuated the poverty of the masses, of whom more than 80 percent remained peasants. Only through a more enlightened economic and social policy could they have become consumers of French goods. The French themselves and the 6,000–7,000 large Vietnamese landowners were the only people in the market for most goods imported from France.

Effects on France of policies in Vietnam

The statistics of the French seem to contradict their own claim that much was done for the Vietnamese in the areas of medical care and education. In 1939 no more than 15 percent of all school-age children received any kind of schooling, and about 80 percent of the population was illiterate, in contrast to precolonial times when the majority of the people possessed some degree of literacy. With its more than 20,000,000 inhabitants in 1939, Vietnam had one university, with fewer than 700 students. Only a small number of Vietnamese children were admitted to the lycées (secondary schools) for the children of the French.

Medical care was well organized for the French in the cities, but in 1939 there were only two physicians for every 100,000 Vietnamese, compared with 76 in Japan and 25 in the Philippines.

Two other aspects of French colonial policy are important in considering the attitude of the Vietnamese people, especially their educated minority, toward the colonial regime: one was the absence of any kind of civil liberties for the native population; the other, the exclusion of the Vietnamese from the modern sector of the economy, especially industry and trade. Not only rubber plantations, mines, and industrial enterprises were in foreign hands—French, where the business was substantial, and Chinese on the lower levels—but all other business too, from local trade to the great export–import houses. The social consequence of this policy was that, apart from the landlords, no property-owning native middle class developed in colonial Vietnam. Thus capitalism appeared to the Vietnamese to be a product of foreign rule, a fact that, together with the lack of democracy, profoundly influenced the nature and orientation of the national resistance movements.

THE MOVEMENTS OF NATIONAL LIBERATION

arly
esistance
French
ule

The anticolonial movement of the Vietnamese people can be said to have started with the establishment of French rule. Most local Mandarins of Cochinchina refused to collaborate with the French; instead they led guerrillas, composed of the remnants of the defeated armies, in attacks on French outposts and early in 1863 in a general uprising that the French were able to suppress within a few weeks. A much broader movement of resistance developed in Annam in 1885, led by Mandarins of the court; it was not defeated until 1888. A few years later armed resistance was resumed in Annam under the leadership of the great scholar Phan Dinh Phung, whose rebellion collapsed only after his death in 1895. Equally troublesome for the colonial administration was a chain of uprisings in Tonkin, the suppression of which took the French 14 years (from 1883 to 1897).

Apart from its local nature, the main characteristic of the national movement during this first phase of resistance was its political orientation toward the past. Its leaders were Mandarins and scholars. Filled with ideas of precolonial Vietnam, they wanted to be rid of the French in order to re-establish the old imperial and Mandarin order. Because this aspiration could have little meaning for the generation that grew to maturity after 1900, the national movement led by the royalist Mandarins, after being defeated militarily, was also dead politically.

new
ational
ovement

Phan Boi Chau. A new national movement arose soon after the turn of the century. Its most prominent spokesman was Phan Boi Chau, with whose rise the old traditionalist opposition gave way to a modern nationalist leadership that rejected French rule but not Western ideas, science, and technology. Although still a monarchist at the outset, Chau was a modern thinker for his time; his aim was a new Vietnam under a progressive emperor. He chose Prince Cuong De, a direct descendent of Emperor Gia Long, as his candidate.

In 1905 Chau went to Japan, and Prince Cuong De followed him in 1906. Abroad they founded the Association for the Modernization of Vietnam, thus reviving their country's name, of which the French had deprived them. Chau's plan, mildly encouraged by some Japanese statesmen, was to free Vietnam with Japanese help; he smuggled hundreds of young Vietnamese into Japan, where they studied sciences and underwent training for clandestine organization, political propaganda, and terrorist action. Inspired by Chau's effective writings, nationalist intellectuals in Hanoi opened the Free School of Tonkin in 1907, which soon became a centre of anti-French agitation and was consequently suppressed after a few months. Also under the inspiration and guidance of Chau's followers, mass demonstrations demanding a reduction of high taxes took place in many cities in 1908. As a symbol of their rejection of the past, a movement to cut off the men's traditional long hair was started and spread quickly over the whole country. Hundreds of demonstrators and suspected organizers were arrested—some were condemned to death, others sent to the island of Poulo Condore, which the French turned into a concentration camp for Vietnamese nationalists.

Chau was forced to abandon the hope that Vietnam could be freed with Japanese help when, after Japan had received a loan from France, he and Prince Cuong De were expelled from Japan in 1910. They went to China, where a revolution had broken out in Canton. There in 1912 they set up a republican government in exile, with Prince Cuong De as president. Chau also united a number of nationalist groups in the League for the Restoration of Vietnam, but a temporary setback in the Chinese Revolution enabled the French to have Chau thrown into prison in China. He was freed only in 1917 after the triumph of the revolution in southern China. Chau continued to work from exile, but his movement declined rapidly after World War I. In 1925 he was kidnapped by French agents in Shanghai, taken to Vietnam, and condemned to death; the liberal governor general Varenne changed the sentence to life confinement in his house in Hue. There Chau died in 1940.

During the years between the outbreak of World War I and the arrest of Chau, the movement for national liberation, disorganized and persecuted by the colonial regime, nevertheless manifested its existence in a variety of terrorist acts, demonstrations, and local revolts. In 1916, the 18-year-old emperor Duy Tan led an unsuccessful revolt, whose participants were either executed or deported; the Emperor himself was exiled to Reunion Island. In 1917 the indigenous garrison at Thai Nguyen staged an uprising and took the town, but it was quickly defeated by French troops from Hanoi. After 1920 a number of prominent intellectuals, among them Pham Quynh, pursued the hope of obtaining political concessions from the colonial regime through collaboration with the French. In the south, Bui Quang Chieu, who was one of the first Vietnamese to become a French citizen, was allowed to found the Constitutionalist Party of Cochinchina. But the total failure of these "reformist" efforts led to a revival of clandestine and revolutionary groups, especially in Annam and Tonkin; among these were the Vietnam Revolutionary Party, founded in Annam in 1925, and the more important Vietnamese Nationalist Party (Viet Nam Quoc Dan Dang, founded in 1927 and usually referred to as the VNQDD), whose leader, Nguyen Thai Hoc, was a 23-year-old teacher. The VNQDD preached terrorist action and penetrated the garrisons of indigenous troops with a plan to oust the French in a military uprising. On the crucial night of February 9–10, 1930, the troops of only one garrison, Yen Bai in Tonkin, killed their French officers, but they were overwhelmed a day later and summarily executed. A wave of repression followed that took hundreds of lives and sent thousands into concentration camps; Nguyen Thai Hoc and 12 of his collaborators were arrested and beheaded. The VNQDD was virtually destroyed, and for the next 15 years it existed mainly as a group of exiles in China supported by the Kuomintang (Chinese Nationalist Party).

Ho Chi Minh. For yet another reason, the year 1930 was an important one in the history of the national liberation movement. In 1925, the year of Chau's arrest, a new figure, destined to become the most prominent leader of the national revolution, had appeared on the scene as an exile in Canton. He was Nguyen Ai Quoc, better known as Ho Chi Minh, the name he adopted in 1943. In June 1925 Ho Chi Minh founded the Revolutionary League of the Youth of Vietnam, which became the nucleus of the Vietnamese Communist Party.

Ho Chi Minh had left Vietnam as a young seaman in 1912 and travelled widely before settling in Paris in 1917; he joined the Communist Party of France in 1920 and later spent several years in Moscow and China in the service of the international Communist movement. After making his Revolutionary League the most influential of all clandestine resistance groups, he succeeded in 1930 in forming the Indochinese Communist Party from a number of competing Communist organizations. In May 1930 the Communists exploited the existence of near starvation conditions over large regions of central Vietnam by staging a broad peasant uprising, during which numerous Vietnamese officials as well as many landlords were killed and "Soviet" administrations were set up in several provinces of Annam. It took the French until the spring of 1931 to suppress this movement and to re-establish, in an unparalleled wave of terror, their own control.

Unlike the dispersed and disoriented leadership of the VNQDD and some smaller nationalist groups, the Communist Party, relying on hundreds of cadres trained in Russia and China, recovered quickly from the setback of the entire liberation movement in 1930–1931. After 1936, when the French Popular Front government extended some political freedoms to the colonies, the clandestine apparatus of the Communist Party skillfully exploited all opportunities for the creation of legal front organizations, through which the party's influence on intellectuals, workers, and peasants was increased. In the south the Communists had to compete with a strong Trotskyite faction, and the spread of their influence was somewhat checked in regions where religiopolitical sects dominated—the Cao Dai, founded in 1926, and the Hoa Hao, which arose in 1939. But when all political freedoms were again suspended at the outbreak of World War II, the Communist Party, unified and in command of a well indoctrinated and disciplined following, was no doubt the most effective faction in the entire movement of national liberation.

WORLD WAR II AND INDEPENDENCE (1940–45)

Wartime liberation activities

For five years during World War II, Indochina was a French-administered possession of Japan. On September 22, 1940, Adm. Jean Decoux, the French governor general appointed by the Vichy government after the fall of France, concluded an agreement with the Japanese that permitted the stationing of 30,000 Japanese troops in Indochina and the use of all major Vietnamese airports by the Japanese Army. The agreement made Indochina the most important staging area for all Japanese military operations in Southeast Asia. The French administration cooperated with the Japanese occupation forces and was ousted only toward the end of the war (in March 1945), when the Japanese began to fear that the French forces might turn against them at the hour of approaching defeat. After the French had been disarmed, Bao Dai, the last French-appointed emperor of Vietnam, was allowed to proclaim the independence of his country and to appoint a Vietnamese national government at Hue, but all real power remained in the hands of the Japanese Army of occupation.

Apart from Cochinchina, where the religiopolitical sects and a few other small groups maintained themselves under Japanese protection, the movement of national liberation during these years was confined largely to the activities of the Vietnamese exiles in southern China. There, in May 1941, Ho Chi Minh succeeded in forming the Viet Nam Doc Lap Dong Minh Hoi, known subsequently as the Viet Minh, which was a united front of several nationalist groups under Communist direction. Under Nationalist Chinese pressure, a counterfront was organized under the name of Dong Minh Hoi ("Vietnam Revolutionary League"). On orders of Chiang Kai-shek, Ho Chi Minh was imprisoned in 1942. But the anti-Communist leadership of the Dong Minh Hoi proved incompetent, and the Allies—Chinese as well as U.S.—who needed intelligence about the Japanese in Vietnam and help in rescuing downed pilots, began more and more to rely on the Viet Minh. Under the direction of Vo Nguyen Giap, the Viet Minh organized a tight net of political agents and groups of guerrillas in northern Tonkin and thus made themselves useful to the Allies. Ho Chi Minh was released in 1943, made head of the Dong Minh Hoi, and given financial assistance and even some weapons from the United States OSS (Office of Strategic Services) for the anti-Japanese activities of the Viet Minh inside Vietnam. Ho Chi Minh used his position to advance the cause of the Viet Minh; he himself entered Vietnam in October 1944. When the Japanese surrendered in August 1945, the Communist-led Viet Minh ordered a general uprising. As the only group whose leaders were in the country and heading a well-organized political and military movement, the Viet Minh was able in the second half of August 1945 to take power in Hanoi without meeting any significant opposition. The Bao Dai-appointed government at Hue resigned on August 22, and the Emperor himself abdicated in favour of the government set up under Ho Chi Minh in Hanoi. A Provisional Executive Committee of South Vietnam, installed in Saigon and dominated by the Viet Minh, placed itself under the authority of the Hanoi government on August 25. When Ho Chi Minh proclaimed the independence of Vietnam on September 2, it was evident that in the struggle for power among the various groups in the movement for national liberation, the Communists had gained the upper hand.

The French—whose claim to reoccupy Indochina had been confirmed by the Potsdam Conference of the victorious Allies—rather than Vietnamese opponents prevented the Viet Minh from establishing themselves firmly in the whole country. But in denying to Ho Chi Minh's government the control of Vietnam, the French denied to all national liberation movement groups their common goal—the independence of Vietnam.

French oppositio

The Potsdam agreement provided that the British disarm the Japanese in the south of Vietnam and the Chinese disarm them in the north. The Chinese under Chiang Kai-shek did not favour the return of the French to Indochina, and Chinese occupation forces, after entering Tonkin in September 1945, therefore refrained from interfering with the established government. The British, who arrived in the south on September 12 after freeing French soldiers interned by the Japanese, not only armed them but also supported them in their struggle to depose the local Vietnamese administration set up by the Committee of the South. The French reconquest of Vietnam, which started on September 23, drove the Vietnamese—Communists and anti-Communists—into armed resistance. With the arrival of reinforcements at the end of September, the French broke this resistance after several weeks of fighting in and around Saigon. At the beginning of 1946 the French controlled the cities and highways of the south, but armed guerrilla resistance, morally and materially supported by the Hanoi government, continued all over Cochinchina and southern Annam. (J.Bu.)

THE FIRST INDOCHINA WAR (1946–54)

Negotiations between the French and Ho Chi Minh led to an agreement in March 1946 that appeared to promise a peaceful solution. Under the agreement France would recognize the Viet Minh government and give Vietnam the status of a free state within the French Union. French troops were to remain in Vietnam, but they would be withdrawn progressively over five years. For a period in early 1946 the French cooperated with Ho Chi Minh as he consolidated the Viet Minh's dominance over other nationalist groups, in particular those politicians who were backed by the Chinese Nationalists.

Despite tactical cooperation between the French and the Viet Minh, their policies were irreconcilable; the French aimed to re-establish colonial rule, while Hanoi wanted total independence. French intentions were revealed in the decision of Adm. Georges-Thierry d'Argenlieu, the high commissioner for Indochina, to proclaim Cochinchina an autonomous republic in June 1946. Further negotiations did not resolve the basic differences between the French and the Viet Minh. The First Indochina War began with the French bombardment of Haiphong on November 23, 1946, which caused at least 6,000 civilian casualties, and the subsequent Viet Minh attempt to overwhelm French troops in Hanoi on December 19.

Outbreak of war

Between 1946 and 1951 the Viet Minh waged an increasingly successful guerrilla war, aided after 1949 by the new Communist government of China. Despite large amounts of U.S. financial and military aid, the French were on the defensive by 1952.

Initially confident of victory, the French long ignored the real political cause of the war—the desire of the Vietnamese people, including their anti-Communist leaders, to achieve unity and independence for their country. French efforts to deal with this problem were devious and ineffective. Reuniting Cochinchina with the rest of Vietnam in 1949, they appointed a former emperor, Bao Dai, as chief of state. Anti-Communist nationalists, among them the Roman Catholic leader Ngo Dinh Diem, denounced these maneuvers, and leadership of the struggle for independence from the French remained with Ho Chi Minh.

By the end of 1953 the Viet Minh controlled much of Vietnam and neighbouring Laos. The deteriorating French military position became dramatically worse in March 1954, when their strong garrison at Dien Bien Phu was besieged by Viet Minh forces under Gen. Vo Nguyen Giap. A mounting demand in France for an end to the war in Indochina and the refusal of the U.S. Congress to intervene to save the French at Dien Bien Phu, which fell on May 7, 1954, led to French readiness to negotiate to end the war at an international conference already in session in Geneva.

THE TWO VIETNAMS (1954–65)

An agreement concluded at Geneva on July 21, 1954, and signed by French and Viet Minh representatives provided for a cease-fire and for a temporary division of the country at the 17th parallel of latitude into two military zones. All Viet Minh forces were to withdraw north of the 17th parallel, and all French and State of Vietnam troops south of the parallel; permission was granted for refugees to move from one zone into the other within a given time limit. An international commission was established, composed of Canadian, Polish, and Indian members under an Indian chairman, to supervise the execution of the agreement.

ɪrth
ɜtnam

This agreement left the Hanoi regime—recognized since 1950 by China and the Soviet bloc governments as the Democratic Republic of Vietnam—in control of only the northern half of the country. But the Viet Minh leaders had been convinced before they went to Geneva that total victory was within reach, and they agreed to this division only because they regarded it as provisional. Most of the conference participants—which included, in addition to the French and the Indochinese, delegations of the United States, the United Kingdom, the Soviet Union, and the People's Republic of China—approved a Final Declaration, which provided for all-Vietnamese elections supervised by the commission in July 1956 to unify the country; Viet Minh leaders appeared certain to win these elections. The U.S. and South Vietnam would not approve the unsigned Final Declaration.

In the meantime, reconstruction of war-ravaged North Vietnam was begun with great energy, and great sacrifices were imposed on the population. Assistance from China, the Soviet Union, and other Communist states enabled the Hanoi regime to embark on an ambitious program of industrialization. Except for some local revolts in 1956 against a brutally conducted campaign to collectivize agriculture, no serious social or political obstacles interfered with the north's steady economic progress between 1955 and 1965.

uth
ɜtnam

In contrast to the north, where the leaders in 1965 were those who had made the revolution 20 years earlier, South Vietnam went through several periods of turbulent political changes during the same span of time. On June 16, 1954, Emperor Bao Dai called upon the Roman Catholic leader Ngo Dinh Diem to form a new government. Diem succeeded against tremendous odds in mastering the existing political chaos in the south and, with political and profuse financial aid from the United States, in stabilizing his anti-Communist regime. He eliminated the pro-French leadership of the army and abolished the local autonomy of the religio-political sects. He also resettled an estimated 900,000 refugees from the north. In a government-controlled referendum in October 1955 that removed Bao Dai as chief of state, Diem made himself president of the Republic of Vietnam.

Diem's early success did not continue. Plans for land reform were sabotaged by entrenched interests. With U.S. financial backing the regime's chief energies were directed toward building up the army and a variety of intelligence and security forces to counter the still influential Viet Minh. Totalitarian methods were directed against all who were regarded as opponents, and the favouritism shown to Roman Catholics alienated the Buddhist population. Loyalty to the President and his family was made a paramount duty and Diem's brother, Ngo Dinh Nhu, founded a secret party to spy on officials, army officers, and prominent private citizens. Under these conditions, a Communist-led insurgency began in 1957, a year after the constituent assembly in Saigon reiterated Diem's refusal to participate in the all-Vietnamese elections described in the Final Declaration in Geneva. The insurrection appeared close to success when Diem's army overthrew him on November 1, 1963. Diem and his brother Nhu were killed following the coup, and there were nine changes of government before the military firmly seized control in June 1965 under Air Vice Marshal Nguyen Cao Ky.

THE SECOND INDOCHINA WAR

The military regime that brought an end to the period of rapid changes of government in Saigon was not noticeably different from the dictatorship of Ngo Dinh Diem. The militant Buddhists who had helped overthrow Diem strongly opposed Ky's government; but he was able to break their resistance, centred at Hue and Da Nang, in late spring 1966. Civil liberties were restricted; political opponents—denounced as "neutralists" and as helpful to the Communists—were imprisoned, usually without trial; and political parties were allowed to operate only if they did not openly criticize government policy.

The
military
regime

This character of the regime remained largely unchanged after the presidential elections on September 3, 1967, which the candidates of the military forces—Gen. Nguyen Van Thieu and his running mate Ky—won with only 35 percent of the vote (the opposition, whose 65 percent of the vote was split among 10 civilian candidates, charged that the victory was fraudulently obtained).

No less evident than the oppressive nature of the Thieu-Ky regime was its inability to cope with the Communist-led insurrection, which appeared to be even closer to victory in 1965 than it had been before the fall of Diem. In December 1960 the various groups engaged in the armed struggle against the Saigon regime had formed, under Communist direction, the National Front for the Liberation of the South (NFL; also called National Liberation Front, NLF), whose program—neutralization of South Vietnam, withdrawal of all foreign troops, and gradual reunification with the north—was endorsed by Hanoi in January 1961. With an openly Communist People's Revolutionary Party, founded in 1962, the Communists maintained their dominant position within the NFL. A steady infiltration of weapons and military experts from the north and effective recruitment in the south rapidly increased the number of NFL fighters, commonly called Viet Cong. The fighting strength of the Viet Cong grew from about 30,000 men in 1963 to about 150,000 in 1965 when, in the opinion of many U.S. military experts, the survival of the Saigon regime was most seriously threatened.

American-
ization
of the
war

American involvement in Vietnam. This danger set in motion early in 1965 the process described as the Americanization of the Vietnamese War. Until 1960 the United States had supported the Saigon regime and its army only with military equipment, financial aid, and, as permitted by the Geneva agreement, 700 advisers for the training of the army. The number of advisers had increased to 17,000 by the end of 1963, and they were joined by a rising number of U.S. helicopter pilots. All this assistance, however, proved insufficient to halt the advance of the Viet Cong, and at this critical point, Pres. Lyndon B. Johnson of the U.S., in February 1965, ordered the bombing of North Vietnam, hoping to prevent further infiltration of arms and men into the south. The bombing failed to stop Hanoi's support for the Viet Cong. On the contrary, according to U.S. military intelligence, the Hanoi government, four weeks after the start of the bombing, began to infiltrate its first regular army units, a battalion of about 400 to 500 men.

Four weeks after the systematic bombing of the north had begun, the United States sent its first combat troops to Vietnam—3,500 U.S. marines landed at Da Nang on March 7, 1965. By July 1965 the number of U.S. combat troops had reached 75,000; it continued to climb from month to month until it stood at more than 510,000 early in 1968. The Americans fought at the side of some 600,000 regular Vietnamese troops and several hundred thousand regional and local defense forces. (They were also joined by nearly 50,000 South Koreans and smaller numbers of troops from Thailand, New Zealand, and Australia.) The number of regular North Vietnamese army troops in the south was said to have been 50,000 at this time, and all Viet Cong combatants numbered about 230,000. (All these figures, as well as those of enemy casualties, are estimates thought by most observers to be highly unreliable.)

Nearly three years of intensive bombing in the north and of fighting in the south with the enormous quantity of modern equipment the U.S. was able to supply had devastated vast regions of Vietnam, both north and south, but seemed unable to weaken the will and strength of the NFL and North Vietnam. This situation became evident in the so-called Tet Offensive of February 1968, during which the Viet Cong and North Vietnamese attacked more than 100 cities and military bases, holding on to some for several weeks. After that, a growing conviction in Washington that a military victory was impossible and that continuation of the war at the current levels was no longer politically acceptable led President Johnson to reject the request of his commander in Vietnam, Gen. William Westmoreland, for an additional 206,000 U.S. soldiers. Instead, the President ordered a restriction of the bombing in the north, and this decision opened the way for U.S. negotiations with Hanoi, which started in Paris on May 13, 1968. After the bombing was halted over the entire north in November 1968, the Paris talks were enlarged to include representatives of the NFL and the Saigon regime.

Negotiations and U.S. troop withdrawals

Although a gradual withdrawal of U.S. troops was begun by Pres. Richard M. Nixon, the war, with United States participation, widened in Laos and, in April 1970, was taken by the U.S. into Cambodia, thus becoming again a truly Indochinese war. By April 1970, 115,000 U.S. troops had been withdrawn, and Nixon announced a plan to withdraw another 150,000 over the next 12 months. The withdrawal of all United States troops was made contingent, however, on the so-called Vietnamization of the war—the ability of the Saigon regime and its armed forces to contain the fighting, while peace talks went on in Paris and renewed U.S. bombing on a scale unprecedented in history continued all over Indochina.

Withdrawal of American troops. Finally, in January 1973, a cease-fire agreement was signed by the United States and all three Vietnamese parties. It provided for complete withdrawal of U.S. troops within 60 days and gave the Provisional Revolutionary Government (formed in 1969 by the NFL and other groups opposed to the regime) equal recognition with Saigon. (J.Bu./M.E.O.)

The signature of the Paris agreements did not bring an end to the fighting in Vietnam, however. The Saigon regime made a determined attempt to eliminate Communist power in the Mekong delta, and by late 1973 the northern leadership decided to resolve the situation through military means. Fierce fighting took place during 1974, and the Communists made important advances in the highland region of southern Vietnam. In early 1975 the Saigon regime still had control of the coastal littoral from north of Hue to the Mekong delta, but the strategic balance soon tipped sharply against the Saigon forces. The fall of Ban Me Thuot in the highlands was followed by the loss of Hue and Da Nang in March. Xuan Loc fell to the Communists on April 21, and the Communists entered the southern capital on April 30. The Second Indochina War was finally at an end.

REUNIFICATION

Following the Communist victory, Vietnam remained theoretically divided, with the Provisional Revolutionary Government responsible for the territory of southern Vietnam. After national elections in April 1976, formal reunification took place, and the actual situation, in which power lay with the Vietnamese Communist Party and the government in Hanoi, was ratified. The Socialist Republic of Vietnam was officially proclaimed on July 2, 1976.

Problems of peace

Vietnam at peace faced formidable problems. In the south alone, 57 percent of the population had been made homeless between 1965 and 1974, and more than 16 percent of the population had been killed or wounded during the same period. The costs in the north were probably higher. Plans to reconstruct the country following the devastation of the long years of war called for the expansion of industry in the north and of agriculture in the south. Within two years of the Communist victory, however, it became clear that Vietnam faced major difficulties in achieving its goals.

With only limited personnel trained to deal with peacetime problems, the Vietnamese government encountered considerable resistance to its policies, particularly in the vast urban concentration of Saigon (renamed Ho Chi Minh City in 1976). The former commercial community, the members of which were mostly ethnic Chinese, sought to avoid cooperating in the implementation of new socialist economic measures, and many ethnic Vietnamese city-dwellers were reluctant to become pioneer agricultural workers in "new economic zones." Vietnam also suffered major floods and drought that severely decreased rice production. One result of these circumstances was the flight of 545,000 refugees from both southern and northern Vietnam by 1980. These refugees were in addition to the 135,000 who had fled at the time of the fall of Saigon in 1975.

Internal difficulties were compounded by Vietnam's problems in the field of foreign affairs. Perhaps unrealistically, Vietnam had hoped to gain U.S. economic aid following the end of the Second Indochina War. When such aid was not forthcoming, and in circumstances of acute economic distress, Vietnam turned to the Soviet Union and developed an even closer relationship with that country. In June 1978 Vietnam became a member of the Council for Mutual Economic Assistance (Comecon), the east European Communist economic bloc dominated by the Soviet Union. Subsequently, in November 1978, Vietnam signed a 25-year treaty of friendship and cooperation with the Soviet Union. This strengthening of both political and economic ties with the Soviet Union took place at a time when Vietnam's relations with the People's Republic of China and Cambodia (Kampuchea) were rapidly deteriorating. Savage fighting between Vietnam and Cambodia grew more intense through 1978 and concurrently Vietnam and China traded charges of subversion and armed aggression. It was in these circumstances that the first significant exodus of ethnic Chinese refugees from northern Vietnam into China took place.

Invasion of Kampuchea

Vietnam invaded Kampuchea at the end of 1978. Following a short campaign, Vietnam overcame resistance from the regime of Pol Pot in the central Cambodian region and installed a new protégé regime in Phnom Penh in early January 1979. Resistance to the Vietnamese and their Cambodian protégés continued on a significant scale, however, and it was not until the middle of 1979 that the remnants of Pol Pot's forces were driven back to remote regions in the northern and western parts of Cambodia.

China's invasion o
Vietnam

With Vietnam heavily engaged in Cambodia, where some 200,000 Vietnamese troops had been committed, China embarked on a month-long punitive invasion of northern Vietnam beginning in mid-February 1979. This invasion to "teach the Vietnamese a lesson" for their actions in Cambodia involved 100,000 troops, some of whom penetrated as deeply as 30 miles inside Vietnamese territory. In 17 days of campaigning, the Chinese destroyed major Vietnamese towns and inflicted heavy damage on Vietnam's economic infrastructure. The invasion was not a one-sided affair, however. Although taking heavy casualties themselves, the Vietnamese were able to inflict substantial losses on the Chinese.

Events following the Vietnamese invasion of Cambodia and the subsequent Chinese invasion of Vietnam confirmed Vietnam's increasing isolation in Southeast Asia. Apart from the protégé regime in Phnom Penh and the government of Laos, which was also heavily dependent on Vietnamese assistance for its survival, Vietnam was at odds with the remainder of its regional neighbours. Condemnation of Vietnam's presence in Cambodia by the members of the Association of Southeast Asian Nations (ASEAN) stood in the way of Vietnamese efforts to gain international recognition for its regime in Phnom Penh. Peace talks between Vietnam and China aimed at settling the major differences between those two countries were unproductive. (M.E.O.)

Vietnam also experienced continuing major social and economic difficulties. Following intense international pressure to prevent a further outflow of refugees, Vietnam agreed to curtail refugee departures in July 1979. But the continuing number of persons leaving Vietnam during the 1980s, both ethnic Vietnamese and Chinese, underlined the harsh nature of prevailing economic conditions. The cost of maintaining troops in Cambodia and Laos and of keeping combat-ready troops along the Sino-Vietnamese border was particularly heavy. In addition, there were continuing difficulties in integrating the southern part of the country into a socialist economy. Internal security was not seriously threatened by the continued existence of anti-government groups. There was, however, sufficient resistance from discontented minorities, particularly in the central highlands, to add a further burden to the Hanoi government. During 1986 the death of longtime Vietnamese Communist Party boss Le Duan and the resignation of several other top party leaders constituted a major changing of the political guard. (M.E.O./Ed.)

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

BIBLIOGRAPHY

Kampuchea: L.P. BRIGGS, *The Ancient Khmer Empire* (1951), the most comprehensive work in English on the subject, drawn from all the major French scholars; JEAN DELVERT, *Le Paysan cambodgien* (1961), a classic study of Cambodian geography and rural life; DAVID P. CHANDLER, *The Land and People of Cambodia* (1972), a historical treatment aimed at high school students; CLAUDE-GILLES GOUR, *Institutions constitutionnelles et politiques du Cambodge* (1965), an intensive analysis of Cambodia's constitutional institutions; FRANK M. LEBAR (ed.), *Ethnic Groups of Mainland Southeast Asia* (1964), one of the best sources of information about the country's many tribal groups; CHARLES MEYER (ed.), *Cambodge* (1962), a beautifully illustrated volume devoted to all aspects of life in Cambodia; JACQUES MIGOZZI, *Cambodge: Faits et problèmes de population* (1973), a detailed analysis of demographic trends; MILTON OSBORNE, *Politics and Power in Cambodia* (1973), an analysis of Cambodia in the 1950s and 1960s; FRANCOIS PONCHAUD, *Cambodge année zéro* (1977; *Cambodia: Year Zero,* 1978), an account of Democratic Kampuchea; REMY PRUD'HOMME, *L'Economie du Cambodge* (1969), a penetrating analysis of the economy; WILLIAM SHAWCROSS, *Sideshow: Kissinger, Nixon, and the Destruction of Cambodia* (1979), an analysis of U.S. policy toward Cambodia during the Khmer Republic period; D.J. STEINBERG *et al., Cambodia: Its People, Its Society, Its Culture,* rev. ed. (1959), a compendium of information drawn from numerous published sources, somewhat outdated, but still useful; W.E. WILLMOTT, *The Chinese in Cambodia* (1967), one of the rare books devoted to the Chinese in Cambodia and their relations with the Khmer. Much of the important material concerning Cambodian history until the fall of Angkor is in journal articles, mostly in French. Among the more valuable and readily available sources are: G. COEDES, *Les États hindouisés d'Indochine et d'Indonésie,* rev. ed. (1964; Eng. trans., *The Indianized States of Southeast Asia,* 1968), *Les Peuples de la Péninsule Indochinoise* (1962; Eng. trans., *The Making of South East Asia,* 1966), and *Pour mieux comprendre Angkor* (1943; Eng. trans., *Angkor: An Introduction,* 1963), three studies by the most distinguished French scholar of Cambodia to the end of the Angkorian period; J. BOISSELIER, *Le Cambodge* (1966), the most complete archaeological survey of the end of the Angkorian period; B.-P. GROSLIER and J. ARTHAUD, *Angkor: Hommes et pierres,* rev. ed. (1965; Eng. trans., *Angkor: Art and Civilization,* 1966), a fine pictorial study with concise text; M. GITEAU, *Les Khmers* (1965; Eng. trans., *Khmer Sculpture and the Angkor Civilization,* 1965), by a former curator of the Phnom Penh National Museum; L.P. BRIGGS, *The Ancient Khmer Empire* (1951), a detailed survey of Cambodian history to the fall of Angkor, now in need of revision; O.W. WOLTERS, "The Khmer King at Basan (1371–3) and the Restoration of the Cambodian Chronology during the Fourteenth and Fifteenth Centuries," *Asia Major,* 12:44–89 (1966), an important journal article suggesting a revision of the date of the fall of Angkor; and D.G.E. HALL, *A History of South-East Asia,* 3rd ed. (1968), for a summary of Cambodian history. B.-P. GROSLIER, *Angkor et le Cambodge au XVI^e^ siècle d'après les sources portugaises et espagnoles* (1958), discusses the economic basis of Angkorian times and the "Spanish interlude"; D.P. CHANDLER, *Cambodia Before the French; Politics in a Tributary Kingdom, 1794–1848* (1974), a valuable study; J. MOURA, *Le Royaume du Cambodge,* 2 vol. (1883), an early French study; A. LECLERE, *Histoire du Cambodge*... (1914), a monumental but flawed pioneering history; M.E. OSBORNE, *The French Presence in Cochinchina and Cambodia: Rule and Response (1859–1905)* (1969), treats the first 50 years of French rule; D.J. STEINBERG *et al., Cambodia: Its People, Its Society, Its Culture,* rev. ed. (1959), a useful but dated survey; R.M. SMITH, "Cambodia," in G.M. KAHIN (ed.), *Governments and Politics of Southeast Asia,* 2nd ed. (1964); *Cambodia's Foreign Policy* (1965), a sympathetic study of Sihanouk's policies; M. LEIFER, *Cambodia: The Search for Security* (1967), critically evaluates foreign policy issues; W.E. WILLMOTT, *The Chinese in Cambodia* (1967), a valuable study of the Chinese minority in Cambodia; M.E. OSBORNE, *Politics and Power in Cambodia; The Sihanouk Years* (1973); and NORODOM SIHANOUK, *War and Hope* (1980), a plan to rescue Cambodia from its invaders.

Laos: RENE DE BERVAL *et al., Présence du royaume Laos, pays du million d'éléphants et du parasol blanc* (1956; Eng. trans., *Kingdom of Laos: The Land of the Million Elephants and of the White Parasol,* 1959), a general symposium by French and Laotian writers on the country's history, economy, society, and culture; FRANK M. LE BAR and ADRIENNE SUDDARD (eds.), *Laos: Its People, Its Society, Its Culture* (1960), a general survey by American scholars; T.D. ROBERTS (ed.), *Area Handbook for Laos* (1972), a highly readable survey. E.H.G. DOBBY, *Southeast Asia,* 10th ed. (1967), a standard work; C. ROBEQUAIN, *L'Évolution économique de l'Indochine française* (1939; Eng. trans., *Economic Development of French Indo-China,* 1944), an older but useful study. J.M. HALPERN, *Government, Politics and Social Structure in Laos: A Study of Tradition and Innovation* (1964), a detailed treatment by a leading authority; *Economy and Society of Laos: A Brief Survey* (1964), a good detailed overview; P. KUNSTADTER (ed.), *Southeast Asian Tribes, Minorities, and Nations,* 2 vol. (1967), with four excellent chapters on the peoples of Laos; FRANK M. LE BAR, G.C. HICKEY, and J.K. MUSGRAVE, *Ethnic Groups of Mainland Southeast Asia* (1964), a good general treatment of the peoples of Laos. A. SCHAAF, C. HART, and R.H. FIFIELD, *The Lower Mekong: Challenge to Cooperation in Southeast Asia* (1963); UNITED NATIONS, *Economic and Social Survey of Asia and the Pacific,* annual (1948–), and *Statistical Yearbook for Asia and the Pacific,* annual (1970–), are reliable sources of economic data and analysis. A.J. DOMMEN, *Conflict in Laos: The Politics of Neutralization* (1964); P.F. LANGER and J.J. ZASLOFF, *North Vietnam and the Pathet Lao* (1970); SISOUK NA CHAMPASSAK, *Storm over Laos: A Contemporary History* (1961). P.B. LAFONT, *Bibliographie du Laos* (1964), a listing of articles and books published from 1663 to 1960; *Annales du Laos* (1926), a French translation of Lao documents; G. COEDES, *Les Peuples de la péninsule indochinoise* (1962); D.G.E. HALL, *A History of South-East Asia,* 3rd ed. (1968); P. LE BOULANGER, *Histoire du Laos français: Essai d'une étude chronologique des principautés laotiennes* (1931), a history of Laos from 1353 to 1930; W.G. BURCHETT, *Mekong Upstream* .(1957), on the origins of the Pathet Lao movement, and J.J. ZASLOFF, *The Pathet Lao: Leadership and Organization* (1973). A.J. DOMMEN, *Conflict in Laos: The Politics of Neutralization* (1964); B.F. FALL, *Anatomy of a Crisis: The Laotian Crisis of 1960–61* (1969); N.S. ADAMS and A.W. MCCOY, *Laos: War and Revolution* (1970); D. LANCASTER, *The Emancipation of French Indo-China* (1961); K.P. LANDON, "Thailand's Quarrel with France in Perspective," *Far Eastern Quarterly,* 1:25–42 (1941); A. PAVIE *et al., Mission Pavie: Géographie et voyages,* 7 vol. (1900–19); S. PENSRI, *Les Relations entre la France et la Thailande (Siam) au XIX^e^ siècle, d'après les archives des Affaires Étrangères* (1962), on the French penetration; K.D. SASORITH, *Contribution à l'histoire du mouvement d'indépendance nationale Lao* (1948), on the Lao Issara movement, by its chief spokesman; H. TOYE, *Laos: Buffer State or Battleground* (1968); D.K. WYATT,"Siam and Laos 1767–1827," *Journal of South East Asian History,* 4:13–32 (1963); and *Iron Man of Laos,* by PRINCE PHETSARATH RATANAUONGSA (1978), in effect, a history of Laos during the Vichy rule, Japanese domination, and immediate postwar period.

Malaysia: S.T. ALISJAHBANA *et al.* (eds.), *The Cultural Problems of Malaysia in the Context of Southeast Asia* (1967), articles on aspects of the culture, religion, and languages of the peoples of Malaysia; SUDHIR ANAND, *Inequality and Poverty in Malaysia: Measurement and Decomposition* (1982), a study that discusses poverty and income distribution in Malaysia; the national atlas of Malaysia, *Atlas Kebangsaan Malaysia* (1977), features a wide range of geographic and demographic information; JASBIR CHHABRA *et al., The World Tin Economy: An Econometric Analysis* (1978), an account of particular relevance to Malaysia; C.A. FISHER, *South-East Asia* (1964), an important work on the geography of Southeast Asia, with a substantial section on the states of Malaysia; D.W. FRYER, *Emerging Southeast Asia,* 2nd ed. (1979), a wide-ranging analytical study of the economy of Southeast Asian countries, with a perceptive chapter on Malaysia; J.C. JACKSON, *Sarawak: A Geographical Survey of a Developing State* (1968), a comprehensive, well-documented survey; L.W. JONES, *The Population of Borneo* (1966), a demographic study of the peoples of Sarawak, Sabah, and Brunei, based on 1960 census statistics; V. KANAPATHY, *The Malaysian Economy: Problems and Prospects* (1970), a series of essays by an economist on manpower, agriculture, industrialization, financial institutions, and international trade; FRANK M. LEBAR (ed.), *Ethnic Groups of Insular Southeast Asia,* vol. 1, *Indonesia, Andaman Islands, and Madagascar* (1972), a discussion of the ethnic groups of East Malaysia; YONG LENG LEE, *North Borneo (Sabah): A Study in Settlement Geography* (1965), an examination of the problems and features of settlement in an equatorial environment, *Population and Settlement in Sarawak* (1970), a well-documented

study of the peoples and their impact on the landscape of Sarawak; CHONG-YAH LIM, *Economic Development of Modern Malaya* (1968), a detailed study of the evolution and growth of the West Malaysian economy from 1874 to 1963. NENA VREELAND *et al.* (eds.), *Area Handbook for Malaysia* (1977), a comprehensive compilation of basic information on the social, political, and economic background and national security of Malaysia; DIPAK MAZUMDAR, *The Urban Labor Market and Income Distribution: A Study of Malaysia* (1981), an analysis of Malaysia's urban work force; JIN-BEE OOI, *Peninsular Malaysia,* new ed. (1976), a comprehensive study of the physical environment, resources, population, and economy of West Malaysia; N.J. RYAN, *The Cultural Heritage of Malaya,* 2nd ed. (1971), a well-written, neatly organized book on the culture of the main ethnic groups of West Malaysia; B.C. STONE (ed.), *Natural Resources in Malaysia and Singapore* (1969), proceedings of a symposium held in 1967 containing numerous articles; GUNGWU WANG (ed.), *Malaysia: A Survey* (1964), studies on the history, geography, politics, economy, and societies of Malaysia; E.L. WHEELWRIGHT, *Industrialization in Malaysia* (1965), an enquiry into the industrial structure and the industrialization policies of Malaysia (and Singapore); KEVIN YOUNG, WILLEM C.F. BUSSINK, and PARVEZ HASAN, *Malaysia: Growth and Equity in a Multiracial Society* (1980), an argument for rapid growth in Malaysia to overcome the nation's economic and social problems. For contemporary information, see *Information Malaysia* (annual; formerly the *Malaysia Yearbook*). K.G. TREGONNING, *Southeast Asia: A Critical Bibliography* (1969), numerous annotated entries devoted to recent books and articles, including many bibliographical references on Malaya, and *A History of Modern Malaysia and Singapore,* rev. ed. (1972), a substantial survey; P. WHEATLEY, *The Golden Khersonese: Studies in the Historical Geography of the Malay Peninsula Before A.D. 1500* (1961, reprinted 1973), a brilliant analysis; J.M. GULLICK, *Malaysia* (1969), a comprehensive and valuable account; V. PURCELL, *The Chinese in Southeast Asia,* 2nd ed. (1965); W.R. ROFF, *The Origins of Malay Nationalism* (1967), an examination of the growth of the dominant political factor in the peninsula; CHONG-YAH LIM, *Economic Development of Modern Malaya* (1968); E. THIO, *British Policy in the Malay Peninsula, 1880–1910* (1969), a scholarly examination of the expansion of British control; C.N. PARKINSON, *British Intervention in Malaya, 1867–1877* (1960); J.-B. OOI, *Peninsular Malaysia,* new ed. (1976), a comprehensive geography; R. ALLEN, *Malaysia: Prospect and Retrospect* (1968), a discussion of the effects of colonial rule on Malaya.

Myanmar: E.H.G. DOBBY, *Southeast Asia,* 11th ed. (1973), gives a general description of Myanmar as a part of Southeast Asia and also contains a special section on the country itself, with detailed maps. O.H.K. SPATE, *Burma Setting* (1943), gives a brief but detailed introduction to the physical and social geography and some details of the economic geography up to the outbreak of World War II. H.L. CHHIBBER, *Physiography of Burma* (1933, reprinted 1975) and *Geology of Burma* (1934); and the *British Burma Gazetteer,* 2 vol. (1880, reprinted as the *Gazetteer of Burma,* 1983), are standard works. R. KAULBACK, *Salween* (1939); and L.D. STAMP, "Irrawaddy River," *Geogrl. J.,* 95:329–356 (1940), describe the drainage system. J.R. ANDRUS, *Burmese Economic Life* (1948); and J.S. FURNIVALL, *Political Economy of Burma,* 2nd rev. ed. (1938), describe the economy during the colonial period. FRANK N. TRAGER, *Burma: From Kingdom to Republic* (1966), contains an account of the economy before and after independence, up to the years of extensive nationalization in 1962–63. MAUNG HTIN AUNG, *A History of Burma* (1967); and MELFORD E. SPIRO, *Kinship and Marriage in Burma* (1977), give the historical background and contain accounts of Myanmar society and culture. JOSEF SILVERSTEIN, *Burmese Politics: The Dilemma of National Unity* (1980), is a useful analysis of contemporary politics. U AUNG THAW, "Neolithic Culture of Padhalin Caves," *Journal of Burma Research Society,* vol. 52 (1969), contains a good account of the Stone Age in Myanmar. See also the articles by G.H. LUCE on the Pyus, Mons, and Early Burmans, *ibid.,* vol. 27 (1932), vol. 34 (1950), and vol. 43 (1959); to be read with MAUNG HTIN AUNG, *Burmese History Before 1287: A Defence of the Chronicles* (1970). A.P. PHAYRE, *A History of Burma* (1883, reissued 1969); and G.E. HARVEY, *A History of Burma* (1925, reissued 1969), are standard works for the pre-British period (*i.e.,* to 1824). G.E. HARVEY, *British Rule in Burma, 1824–1942* (1946, reprinted 1974), attempts to justify British rule. DOROTHY WOODMAN, *The Making of Burma* (1962); MAUNG MAUNG, *Burma in the Family of Nations,* 2nd ed. rev. (1957); and MAUNG HTIN AUNG, *The Stricken Peacock: Anglo-Burmese Relations, 1752–1948* (1965), deal with the history of Anglo-Myanmar relations. The history for the post-independence period is given in H. TINKER, *The Union of Burma,* 4th ed. (1967); FRANK N. TRAGER, *Burma: From Kingdom to Republic* (1966, reprinted 1976); JOSEF SILVERSTEIN, *Burma: Military Rule and the Politics of Stagnation* (1977), and *Burmese Politics: The Dilemma of National Unity* (1980); and DAVID I. STEINBERG, *Burma's Road Toward Development: Growth and Ideology Under Military Rule* (1981), and *Burma: A Socialist Nation of Southeast Asia* (1982). D.G.E. HALL, *A History of Southeast Asia,* 4th ed. (1981), contains chapters on Myanmar history that give the British view.

Singapore: E.H.G. DOBBY, "Singapore, Town and Country," *Geogrl. Rev.,* 30:84–109 (1940); OOI JIN-BEE and CHIANG HAI DING (eds.), *Modern Singapore* (1969). Current information may be found in the (Singapore) *Monthly Digest of Statistics,* the *Singapore Year Book,* and the *Yearbook of Statistics.* R. HO, "The Physical and Human Background," in *Impact of Man on Humid Tropics Vegetation* (1960); S. NIEUWOLT, "Uniformity and Variation in an Equatorial Climate," *J. Trop. Geogr.,* 27:23–39 (1968); I.E.M. WATTS, "Rainfall of Singapore Island," *Malay. J. Trop. Geogr.,* 7:1–68 (1955). W. NEVILLE, "The Areal Distribution of Population in Singapore," *J. Trop. Geogr.,* 20:16–25 (1965); YOU POHSENG, "The Population of Singapore, 1966," *Malay. Econ. Rev.,* 12:59–96 (1967). C.T. EDWARDS, *Public Finances in Malaya and Singapore* (1970); FIRST NATIONAL CITY BANK, *Singapore: Foreign Investment Guide* (1969); GOH KENG SWEE, *Decade of Achievement* (1970); H. HUGHES and YOU POH-SENG, *Foreign Investment and Industrialization in Singapore* (1969). A. JOSEY, *Lee Kuan Yew and the Commonwealth* (1969); TAE YUL NAM, "Singapore's One-Party System," *Pacif. Affairs,* 42:465–480 (1969–70). M. FREEDMAN and M. TOPLEY, "Religion and Social Realignment among the Chinese in Singapore," *J. Asian Stud.,* 21:3–23 (1961); C. GAMBA, "Some Social Problems in Singapore," *Aust. Q.,* 26:99–106 (1954); B.B. KAYE, *Upper Nankin Street, Singapore: A Sociological Study of Chinese Households Living in a Densely Populated Area* (1966); SINGAPORE HOUSING AND DEVELOPMENT BOARD, *Annual Report.*

Thailand: The annual *Thailand Official Year Book* is the best source for current information. An excellent, well-illustrated pamphlet of the TRIBAL RESEARCH CENTER OF THAILAND, "Tribesmen and Peasants in North Thailand, 1967" (1969), briefly describes all the tribesmen in Thailand. The ROYAL THAI SURVEY DEPARTMENT, *Thailand National Resources Atlas* (1969), is the only atlas that can be used authoritatively as a reference. MASASHICHI NISHIO, "Public Health in Thailand," *Southeast Asian Studies,* vol. 11, no. 1 (1964; Eng. trans. JPRS 32217, 1965), based on scholarly research, gives a true picture of health conditions in Thailand. On physical geography, see ROBERT L. PENDLETON, *Report to Accompany the Provisional Map of the Soils and Surface Rocks of the Kingdom of Siam* (1953); and the SIAM, MINISTRY OF COMMERCE AND COMMUNICATIONS, *Nature and Industry* (1930). General histories include: W.A.R. WOOD, *A History of Siam,* 2nd ed. (1933), essentially a chronicle of reigns; and PRINCE CHULA CHAKRABONGSE, *Lords of Life* (1960), a sympathetic account stressing the Bangkok period. A general account of the Chinese minority is G.W. SKINNER, *Chinese Society in Thailand: An Analytical History* (1957). On the Ayutthaya period, see CHARNVIT KASETSIRI, *The Rise of Ayudhya: A History of Siam in the 14th and 15th Centuries* (1977); GEORGE V. SMITH, *The Dutch in Seventeenth-Century Thailand* (1977); and E.W. HUTCHINSON, *Adventurers in Siam in the Seventeenth Century* (1940). The early Bangkok period is discussed in KLAUS WENK, *The Restoration of Thailand under Rama I, 1782–1809* (1968); WALTER F. VELLA, *Siam under Rama III, 1824–1851* (1957); and AKIN RABIBHADANA, *The Organization of Thai Society in the Early Bangkok Period, 1782–1873* (1969). For developments that followed the opening of the country to the West, see ABBOT LOW MOFFAT, *Mongkut: The King of Siam* (1961); JAMES C. INGRAM, *Economic Change in Thailand Since 1850* (1955); and DAVID K. WYATT, *The Politics of Reform in Thailand* (1969); TEJ BUNNAG, *The Provincial Administration of Siam, 1892–1915* (1977); WALTER F. VELLA, *Chaiyo!* (1978); and THAWATT MOKARAPONG, *History of the Thai Revolution* (1972). JOHN L.S. GIRLING, *Thailand: Society and Politics* (1981), a comprehensive overview.

Vietnam: A general view of the geography of Vietnam can be found in E.H.G. DOBBY, *Southeast Asia* (1950, 11th ed., 1973), while more regional information is contained in P. GOUROU, *Les paysans du delta Tonkinois* (1965), and F.R. MOORMAN, *The Soils of the Republic of Viet-Nam* (1961). Information on Vietnamese forests is included in L. WILLIAMS, *Vegetation of Southeast Asia: Studies of Forest Types, 1963–1965* (1965?). H. DE MONESTROL, *Chasses et faune d'Indochine* (1952) deals with wild game and hunting, while P. WILDASH, *Birds of South Vietnam* (1968) is a comprehensive ornithological work. J. BUTTINGER, *The Smaller Dragon: A Political History of Vietnam* (1958) and *Vietnam: A Dragon Embattled* (2 vol. 1967); G. COEDES, *Les Peuples de la péninsule Indochinoise* (1962; *The Making of South East Asia,* 1966); LE-THANH-KHOI, *Le Viêt-Nam: Histoire et civilisation* (1955); D.G.E. HALL, *A History of South-East Asia* (1955, 3rd ed., 1968); and D. DUNCANSON, *Government and Revolution in Vietnam* (1968). More specific historical books are L. BEZACIER, *Le Viêt-*

Nam (1972), on prehistory, and A. WOODSIDE, *Vietnam and the Chinese Model: A Comparative Study of Nguyên and Ch'ing Civil Government in the First Half of the Nineteenth Century* (1971), which focuses on the reigns of Gia Long and his son, Minh Mang. J.F. CADY, *The Roots of French Imperialism in Eastern Asia* (1954); V. THOMPSON, *French Indo-China* (1937); and D. LANCASTER, *The Emancipation of French Indochina* (1961). M. OSBORNE, *The French Presence in Cochinchina and Cambodia* (1969) concerns the late 19th and early 20th centuries in southern Vietnam and Cambodia, while C. ROBEQUAIN, *L'Évolution économique de l'Indochine français* (1939; *The Economic Development of French Indo-China,* 1944) treats French colonial economic policies. Sources on the Indochina War include P. MUS, *Viêt-Nam: Sociologie d'une guerre* (1952); E. HAMMER, *The Struggle for Indochina* (1954); P. DEVILLERS, *Histoire du Viêt-Nam de 1940 à 1952* (1952); and P. DEVILLERS and J. LACOUTURE, *La fin d'une guerre: Indochine 1954* (1960; *End of a War: Indochina, 1954,* 1969). D. WARNER, *The Last Confucian* (1963, rev. ed., 1964); G. MCT. KAHIN (ed.), *Governments and Politics of Southeast Asia* (1959, 2nd ed., 1964); R. SCIGLIANO, *South Vietnam: Nation Under Stress* (1964); R. SHAPLEN, *The Lost Revolution: The U.S. in Vietnam, 1946–1966* (1966); and D. HALBERSTAM, *The Making of a Quagmire* (1965). WILLIAM J. DUIKER, *The Communist Road to Power in Vietnam* (1981), a history of the rise of the communist party in Vietnam. Some works on the Vietnam War are G. MCT. KAHIN and J.W. LEWIS, *The United States in Vietnam* (1967); R. SHAPLEN, *The Road From War: Vietnam 1965–1970* (1970); D. WARNER, *Not With Guns Alone* (1977; *Certain Victory: How Hanoi Won the War,* 1978); and G. LEWY, *America in Vietnam* (1978). W. BURCHETT, *North of the Seventeenth Parallel* (1956, rev. ed., 1957); B. FALL, *Le Viet-Minh: La République Démocratique du Viet-Nam, 1945–1960* (1960); HOANG-VAN-CHI, *From Colonialism to Communism: A Case History of North Vietnam* (1964); J. LACOUTURE, *Ho Chi Minh* (1967, rev. ed., 1977; Eng. trans., *Ho Chi Minh,* 1968); P.J. HONEY (ed.), *North Vietnam Today* (1962); and D. PIKE, *Viet Cong: The Organization and Techniques of the National Liberation Front of South Vietnam* (1966). A general work on traditional Vietnamese society is P. HUARD and M. DURAND, *Connaissance du Viêt-Nam* (1954). A comprehensive work on traditional religious practices is L. CADIERE, *Croyances et pratiques religieuses de Vietnamiens,* 3 vol. (1958). HUYNH SANH THONG, *The Heritage of Vietnamese Poetry* (1979), and (ed.), *The Tale of Kieu* (1973), are analytical studies of Vietnamese traditional poetry. More recent works on Vietnamese society, economy, and politics include J. HENDRY, *The Small World of Khanh Hau* (1964); G.C. HICKEY, *Village in Vietnam* (1964); P. GHEDDO, *Cattolici e Buddisti nel Vietnam* (1968; *The Cross and the Bo-Tree: Catholics and Buddhists in Vietnam,* 1970). NGHIEM-DANG, *Viet-Nam: Politics and Public Administration* (1966); R. SANSOM, *The Economics of Insurgency in the Mekong Delta of Vietnam* (1970); J.C. SCOTT, *The Moral Economy of the Peasant: Rebellion and Subsistence in Southeast Asia* (1976); S.L. POPKIN, *The Rational Peasant: The Political Economy of Rural Society in Vietnam* (1979); and B. FALL, *The Two Viet-Nams: A Political and Military Analysis* (1963, 2nd rev. ed., 1967). Books on the ethnic groups of Vietnam include FRANK M. LEBAR, G.C. HICKEY, and J. MUSGRAVE, *Ethnic Groups of Mainland Southeast Asia* (1964); G. CONDOMINAS, *Nous avons mangé la forêt* (1950; *We Have Eaten the Forest,* 1977); J. DOURNES, *Coordonnées: Structures Jörai familiales et sociales* (1972); and P.B. LAFONT, *Toloi Djuat: Coutumier de la tribu Jarai* (1963). GEORGES COEDES, *Les Peuples de la Péninsule Indochinoise* (1962; Eng. trans., *The Making of Southeast Asia,* 1964), an authoritative history of the evolution of the peoples and nations of Indochina; L. BEZACIER, *Le Viêt-Nam* (1972), a major survey of Vietnam's archaeology; PIERRE HUARD and MAURICE DURAND, *Connaissance du Viêt-Nam* (1954), the most comprehensive study of Vietnamese civilization; LE THANH KHOI, *Le Viet-Nam: Histoire et civilisation* (1955), the first modern history of Vietnam by a leading Vietnamese historian; JOSEPH BUTTINGER, *The Smaller Dragon* (1958), a political history to 1900; GEORGES MASPERO, *The Kingdom of Champa* (1949), a translation of ch. 1 of *Le Royaume du Champa,* by the leading authority on the history of Champa; J.F. CADY, *The Roots of French Imperialism in Eastern Asia* (1954), a study of French policy of the two decades prior to military intervention in Vietnam and during the first decade of French rule in Indochina; M.E. OSBORNE, *The French Presence in Cochinchina and Cambodia* (1969), describes the establishment of French power in southern Vietnam up to the early 20th century; VIRGINIA THOMPSON, *French Indo-China* (1937), the standard work in English on the French colonial regime in Indochina. DAVID G. MARR, *Vietnamese Tradition on Trial: 1920–1945* (1981), an intellectual history. CHARLES ROBEQUAIN, *The Economic Development of French Indochina* (1944), the basic work on the French colonial economy; PHILIPPE DEVILLERS, *Histoire du Viêtnam de 1940 à 1952* (1952), the best account of French policy and Vietnamese reaction to it that led to the Indochina war; JOSEPH BUTTINGER, *Vietnam: A Dragon Embattled,* vol. 1, *From Colonialism to the Vietminh,* vol. 2, *Vietnam at War* (1967), a history of the period from the early years of French rule in Vietnam to the fall of the Diem regime in 1963; DONALD LANCASTER, *The Emancipation of French Indochina* (1961), a well-documented survey of French Indochina from the conquest to the establishment of the two Vietnams and the early years of the Diem regime; K.C. CHEN, *Vietnam and China, 1938–1954* (1969), the definitive work on North Vietnamese–Chinese relations before the 1954 Conference of Geneva; J.T. MCALISTER, JR. and PAUL MUS, *The Vietnamese and Their Revolution* (1970), excerpts from Paul Mus's untranslated profound study, *Viet-Nam: Sociologie d'une guerre* (1952); J.T. MCALISTER, JR., *Vietnam: The Origins of Revolution* (1969), an up-to-date history and perceptive analysis; W.J. DUIKER, *The Rise of Nationalism in Vietnam, 1900–1941* (1976), traces the rise to dominance of the Vietnamese Communists; M.E. GETTLEMAN (ed.), *Viet Nam: History, Documents, and Opinions on a Major World Crisis* (1965), a wide-ranging collection of documents relating the history of Vietnam to the contemporary conflicts; PHILIPPE DEVILLERS and JEAN LACOUTURE, *End of War: Indochina, 1954* (1969), a scholarly, detailed description of the failure of the French in 1954 and the Geneva Conference. B.B. FALL, *The Two Viet-Nams: A Political and Military Analysis* (1963), an account of both North and South Vietnam by an outstanding authority; NGO VINH LONG, *Before the Revolution: The Vietnamese Peasants Under the French* (1973), provides telling insights into the life of the peasantry; A.B. WOODSIDE, *Vietnam and the Chinese Model* (1971), a major study of Vietnam's complex historical relations with China, and *Community and Revolution in Modern Vietnam* (1976), a study of the evolution of modern Vietnamese society; D.G. MARR, *Vietnamese Anticolonialism, 1885–1925* (1971), an account of early reaction against the French; G. PORTER, *A Peace Denied: The United States, Vietnam, and the Paris Agreement* (1975), a detailed account of the negotiations to end the war; J. RACE, *War Comes to Long An: Revolutionary Conflict in a Vietnamese Province* (1972), a detailed analysis of the war in a major southern province; J. LACOUTURE, *Ho Chi Minh* (1967, rev. ed., 1977; Eng. trans., *Ho Chi Minh,* 1968), the best biography available in English translation; ROBERT SCIGLIANO, *South Vietnam: Nation Under Stress* (1963), the best political analysis of South Vietnam under the regime of Ngo Dinh Diem; ROBERT SHAPLEN, *The Lost Revolution: The U.S. in Vietnam, 1946–1966* (1966), a description of the failure of anti-Communist nationalism, especially under Diem; G.M. KAHIN and J.W. LEWIS, *The United States in Vietnam* (1967), a history and critical analysis of American involvement in Vietnam; DOUGLAS PIKE, *Viet-Cong: The Organization and Techniques of the National Liberation Front of South Vietnam* (1966), the standard work on this subject.

(L.C.O./D.P.Ch./M.E.O./M.H.Au./J.Si./P.-B.L./O.J.B./
K.G.T./Rt.H./P.P.A./W.F.V./G.C.H./J.Bu.)

Southeast Asian Arts

The term Southeast Asia refers to the huge peninsula of Indochina and the extensive archipelago of what is sometimes called the East Indies. The region can be subdivided into mainland Southeast Asia and insular Southeast Asia. The political units contained in this region are Burma, Thailand, Laos, Kampuchea (Cambodia), Vietnam, Malaysia, Singapore, Indonesia, and the Philippines. The Philippines originally was not included because Philippine history has not followed the general historical pattern of Southeast Asia, but, because of its geographic position and the close affinities of its primitive cultures with the primitive cultures of Southeast Asia, it is now usually regarded as the eastern fringe of Southeast Asia. A common geographic and climatic pattern prevails over all of Southeast Asia and has resulted in a particular pattern of settlement and cultural development. Mountain people generally have a cultural level less developed than that of the valley dwellers. As a consequence, Southeast Asia is culturally fragmented. (For a discussion of the traditional cultures of the area, see ASIA: *Traditional cultures: Southeast Asia.*)

The article is divided into the following sections:

The cultural setting of Southeast Asian arts

Southeast Asia has been the crossroads of many races who have been contending against each other for centuries. At present, all the peoples of Southeast Asia are Mongoloid in racial origin. The first to come were the Austronesians (Malayo-Polynesians), sometimes described as Proto-Malays and Deutero-Malays. At one time they occupied the eastern half of mainland Southeast Asia, but later they were pushed toward the south and the islands by the Austro-Asiatics. At present, peoples of Austronesian origin occupy Malaysia, the Republic of Indonesia, and the Republic of the Philippines. There were three main Austro-Asiatic races, the Mons, the Khmers, and the Viet-Muong. The Mons were at one time dominant, but they lost their racial identity in the 18th century and became absorbed by the Burmese and the Tais; only a few thousand Mons are now found living near the Burma–Thailand border. The Khmers from the 9th century to the 15th built a great empire, but much of its territory was lost to its neighbours so that only the small kingdom of Cambodia remains today. The Viet-Muongs now occupy Vietnam. A Tibeto-Burmese tribe, the Pyu, founded an empire of city-kingdoms in the Irrawaddy Valley in the early centuries of the Christian Era, but the Pyus disappeared, and the Burmese, taking the leadership, founded their kingdom of Pagan and have occupied Burma up to the present day. In the 13th century the Tai-Shans lost their kingdom of Nanchao in Yunnan, China, and entered the Mae Nam Chao Phraya Valley to found kingdoms that gradually evolved into the kingdoms of Siam (Thailand) and Laos.

EXTERNAL INFLUENCES

The India influence

In Southeast Asia winds of change often came as storms. Indian commerce expanded into Southeast Asia in the early centuries of the Christian Era and, in spite of its peaceful nature, caused revolutionary changes in the life and culture of the peoples of the region. The Indians would sojourn in the region in small numbers for two or three monsoons only. The success of their commercial venture and the safety of their persons depended entirely on the goodwill of the inhabitants. At that time, the Indians were at a higher level of culture, and they brought new ideas and new art traditions. Since these ideas had some affinity with indigenous ideas and art forms, the natives, who had already reached a high level of culture themselves, accepted them but were not overwhelmed by an influx of new traditions. The Hindu–Buddhist culture of the Indians made a tremendous impact and came to form the second layer of culture in Southeast Asia, but the first layer of native ideas and traditions has remained strong to the present day.

European influence

Changes often came to Southeast Asia, usually because it possessed a commodity that was in great demand by the rest of the world. The Indians came because they were looking for fresh sources of gold after the Roman imperial source had run dry. In the 15th, 16th, and the 17th centuries, insular Southeast Asia attracted Islāmic merchants from India and farther west and later the Portuguese and the Dutch as a rich source of spices. As with the Hindu–Buddhist merchants of the past, the Islāmic traders came not as missionaries, though they did spread their religion in the region. The Portuguese came as conquerors and as militant missionaries of their Roman Catholic form of Christianity, and, for those reasons, their cultural traditions were found unacceptable to the natives. In the 17th century the Dutch came as conquerors and colonists for whom the attraction was first spices and then coffee, rubber, and petroleum. Since mainland Southeast Asia produced no spices for export, it was less vulnerable to the navies of Portugal and the Netherlands, so the region was not greatly affected by the Muslims, Portuguese, and Dutch. In the 19th century, Britain and France became interested in mainland Southeast Asia as the back door to China and sought to possess it as a colony. By the end of the 19th century, Burma had fallen to Britain, Siam was allowed to retain its independence only with the tacit permission of the two powers, and the rest had fallen to France. When in the mid-20th century the whole of Southeast Asia became free again, it was found that the arts of the peoples had remained merely static during the years of colonial rule; in some regions they had been driven underground by long neglect, but European culture and European art forms had made only a small impact.

INDIGENOUS TRADITIONS

The peoples of Southeast Asia were once thought to have shared a lack of inventiveness since prehistoric times and to have been "receptive" rather than "creative" in their contacts with foreign civilizations. Later excavations and discoveries in Burma and Thailand, however, inspired some scholars to argue against the accepted theory that civilization came to Southeast Asia from China in prehistoric times; rather, these scholars contend, the peoples of mainland Southeast Asia were cultivating plants, making pottery, and working in bronze about the same time as the peoples of the ancient Middle East, and therefore civilization spread from mainland Southeast Asia to China and India. Southeast Asians have never produced any theory of art or literary or dramatic criticism, for they are always more concerned with doing the actual work of producing beautiful things. Because the Southeast Asians, especially in the western half of the mainland, worked on nondurable materials, it is not possible to trace the development and evolution of art forms stage by stage. The region has always been thickly forested, so it was natural that the first material to be used for artistic purposes should have been wood. The wood-carving tradition, begun in primitive times, was retained even when they learned to work with metals and with stone and continued to flourish long after the great age of stone sculpture and stone architecture, which ended in the 13th century. Proto-Neolithic paintings discovered in a cave near the Salween River in the western Shan state of Burma have very close affinity with the later carvings on posts of houses among the Nāgas on the western hills of Burma. Similarly, cave paintings of a pair of human hands with open palms, one holding the sun and the other holding a human skull, are reflected in the later aesthetic tradition of Southeast Asia: the sun symbol is found as an art motif all over the region, and a suggestion of awe, triumph, and joy at acquiring a human head is found in carvings under the eaves of the Nāga houses. The cave painting testifies to the continuity of the magico-religious tradition connected with all the arts of the area.

The art of casting the bronze drums found at Dong Son, near Hanoi, which are similar to the bronze drums used by mountain tribes throughout Southeast Asia, was thought to have come from China, but recent excavations in Thailand proved that the drums and the so-called Dong Son culture itself are native to mainland Southeast Asia. In any case, the continuity of the aesthetic tradition of Southeast Asia can be seen in the bronze drums that were cast by the Karens of Burma for centuries until the early years of the 20th century. The mountains of mainland Southeast Asia provided gold, silver, and other metals, and the art of metalworking must have developed quite early. Silver buttons, belts, and ornaments now made and worn by the hill peoples in Southeast Asia have behind them a very ancient tradition of workmanship. The same artistic tradition is found in textile designs.

Music, dance, and song were originally associated with tribal rituals. From the beginning, the main characteristic of Southeast Asian music and dance has been a swift rhythm. The slow and stately dances of the Siamese court were of Indian origin; when they were introduced into Burma in the 16th century, the Burmese quickened the tempo, but, even with that modification, the dances were still called Siamese dances to distinguish them from the native ones. In their oral literature—namely, in folk songs and folktales—the emphasis is on gaiety and humour. Typically, Southeast Asians do not like an unhappy ending.

ıskrit
l Pāli
rature

The role of royal patronage and religious institutions. In all the regions of Southeast Asia, the arts flourished under the patronage of the kings. About the time of the birth of Christ, tribal groups gradually organized themselves, after some years of settled life as rice cultivators, into city-kingdoms, or conglomerations of villages. A king was thus little more than a paramount tribal chieftain. Since the tribes had been accustomed to worshipping local spirits, the kings sought a new spirit that would be worshipped by the whole community. One reason why the gods of Hinduism and Buddhism were found so readily acceptable to Southeast Asia was this need for new national gods. The propagation of the new religions was the task of the kings, and consequently the period from the 1st to the 13th century was a great age of temple building all over Southeast Asia. Architecture, sculpture, and painting on the temple walls were the arts that flourished. In the ancient empires of eastern Indochina and the islands, scholars of Sanskrit, the language of the sacred works of Hinduism, became part of the king's court, producing a local Sanskrit literature of their own. This literary activity was confined to the hereditary nobility and never reached the people, except in stories from the great Hindu epics *Mahābhārata* and *Rāmāyaṇa*. Because the Hindu religious writings in Sanskrit were beyond the reach of the common people, Hinduism had to be explained to them by Hindu stories of gods and demons and mighty men. On the other side of the peninsula, in the Pyu-Burmese empire of Prome, which flourished before the 8th century, there was no such development—first, because Hinduism was never widely accepted in Burma and, second, because the more open Burmese society developed neither the institution of a god-king nor that of a hereditary nobility. Although Pāli scholars surrounded the king in later Pagan, Pāli studies were pursued not at the court but at monasteries throughout the kingdom so that even the humblest villager had some faint contact with Pāli teachings. While the courts of the kings in Cambodia and Java remained merely local centres of Sanskrit scholarship, Pagan became a centre of Pāli learning for Buddhist monks and scholars even from other lands. As in the case of stories from the Indian epics, stories of the *Jātaka*s (birth stories of the Buddha) were used to explain Buddhism to the common people, who could not read the scriptures written in Pāli. Just as scenes from the great epics in carving or in fresco adorned the temples in Cambodia and Java, scenes from the *Jātaka*s adorned the Pagan temples.

Musicians of the Pyu kingdom played before the Emperor of China in 801, and the various musical instruments at the performance have their counterparts at the present day, not only in Burma but throughout Southeast Asia. At Pagan the people were so fond of music that even the collection of taxes became an occasion to dance and sing, and a royal official, endowing a temple, inscribed a prayer asking that in all his future existences until he reached Nirvāṇa "might he be woken up every morning to the strains of music sweetly played on flute and violin." In spite of this love for music and dance, no dramatic art seems to have developed in Burma, perhaps because Sanskrit, in which there was a dramatic tradition, was not studied. In contrast, at the courts of Cambodia and Java, the Sanskrit drama, Hindu dances, and native dance traditions combined and produced the court opera ballets. These dramatic elements later reached the common people by way of the shadow play.

Significance of the temples

The patronage of the king and the religious enthusiasm of the common people could not have produced the great temples without the enormous wealth that suddenly became available in the region following the commercial expansion. With the Khmer and Javanese empires, the wealth was produced by a feudalistic society, and so the temples were built by the riches of the king and his nobles, combined with the compulsory labour of their peasants and slaves, who probably derived some aesthetic pleasure from their work because of their religious fervour. Nonetheless, their monuments, such as Borobuḍur, in Java, and Angkor Wat, in Cambodia, had an atmosphere of massive, all-conquering power. At Pagan, where wealth was shared by the king, the royal officials, and the common people, the temples and the monasteries were built by all who had enough not only to pay the artisans their wages but also to guarantee their good health, comfort, and safety during the actual construction. The temples were dedicated for use by all monks and lay people as places of worship, meditation, and study, and the kings of Pagan did not build a single tomb for themselves. The Khmer temple of Angkor Wat and the Indonesian temple of Borobuḍur were tombs in that the ashes of the builders would be enshrined therein; the kings left stone statues representing them as gods for posterity to worship, whereas at Pagan there was only one statue of a king, and it depicted him on his knees with his hands raised in supplication to the Buddha. Consequently, the atmosphere that pervaded the temples of Pagan was one of joy and tranquillity.

This golden age of wealth and splendour in Southeast Asia ended in the 13th century with a sudden violence, when Kublai Khan's Tatar Chinese armies destroyed both the Burmese and the Khmer empires and his navy attacked Vietnam and Java. The tiny kingdoms that subsequently sprang up all over Southeast Asia continually fought among themselves; their kings were neither powerful nor rich, and the royal courts became centres of military planning and political intrigue. During the 13th and 14th centuries, in the new Javanese kingdom of Majapahit and the new Burmese kingdom of Ava, vernacular literatures came into being. Again, differences in social structure had aesthetic repercussions. In Majapahit the king was powerful and gave his patronage to the newly arisen literature, confining it to the court. At Ava the vernacular literature bloomed throughout the kingdom, and the king, lacking power and prestige, prevailed upon some established writers to join the court circles and give them glamour.

After Majapahit, a new cultural force—namely, Islām—reached insular Southeast Asia, and over the two layers of primitive and Hindu–Buddhist cultures was added the third layer of Islām. In mainland Southeast Asia, a new Burmese empire arose over the ruins of the old and continued its task of spreading Buddhism. Hindu tradition reached the Burmese court secondhand in the 18th century as the result of the Burmese conquest of Siam and was one of the factors that contributed to the rise of a Burmese drama. On the other side of the peninsula, Vietnam, reconquered by China, fell more and more under the influence of Chinese culture. After a short period of Islāmic bloom, native culture in insular Southeast Asia was subjected to alien rule. In Burma and Siam alone among the states of Southeast Asia, native arts continued to flourish because, after centuries of warfare, they finally emerged as strong kingdoms.

Predominant artistic themes. The predominant themes of Southeast Asian arts have been religion and national history. In religion the main interest was not so much in actual doctrine but in the life and personality of the Buddha and the personalities and lives of the Hindu gods. In national history the interest was in the legendary heroes of the past, and this theme appeared only after the great empires had fallen and the memories of their glory and power remained. The Buddha image, which went through various stages of development, remained the favourite motif of sculpture and painting. The depiction of scenes from his previous lives in fresco and relief sculpture also had the purpose of teaching the Buddhist ethics to the people, as the *Jātaka*s emphasized certain moral virtues of the Buddha in his previous lives; it also gave an opportunity to the artist to introduce local colour by using, as background, scenes from his own contemporary time. The depiction of scenes from the Hindu epics also had the same purpose and gave the same opportunity to the artist. Many figures from the Buddhist and Hindu scriptures, such as gods and goddesses, heroes and princesses, hermits and magicians, demons and dragons, flying horses and winged maidens, became fused with similar native figures, and, gradually, folklore plots became merged in the general religious themes.

The *nāga*s

The *nāga,* a superhuman spirit, was taken from Buddhist and Hindu texts and merged with native counterparts, with the result that different images of the *nāga* appeared in various regions. The Burmese *nāga* was a snake with a crested head. The Mon *nāga* was a crocodile, and the Khmer and Indonesian *nāga* was conceived as a nine-headed snake. The demons of various kinds from all over Southeast Asia became merged under one name of Pāli-Sanskrit origin, *yakkha* or *yakṣa,* but they retained their separate identities in sculpture and paintings of their own different countries. The lion, which was unknown to the monsoon forest but was a figure of Hindu and Buddhist mythology, evolved into a native symbol and art motif. The primitive worship of the snake-dragon as a god of fertility was retained in the Khmer Empire; the nine-headed *nāga* became a symbol of security and of royalty, and stone *nāga*s guarded the palaces and temples. Buddhism frowned upon *nāga* worship. In Burmese and Mon sculpture the *nāga* was always shown as a servant of the Buddha, putting his body in coils to make a seat for his master and raising his great hood as an umbrella over his master's head. According to tradition, the guardian figure of a Mon temple was a two-bodied lion with a man's head, and the guardian figure of a Burmese temple was the crested lion. The Tais made themselves heirs to both the Khmer and the Mon art traditions relating to the *nāga,* but the guardian figure of their temples was the benevolent demon.

Primitive symbols and animal imagery merged with Indian animals and entered the arts. The Pyus embossed the primitive symbol of the sun on their coins as insignia of their power, and the Burmese transformed it into their favourite bird, the peacock, on the excuse that Buddhist mythology associated the peacock with the sun; the Mons adopted the red sheldrake as their symbol, and in Indonesia the mythical bird called Garuḍa, the vehicle of Vishnu, became merged with the local eagle. The figures of these birds also became decorative motifs. Animals of the Southeast Asian forests whose figures had adorned primitive dwellings of wood and thatch were stylized and came to adorn palaces and monasteries. Primitive geometrical patterns mixed with new spirals and curves from India, and Indian floral designs merged with those of trees and fruits and flowers copied from the monsoon forests.

THE UNIQUE AESTHETIC OF THE REGION

The arts of Southeast Asia have no affinity with the arts of other areas, except India. Burma was always an important route to China, but Burmese arts showed very little Chinese influence. The Tais, coming late into Southeast Asia, brought with them some Chinese artistic traditions, but they soon shed them in favour of the Khmer and Mon traditions, and the only indications of their earlier contact with Chinese arts were in the style of their temples, especially the tapering roof, and in their lacquer ware. Vietnam was a province of China for 1,000 years, and its arts were Chinese. The Hindu archaeological remains in southern Vietnam belong to the ancient kingdom of Champa, which Vietnam conquered in the 15th century. The Buddhist statues in northern Vietnam were Chinese Buddhist in style. The essential differences in aesthetic aim and style between the arts of East Asia and those of Southeast Asia could be seen in the contrast between the emperors' tombs of Vietnam and the temple-tombs of Cambodia and Indonesia or the opulent and dignified Buddha images of Vietnam and the ascetic and graceful Buddha images of Cambodia and Burma. Islāmic art, with its rejection of animal and human figures and its striving to express the reality behind the false beauty of the mundane world, also has no affinity with Southeast Asian arts. Both Hinduism and Buddhism taught that the sensual world was false and transitory, but this message found no place in the arts of Southeast Asia. The world depicted in Southeast Asian arts was a mixture of realism and fantasy, and the all-pervading atmosphere was a joyous acceptance of life. It has been pointed out that Khmer and Indonesian classical arts were concerned with depicting the life of the gods, but to the Southeast Asian mind the life of the gods was the life of the peoples themselves—joyous, earthy, yet divine. The European theory of "art for art's sake" found no echo in Southeast Asian arts, nor did the European division into secular and religious arts. The figures tattooed on a Burmese man's thigh were the same figures that adorned a great temple and decorated a lacquer tray. Unlike the European artist, the Southeast Asian did not need models, for he did not strive to be realistic and correct in every anatomical detail. This intrusion of fantasy and this insistence on the joyousness of human life have made Southeast Asian arts unique.

Literature

GENERAL CONSIDERATIONS

Regional distinctions. From the point of view of its "classical" literatures, Southeast Asia can be divided into three major regions: (1) the Sanskrit region of Cambodia and Indonesia; (2) the region of Burma where Pāli, a dialect related to Sanskrit, was used as a literary and religious language; and (3) the Chinese region of Vietnam.

There are no examples of Chinese literature written in Vietnam while it was under Chinese rule (111 BC–AD 939); there are only scattered examples of Sanskrit inscriptions written in Cambodia and Indonesia; yet most of the literary works produced at the court of Pagan in Burma (flourished *c.* 1049–1300) have survived because the texts were copied and recopied by monks and students. But in the 14th–15th centuries, vernacular literatures suddenly emerged in Burma and Java, and a "national" literature appeared in Vietnam. The reasons behind the development of each were the same: a feeling of nationalistic pride at the final defeat of Kublai Khan's invasions, the desire of the people to find solace in literature amidst change and struggles for power, and the lack of wealth and patronage to channel artistic expression into building temples and tombs. In Vietnam and Java literary activity centred on the courts; but in Burma the first writers were the monks and, later, the laymen educated in their monasteries. In the new Burmese kingdom of Ava (flourished after 1364), the Shan kings were proud of their Burmese Buddhist culture, and they appointed the new writers into royal service, with the result that courtiers became writers also. The Tai kings of Laos and Siam led their courts in learning Pāli from the Mons, whom they had conquered, and Sanskrit from the Khmers, whom they harassed; nevertheless, seized with national pride and influenced by the Burmese example, they developed their own vernacular literature. But Cambodia itself declined. Although the monks in the Theravāda Buddhist (*i.e.,* the Southeast Asian form of Buddhism) monasteries produced a few works in Pāli, no vernacular literature emerged until finally Khmer-speaking people (those living in the area comprised approximately of modern Kampuchea) were borrowing many words from the Tais.

For its vernacular literatures, Southeast Asia can be divided into (1) Burma; (2) Thailand, Laos, and Cambodia; (3) Vietnam; (4) Malaysia and Indonesia; and (5) the Philippines (which produced a vernacular literature only in the 20th century, after the imposed Spanish and English languages and literatures had made their impact).

ne era
' royal
tronage

Prestige of the writer. During the time of the kings, a Southeast Asian writer enjoyed patronage and a prestigious position in society. He could not, however, make a living by writing as a profession. Manuscripts had to be written by hand, and only in the case of famous works might one or two duplicates be made, again by hand. There was no question of selling the manuscript. A writer could only hope to attract the notice of his king and obtain a monetary reward or a royal office. By the time that printing presses were introduced, in the colonial period in the 19th century, the kings were gone and with them their writers. Colonial rule overwhelmed and destroyed vernacular literary traditions, leaving intact only oral literatures in the forms of folktales and folk songs. Literary criticism, as understood in Western cultures, had never been known, either in the ancient or modern literatures of Southeast Asia. Apart from a few stray writings on versification, therefore, no works of literary criticism or literary history existed until the colonial period. Even then, the interest of European scholars was chiefly confined to archaeology, and only a few made the attempt to study some special type or period of a vernacular literature (for example, vernacular versions of the *Rāmāyaṇa,* the great Sanskrit epic of India, or of 14th-century Javanese verse). There is a work in French dealing with Thai literature and a work in Burmese dealing with Burmese literature; but apart from these no study of any Southeast Asian literature as a whole has yet been made. For this neglect, native scholars are much to blame. Works on literary criticism and the history of literature would help give perspective to indigenous writers in the 20th century, both to those who want to cling to their native traditions (as is the case with writers in Burma and Thailand) and those who want to make a complete break with the past (as is generally the case in Vietnam, Indonesia, and Malaysia).

PRE-EUROPEAN COLONIAL PERIOD

Burma. The Burmese borrowed many words from Pāli but not to the extent that the Indonesians, Khmers, and the Thais borrowed Sanskrit words. The Burmese language was monosyllabic and tonal, and since there was no accent or stress, the feature that distinguished verse from prose was the regular occurrence of rhyme. They modelled their literature not on classic examples from Pāli or Sanskrit but on their own traditional folk songs.

The 15th century. In the 15th century, four types of verse existed: (1) *pyo* (religious verse), which retold stories of Buddha's birth and teaching and were taken from the *Jātaka*s (a collection of folktales adapted to Buddhist purposes and incorporated into the Pāli canon), to which were added imaginative details and a Burmese background; (2) *linkar* (shorter religious verse), or a devotional poem, characterized by a metaphysical flavour comparable in many ways to that which informs the work of the early 17th-century English poets George Herbert and Robert Herrick; (3) *mawgoon* (historical verse), half ode, half epic, written in praise of a king or prince and developing out of military marching songs; (4) *ayegyin* (lullaby), an informative poem usually addressed to a young prince or princess and written in praise of his royal ancestors.

Poets at
court

Literature in the 15th century is dominated by three monks: Shin Maha Rahta Thara, who wrote for the court of Ava, and Shin Maha Thila Wuntha and Shin Uttamagyaw, both of whom were of village stock and did not go to court but remained on in their village monasteries. Shin Maha Thila Wuntha, in the closing years of his life, turned to prose and wrote a chronicle history of Buddhism. In this period several courtiers, both men and women, also began to achieve some literary success, and the genre called *myittaza* (epistle) first evolved, which is a long prose letter written by a monk and addressed to the king to advise him of his duties.

The 16th century. In the 16th century, the Burmese conquered Siam, and their subsequent knowledge of Thai romantic poems gave rise to a new verse form called the *yadu* (the seasons). They borrowed only the theme, however, and not the form, and they developed it as an emotional poem, passionate, yet with something of the cool intellectual strength of the poems of the English metaphysical poets John Donne and Andrew Marvell. The most famous writers of the *yadu* were two court poets, Phyu and Nyo; a general of the army called Nawaday; and Natshinnaung, king of Toungoo. The wide popularity of the poems eventually gave rise to a mock-heroic form called *yagan* ("Kick the *yadu*").

Golden age of literature. In the early years of the 18th century, U Kala compiled a history of Burma, written in precise and clear prose; the closing years, which coincided with the establishment of the third Burmese Empire, saw a great period of literature. The Thai court, brought as captives to the Burmese capital, introduced to the Burmese poetic romances and their *Rāma* play (based on the *Rāmāyaṇa*). Contact with the Thais stimulated the growth of a Burmese court drama and led to the appearance of Burmese court romances in poetic prose. The king's treasurer, however, made fun of the Thai importations and wrote the *Rama Yagan,* in which the high romance and courtly elegance of the 4th-century-BC *Rāmāyaṇa* ("The Life of Rāma") were given a rustic setting, with hilarious results. From the quiet of their monasteries, the monk Awbatha wrote novel-like rendering of the *Ten Long Jātakas* and the monk Kyeegan Shingyi wrote homely, pithy, and sometimes even humorous myittaza ("epistles") from villagers to their relations in the cities.

The defeat suffered by the Burmese in the Anglo-Burmese War of 1824–26—their first defeat since the time of Kublai Khan in the 13th century—introduced a note of melancholy to Burmese literature. During this first half of the 19th century, many of the new melancholy lyrics were set to music. Two great writers were a product of this period: the dramatist U Kyin U and the courtier U Pon Nya, the greatest writer of the time, whose plays, epistles, and songs are full of humour and zest for life.

Thailand. Until 1824 Thai literature was entirely the province of the king and his court: the king maintained a corps of writers, and it was the custom to attribute authorship of any literary work to the king himself. Thai vernacular literature began with verse, based on Sanskrit models but relying on an elaborate rhyme scheme because the

Elaborate rhyme scheme of Thai poetry

Siamese language was tonal. The two earliest known poems were *Yoon Pai* ("The Defeat of the Yoons"), an epic-ode having similarities to the Burmese *mawgoon* genre, and *Mahajati* ("The Great *Jātaka*"), a poem stressing ethical ideas, similar in form to the Burmese *pyo.* Both poems, written during the period 1475–85, give ample proof that Thai writers, using Sanskrit, Khmer, and Burmese models, could nonetheless produce a truly Thai work.

First golden age: King Narai (1657–88). All literary activity ceased in the 16th century because of the unsettled conditions that prevailed before and after the annexation of the country by the Burmese. Independence was regained toward the close of the century, and under King Narai (1657–88), at his court in Ayutthaya, Siamese literature achieved its first golden age. Narai was himself a great poet, and during his reign new verse forms were evolved. He wrote poetic romances, based on stories from the "Fifty Jātakas," which were in fact folktales belonging to the region retold in Pāli and disguised as *Jātaka*s by an unknown Tai monk. Narai also wrote the final version of the poem of tragic romance, *Pra Lo* ("Lord Lo"), which had first been composed by an anonymous author in a much earlier reign. Among courtier poets of this time, the most famous were Maharajaguru; Si Prat, a wild young gallant who wrote the romantic poem *Aniruddha* (the name of the hero of the poem) and some passionate love songs; Khun Devakavi, author of cradle-songs using many Sanskrit and Khmer words but modelled on the Burmese *ayegyin;* and Si Mahosot, the author of an ode-epic in praise of King Narai. A new genre, the travel poem, also became popular; and the first versions of the plays *Rāma* and *Inao* (based on Hindu–Khmer–Javanese models) were composed by the King and his corps of writers. Perhaps the only prose work of the period was the *History of Ayutthaya* by Luang Prasroeth, which was lost and came to light only in the 20th century. It showed some signs of being influenced by U Kala's *History* (of Burma).

Second golden age: King Rama II (1809–24). Siam was conquered by the Burmese in 1767, and a new dynasty was established in a new capital, Bangkok. Some effort was made to revive the country's culture, largely destroyed following the sack of the old capital of Ayutthaya; and under the poet-king Rama II a second golden age of Thai literature occurred, during which women achieved prominence as poets for the first time. The King, with his writers, composed the final versions of *Rāma* and *Inao* and also a popular romance, "Khun Chang and Khun Pẹn," based on an incident in Thai history. The most famous poets were Prince Paramanuchit, whose ode-epic *Taleng Phai* ("The Defeat of the Mons") testified to his greatness, and Sunthon Phu, the King's private secretary, who was born of humble parents but made his way in the court by the excellence of his poetry. A strongly religious king, Rama III disbanded the corps of writers and discouraged the performance of plays at his court. Sunthon Phu lost his position but wrote his most famous poem, *Phra Aphaimani,* away from the court. A long fantasy-romance, this work can be regarded as the end of court domination in literature. Further, a royal official composed a Thai translation in prose (*Sam Kok*) of the Chinese classic *Romance of the Three Kingdoms.* The author, Pra Klang, was admittedly a royal official; nevertheless, the work was meant for the people rather than the court. It was followed by a spate of imitations and finally resulted in the development of the historical novel.

Final version of *Rāma*

Laos, Cambodia, and Vietnam. Laotian literature was in many respects a dialect branch of Tai literature, and, as in Thailand, it was the creation of the royal court. A number of popular romantic poems and prose lives of famous monks were composed, but their authors were unknown: all works, in fact, were by custom written anonymously.

The kings of Cambodia, fallen from high estate and often mere vassals of Thailand, could not inspire the rise of a vernacular literature. Only in the monasteries was there any literary activity, and this was written in the Pāli language.

In Vietnam, the emperors of the Tran dynasty (13th–14th century) were themselves poets and patronized a new literature—which, nevertheless, was still written in Chinese and was therefore national rather than vernacular. The writings themselves, however, were by no means a mere branch of Chinese literature. The country was afterward conquered once more by China and it was not until it regained independence that, under the patronage of the Le dynasty emperors (15th–16th century), a new age of literature began. Although the Chinese language was still used, some writers were beginning to use the vernacular (employing Chu-nom script, consisting of modified Chinese characters). Nguyen Trai, Emperor Le Thanh Tong, and Nguyen Binh Khiem were the great poets of this period. In 1651 Father Alexandre de Rhodes, a Roman Catholic missionary priest, invented a new romanized script (Quoc-ngu) that became the national script. Literature then began to reach the common people.

The national script

Literary works written before the end of the 18th century have not survived; the best known are those written in the 19th century, before the country became a French colony in 1862. Ho Xuan Huong, Nguyen Cong Tru, Chu Manh Trinh, and Tran Ke Xuong were famous court poets. Nguyen Du (1765–1820) wrote moral tales in verse that appealed not only to the court but to the common people. His most famous work was *Kim Van Kieu,* a poem of 3,253 lines, showing a strong Chinese influence (the plot was taken from a Chinese historical novel, and its ethical basis was both Confucian and Chinese Buddhist). The plays of the period, although written in Vietnamese, followed Chinese dramatic traditions because the Vietnamese theatre was still Chinese in style and practice.

Malaysia and Indonesia. Malaysia and Indonesia together have about 300 different languages and dialects, but they have a single common linguistic ancestor. Before the coming of Islām to the region in the 14th century, Javanese had been the language of culture; afterward, during the Islāmic period, Malay became the most important language—and still more so under later Dutch colonial rule so that, logically, it was recognized in 1949 as the official Indonesian language by the newly independent Republic of Indonesia.

During the period of Indian cultural influence, Sanskrit flourished in the great empires that included both the Malay Peninsula and the islands of present-day Indonesia. In the 11th century, at the court of Emperor Airlangga, a national literature (as distinct from a vernacular literature) emerged. It was written in courtly Javanese mixed with Sanskrit words, and it used Sanskrit metres and poetic style. In the 14th century in Majapahit (the new Javanese Empire that had been established after the final defeat of Kublai Khan's forces) a vernacular literature based on the speech of the common people came into being. The most important work of this new literature was *Nāgarakertāgama* (1365), a long poem in praise of the king (though it was not a product of the court) that also contained descriptions of the life of the Javanese people at the time. Although it employed a number of Sanskrit words, the style and metre were Javanese, not Sanskrit.

The Indian Hindu epics had already been popularized in the Malay Peninsula and in the islands of Indonesia (by way of the shadow-puppet play), and in this period fresh versions began to be written in the new Javanese. Romances, called *hikayat,* both in verse and in prose, also appeared—having as their source native myth and legend. Soon Malay, Balinese, Sundanese, and Madurese vernacular literatures emerged, all dealing with the same themes.

Romance themes from myth and legenc

The coming of Islām coincided with the rise of Malacca and the decay of Majapahit; but the popular fantasy-romances were able to survive by adopting a Muslim, instead of a Hindu, guise. New romances, telling the stories of heroes known to Islām, such as Alexander the Great, Amīr Hamzah, and Muḥammad ibn al-Ḥanafiah, were added to their number, and translations of Persian Muslim stories and of works on Muslim law, ethics, and mysticism further enriched Malay literature.

The finest work of all in the Malay language was the *Malay Annals,* written in about the 15th century. It gave a romanticized account of the history of the kingdom of Malacca and a vivid picture of life in the kingdom. Although a court record that begins with ancestral myths, it goes on to describe latter-day events of the kingdom with realism and humour.

In the Malay Peninsula, the coming of colonial rule did not at once overwhelm the existing native literature. As at the courts of the sultans of the British federated Malay states, the old traditions continued for some time. In Indonesia, however, a complete break was made with the cultural tradition.

EUROPEAN COLONIAL AND MODERN PERIODS

mpact of
olonial
ule

The entire region of Southeast Asia, with the single exception of Thailand, fell under colonial rule, and Thailand itself survived more as a buffer state than as a truly independent kingdom. At the courts of the kings of Laos, Cambodia, and Vietnam, which fell under French suzerainty, and in the palaces of the sultans of the British Malay states, vernacular literatures managed to survive for a time; but since these literatures had long ago ceased to develop—as a result of harassment by the Thais in the case of Laos and Cambodia, by the Portuguese in the case of Malaya, and by the French in the case of Vietnam—they soon became moribund. In all of Southeast Asia, except Burma and Thailand, the vernacular languages themselves lost their status, as the languages of the colonial rulers became the languages of administration and of a new elite. A revival of interest in the native languages and literatures occurred only toward the close of the colonial period, as a consequence of national movements for freedom.

Burma. In Burma, unlike India and other parts of the British Empire, English did not fully replace Burmese as the language of administration. In the almost classless Burmese society the language of the court and of literature was also the language of the people, which prompted the British government to retain Burmese as a second official language and to make both languages compulsory for study in schools and colleges. As a result, no English-speaking elite emerged, English literature did not dazzle native scholars, and, although its growth was retarded, Burmese literature did not disappear. With the intensification of the movement for freedom, around 1920, political tracts, novels, short stories, and poems reflected a political bias against colonial rule. In 1930, at the University of Rangoon, a group of young writers developed a new style of Burmese prose and poetry, a style little influenced by Western literature. In the post-independence period, novels and poems became centred on biographical and historical writings.

Thailand. Administrative and educational reforms introduced by King Mongkut (1851–68) as an answer to the threat of colonial conquest created a liberal atmosphere and a new reading public, and soon many of the old courtly writings were popularized in the form of romantic prose fiction. About 1914, King Vajiravudh, a graduate of Cambridge University, attempted to win back for the palace the leadership in literature; although he produced some fine adaptations of Shakespeare's plays, they made no impact on the people, with whom romantic fiction remained popular. Because of increased contact with the West, after World War II novels and short stories based on western models began to rival the earlier prose romances.

Laos, Cambodia, and Vietnam. Because of France's restrictive colonial educational policy, French language and literature never reached the common people. Moreover, the French-speaking elite, engrossed in French literature, neglected the native literature. With the growing vehemence of the freedom movement in the 1930s, however, there developed in Vietnam a new school of vernacular poetry that was less traditional and more nationalistic. But in the turbulent years that followed, the poets, including Ho Chi Minh himself, became occupied more with war than with literature.

lalaysian
lalay

Malaysia and Indonesia. The first Malaysian newspaper in the vernacular language, which appeared in 1876, introduced a new style of prose, less literary and nearer to the spoken Malay. Becoming immensely popular, the new style was further developed by other newspapers. (Although the early innovators were influenced by the English language, their followers were influenced by Arabic.) In about 1920 this new "Malaysian Malay" finally replaced the old literary Malay. The Translation Bureau, established by the British government in 1926, translated a great number of English books into the new Malay. In Indonesia, also, the old cultural language, literary Javanese, ceased to be used; by the end of the 19th century young Indonesians, overwhelmed by Dutch literature, started to write in Dutch. For example, a young girl, Raden Adjeng Kartini, wrote in Dutch a remarkable series of letters, containing criticism of Indonesian society, that were later collected and published; and a group of young men wrote poems in Dutch, although with an Indonesian background. By about 1920, however, the Dutch government itself had decided for political reasons to discourage further development of a national literature in Dutch, and the nationalist leaders had become eager for a new literature in the native language. This common aim bore fruit in 1933, when a literary journal under the editorship of Takdir Alisjahbana appeared, containing poems and essays written by various authors in the new Malay, which they now called Indonesian. The editor himself later wrote in Indonesian a number of popular novels containing social criticism, which were imitated by other writers. During the Japanese occupation of Indonesia and Malaya, this new Indonesian literature became popular also in Malaya. The adoption of Bahasa Malay (Indonesian) as the official language of Indonesia in 1949 gave further impetus to the development of the vernacular literature in both countries. The new tradition developed after independence, and its outstanding writers in Indonesia were, in poetry, Chairil Anwar and Sitor Situmorang. Important novelists include Ananta Toer and Takdir Alisjahbana.

19th-century intellectual renaissance

The Philippines. Philippine literature had its beginnings in great epics that were handed down orally from generation to generation and sung on festive occasions. When the Philippines became part of the Spanish Empire in the 16th century, printing was introduced, and all the early published works in the vernacular (Tagalog) were of Christian religious subjects. Eventually, some individual romantic legends taken from the epics were published, but they had acquired a European flavour. An outstanding work in the early years of the 19th century was an epic romance called *Florante at Laura* by the first native writer to achieve prominence—Francisco Balagtas—who wrote in Tagalog. In the latter half of the 19th century, an intellectual renaissance coincided with the beginnings of a national movement toward freedom; writers began using Spanish, for their work was part of the nationalist propaganda. The most famous author was José Rizal, who wrote a series of brilliant social novels, beginning with *Noli me tangere* ("Touch Me Not"). Other prominent writers, all essayists, were Mariano Ponce and Rafael Palma. There were poets also—for example, José Palma, whose poem "Filipinas" was later adopted as the national anthem. After the United States had taken over the Philippines, Spanish was gradually replaced by English, and new writers began to use that language as their medium. But before a new national literature could evolve, World War II took a heavy toll of writers, and those who survived became caught up in the political changes that followed. Many still write in English—the Spanish tradition, too, remains strong—but more and more writers are turning to Tagalog for literary expression. (M.H.Au.)

Music

GENERAL CHARACTERISTICS

Society and music. *Rural and urban music.* A general musical division exists between the urban and rural areas of Southeast Asia. Urban centres comprise the islands of Java and Bali and places in Thailand, Laos, Kampuchea, and Burma, where big ensembles of gong families play for court and state ceremonies. Rural areas include other islands and remote places, where smaller ensembles and solo instruments play a simpler music for village feasts, curing ceremonies, and daily activities. In cities and towns influenced by Hindu epics such as the *Rāmāyaṇa* and *Mahābhārata,* shadow and masked plays and dances utilizing music play important communal roles, while in less urbanized areas, in lieu of musical plays, chants and songs in spirit worship and rituals are sung in exclusive surroundings—a ritual procession on the headwaters of

Borneo, a drinking ceremony in the jungles of Palawan, a feast in the uplands of Luzon.

In both regions the physical setting is usually the open air—in temple yards and courtyards, under the shade of big trees, in house and public yards, fields and clearings. Many musical instruments are made of natural products of a tropical environment, and their sounds are products of this milieu. The music of buzzers, zithers, and harps is thus akin to sounds heard in the tropical vegetation of Southeast Asia. In Bali, for example, special ways of chanting and sounds of the jew's harp (also jaw's harp) ensemble (*genggong*) imitate the croaking of frogs and noise of animals.

Relation to social institutions. Music in Southeast Asia is frequently related to ceremonies connected with religion, the state, community festivals, and family affairs. In Java, important Islāmic feasts, such as the birthday of Muḥammad or the fast of Ramaḍān, as well as animistic ceremonies marking the harvest and cycles of human life, are celebrated with shadow plays (*wayang* [*wajang*]). In Bali, the *gamelan gong* orchestra opens ceremonies and provides most of the music for temple feasts. The *gamelan selunding,* an ensemble with iron-keyed metallophones (like xylophones but with metal keys), plays ritual music, and the *gamelan angklung,* so called because it formerly included tube rattles, or *angklung,* is used to accompany long processions to symbolic baths near the river.

In Malaya the court orchestra, or *nobat,* was held almost as sacred as the powers of the sultan himself. Among the Land Dayak and Iban in Borneo, ceremonial chants are sung in feasts related to rice planting, harvesting, and honouring the omen bird *kenyalang* and other spirits.

Musicians' knowledge of the dance

The relation of music to dance and theatre. In the Thai masked play, or *khon,* dancers, chorus, soloists, and orchestra are all coordinated. The musicians know the movements of classical dance and coordinate musical phrases with dance patterns, turns, and movements. In the shadow play, or *nang sbek,* the dancer, who manipulates a leather puppet, must keep his foot movements in time with vocal recitations. During pauses in which the gong ensemble plays an interlude, the dancer must change steps accordingly. In general, when there is solo singing, the instrumental ensemble remains silent or plays only a few instruments in contrast to interludes of acrobatic shows or scenes of fighting, when the full orchestra clangs on all the instruments. In Balinese dancing, body movements, paces, and directions are dependent on drum strokes and signals from a wood block (*keprak*) and cymbals (*tjengtjeng*). The dancers generally rehearse with the musicians to know exactly when choreographic changes take place.

As theatre, the stories of *Rāmāyaṇa* and *Mahābhārata* have different musical supports, depending on the country. In Bali, *Mahābhārata* shadow plays are presented to the accompaniment of a quartet of metallophones known as *gender wayang.* In Cambodia, where the preference is for stories of the *Rāmāyaṇa* (which is called *Ramker* in Cambodia), the music is a full gong ensemble similar to the Thai *pi phat* ensemble, while in Burma, a percussion orchestra of drums and gongs in circular frames accompanies singing, dancing, and dialogues in all types of plays.

Voice quality and dramatic meaning

Musical traditions and practice. *Vocal music.* The role of the voice in music making differs from that of European music in both concept and execution. Men's and women's voices are each not divided into high and low ranges but are used for their colour qualities. In the Javanese shadow play, for example, the narrator (*dalang*) assumes many singing and speaking qualities to depict different characters and scenes. Arjuna, the chief *wayang* hero, is represented with a clear voice, speaking in a single tone. Puppets with bigger bodies are given lower, resonant voices. In Thai masked plays there is no desire to produce full open tones, as in Italian bel canto. A vocal tension accounts for shades of "nasal" singing that can be discerned in commercial recordings of Thai, Javanese, Cambodian, and Vietnamese music. In the Javanese orchestra (gamelan) the voice tries to imitate the nasality of the two-stringed fiddle (*rebab*). In Bali, a particular use of men's voices is in the *ketjak,* a ritual in which groups seated in concentric circles combine markedly pronounced syllables into pulsing rhythmic phrases. In village settings among the Kalinga of Luzon, in the Philippines, singing, speaking, or whispering of vowels is so subtle as to blur the border line between speech and song. On the Indonesian island of Flores, leader–chorus singing, with the chorus divided into two or more parts, is accompanied by a prolonged note (drone) or by a repeated melodic, rhythmic fragment (ostinato). In Borneo, or Mindanao and Luzon in the Philippines, a man or woman may sing an epic or a love song in a natural voice with little or no attempt to nasalize it. Epic singing, with long or short melodic lines, goes on for several nights, and some of the sounds are mumbled to give words and their meanings a particular shading. Further, a sensuousness in the quality of Islāmic singing is achieved through the use of shades of vowel sounds, vocal openings, and a bell-like clarity of tones.

Homogeneous and heterogeneous sound

Instrumental music. Although gong orchestras consisting of gongs, metallophones, and xylophones bind Southeast Asia into one musical cultural group, the types of ensembles and sounds they form may be classified into four areas. Java and Bali make up one unit because of their predominant use of bronze instruments in orchestras that make one homogeneous sound. Thailand, Laos, and Cambodia form another subdivision, with families of musical instruments producing heterogeneous sounds: the bronze group makes slowly decaying sounds, wooden xylophones play short sounds, and a reed blows a penetrating melody accompanied by a fourth group of cymbals, drums, and another gong. The Burmese orchestras differ from the Indonesian and Thai groups by the unique use of a row of tuned drums (sometimes called a drum circle), with sounds consisting of sharp attacks and quick-vanishing waves. The fourth area, Indonesia, Malaysia, and the Philippines, uses several types of suspended and horizontally laid gongs. These gongs produce various combinations of sounds. In Nias, an island west of Sumatra, one group of three heavy suspended gongs plays three rhythms of homogeneous sounds. Suspended gongs with a wide rim and a high knob are played alone, with another gong or with a drum on the Philippine islands of Mindanao and Palawan and the Indonesian island of Kalimantan (Borneo). Gongs laid in a row, called *kulintang,* are melody instruments accompanied by a percussion group. The most developed melodies are found in Mindanao, and the area of distribution extends to Borneo, Sumatra, and Celebes, in Indonesia. The sets of tuned gongs found throughout Southeast Asia are also called gong chimes, gong kettles, and gongs in a row.

Tonal systems. In contrast to the Western diatonic-scale system (based on seven-note scales comprised of whole and half steps) and its association with relatively "fixed" pitches, there prevails a gapped system in Southeast Asia (*i.e.,* scales containing intervals larger than a whole step) with elastic intonation. Examples include the five-tone *slendro* and the seven-tone *pelog* of Java and the seven-tone scale of Thailand. In each of these systems the distances between corresponding tones in two different sets of octaves are not exactly the same. For example, one Javanese *slendro* octave has the following intervals expressed in cents (a unit of pitch measurement; 1,200 cents make 12 semitones or one octave): 246, 241, 219, 254, 246; another has 245, 237, 234, 245, 267. In contrast, two tunings of the Western chromatic scale theoretically always have 12 semitones of 100 cents apiece.

Modal systems

Related to tonal systems are modes, which in Southeast Asia use tones of a particular scale system to form melodies. Associated with a given mode are a hierarchy of pitches, the principal and auxiliary tones, endings of melodic phrases (cadential formulas), ornaments, and the vocal line. Modes express emotions and are applied to different times of the day and night and to particular situations in stage plays. They are clearly present, with local variations, in Java, Vietnam, and Burma but are less distinct in Bali, Thailand, Laos, and Cambodia.

In rural areas a multitude of scales with mixed diatonic and gapped systems and no modes are used.

Musical time and improvisation. Musical time is generally divisible in units of two or four in urban music, but it occurs more freely and without a metric pulse in rural areas, especially in singing. Musical improvisation or the use of variations based on a melodic theme is not universal. It is essential to the playing of the *rebab* and singing in the Javanese gamelan, the tappings on the Burmese circle of drums, and the percussive playing on the *kulintang.* But, in fast playing in the Balinese gamelan, exact repetitions of patterns are necessary, for there is no time for the performer to think of alternative formulas. Similarly, the separate rhythmic patterns of five instrumental parts do not change in the gong (*gangsa*) music of the Ibaloi of Luzon. Repetition is the essence of the music.

HISTORICAL DEVELOPMENTS

Origins. *Early bamboo instruments.* The widespread use of bamboo musical instruments in practically all parts of Southeast Asia points to the antiquity of these instruments and, probably, that of the music they play. A historical citation of mouth organs and jew's harps in the Chinese *Shih Ching* ("Classic of Poetry") shows that these instruments were known in the 8th century BC. Previous to this time, other bamboo musical instruments were probably in use, just as bamboo tools were used in pre-Neolithic times.

)rones
nd
stinato

The music of pre-Neolithic types of bamboo musical instruments, such as are played in the 20th century, may be just as old as these instruments. One general feature that points to this antiquity is the widespread and frequent use of a very simple musical element: a sustained tone (drone) or repetition of one or several tones (ostinato). Sustained tones appear in the mouth organ, where one or two continuous sounds are held by one or two pipes while a melody is formed by the other pipes. Prolonged tones may also be heard in rows of flutes played by one person in Flores. One flute acts as ostinato and the rest make a melody. In group singing, an underlying held tone is common. Repetition of tones occurs in bamboo instruments (jew's harps, percussion tubes and half percussion tubes, zithers, clappers, slit drums) as well as in nonbamboo instruments. In the *kudjapi,* a two-stringed lute, one string is used for the ostinato and the other to pluck the melody. In the log drum, two players play fast rhythms of continuous sounds while another player taps improvised rhythms.

Bronze instruments in gong families of Indonesia, Thailand, and Burma employ repeated sounds acting as ostinati. A widespread and preponderant use of dronelike or repeated sounds in Southeast Asia shows that they are probably an ancient fundamental musical element.

Early bronze instruments. The earliest bronze musical instruments are kettle gongs (deep-rimmed gongs), which date back to about 300 BC and are found in Vietnam, Bali, Sumatra, Borneo, Thailand, and Burma. In Burmese gongs the use of a heavy beater for the centre and a lighter stick to strike the side denotes an opposition of a full and a tiny sound applied today also to the *babandil* and other gong ensembles in Palawan and Borneo.

Gongs that predominate in Southeast Asia are those with a boss, or central beating knob. The many varieties differ according to their shapes, chemical properties, playing position, number in a series, manner of playing, musical function, and sound. Flat gongs without a central boss are not as widely used. They are found in the hills of Thailand, Laos, Cambodia, Vietnam, in some parts of Indonesia, and in the northern Philippines and may have come to Southeast Asia either through China in the 6th century or from the Middle East.

ingling
fluences

Musical traditions. The influence of the great traditions of Asia—Indian, Chinese, Islāmic, and Khmer (Cambodian)—on native Southeast Asian music varies in different countries. From India come principally two ancient Sanskrit epics—the *Mahābhārata* and the *Rāmāyaṇa.* Deep attachment to themes from the *Rāmāyaṇa* pervades the whole Southeast Asian region, except the Philippines, where Indian influence was weakest. Musical instruments attributed to India and appearing in 9th-century reliefs at the Buddhist temple of Borobuḍur and Hindu temple of Prambanan, in Java, are bronze bells, bar zithers, cymbals, conical drums, flutes, shawms, and lutes. They may still be found in several islands of Indonesia. Khmer gong circles, stringed instruments, mouth organs, drums, and oboes still in use in rural Cambodia and Vietnam are depicted in the 12th-century ruins at Angkor Wat in Cambodia. Prehistoric lithophones, or stone chimes, excavated in Vietnam in 1949, may have been the ancestors of kettle gongs. Chinese-type musical instruments (two- and three-stringed fiddles, bells, and drums), the use of the Chinese pentatonic (five-tone) scale, and duple and quadruple time (typical Chinese metres) are used in Vietnam, Burma, Thailand, Laos, and Cambodia, Islāmic musical instruments—drums, two-stringed fiddles (*rebab*), and three-stringed lutes—may be heard in Java, while melismatic singing (many notes to one syllable), especially in Islāmic rituals, is usual among the Malay groups on Borneo.

There are also musical instruments and elements that have developed locally. The mouth organs of Borneo, Laos, and Cambodia are probable ancestors of the Chinese *sheng* and the Japanese *sho* (mouth organs). Jew's harps, tube zithers, ring flutes, buzzers, xylophones, two-stringed lutes, and various types of gongs with boss (knobbed centre) are some of the most typical instruments of Southeast Asia. A probably ancient manner of measuring flute stops in Mindanao—dividing flute segments into proportional lengths to produce the octave, fifth, and other intervals—recalls a very old Chinese account of cutting bamboo tubes into lengths that would sound these same intervals.

In general, music in Southeast Asia is a tradition taught to each succeeding generation without the use of written notation. From exclusive families of musicians in courts, gamelan music was transmitted to the people. Epic and ritual songs are learned by rote and handed down from older to younger generations. Hence, skill in instrumental music is developed by imitation and practice.

Burma. Just as today all types of Burmese plays are accompanied by the traditional Burmese orchestra, the beginnings of Burmese theatre contained a music that, like the theatre, was probably based on primitive religious rituals. Before Indian and Chinese musical influences, the inspirational source of Burmese music and dance was the miracle plays (*nibhatkhin*), which, in turn, were based on singing, dancing, and entertainment in local folk feasts that date back to antiquity. The worship of spirits (*nats*) at Chinese festivals was accompanied by women who, through song and dance, communicated with and were possessed by these spirits. Following this practice, professional entertainers taking the place of women danced, sang, and played instruments during the first *nibhatkhin.* These practices led to the dancing and singing associated with the *pwe,* a popular play for public and courtly entertainment.

Foreign musical influences came from India, China, and Thailand. Indian elements appear in musical terms, theories about scales, and in some musical instruments—oboe, double-headed drums, cymbals, and the arched harp. Chinese influence appears to be older and is apparent in the use of the pentatonic scale and such musical instruments as table zithers (related to the Chinese *ch'in*), a dragon-head lute resembling a Chinese *p'i-p'a,* and two- and three-stringed fiddles. From Thailand and the Khmer civilization of Cambodia probably came both the use of gongs in a circular frame and the dramatization of episodes from the *Rāmāyaṇa.* In the traditional orchestra for state ceremonies, for the theatre, and, formerly, for royalty, three simultaneous variations of the same theme are performed by two sets of melodic percussion—a circle of about 21 tuned drums (*saing-waing*) and a circle of about 21 tuned gongs (*kyi waing*)—and at least one oboe (*hne*) or a flute (*pulwe*). To this is added a playing of a percussion group comprising a double-headed drum (*patma*), a pair of cymbals (*la gwin*), and clappers playing a duple or a quadruple metre. In three rhythmic patterns applied by these percussion groups to specific song types, the strong beats are always marked by the clappers.

Melodic
traits

Melodies played on traditional instruments (*saing-waing,* harp, *pattala* or xylophone) are frequently broken by rests and consist of segments of two, three, or four notes that form phrases, usually of eight or 16 beats. Several phrases make up a number of verses to complete a musical rendition. Melodies, based on modes, are constructed according to the previously discussed elements usually found in the modal music of Southeast Asia. Song types exist in Burmese music and are assigned to specific modes.

The Burmese arched harp (*saung gauk*) has features that may be traced back to pre-Hittite times and the Egyptian 4th dynasty (*c.* 2613–*c.* 2494 BC). Scarcely existent outside of Burma, this instrument has undergone a renascence in the 20th century. A more popular solo instrument is a wooden xylophone *pattala.*

The following instruments may be found among ethnic groups in rural Burma: idiophones, or resonant solids—bamboo jew's harps, clappers, cymbals, wooden slit drums, bronze kettle gongs, drums; membranophones, or vibrating-membrane instruments—goblet drums; chordophones, or stringed instruments—crocodile zithers, monochords with calabash resonators, three- and four-stringed fiddles; aerophones, or wind instruments—lip-valley flutes, ring flutes, panpipes, double-reed winds, buffalo horns, and mouth organs.

Thai ensembles

Thailand, Laos, and Cambodia. Although their individual political histories differ, the music practiced in Thailand, Laos, and Cambodia is almost identical. The musical instruments and forms of this region spring from the same sources: India, the indigenous Mon-Khmer civilizations, China, and Indonesia. In Thailand, three types of orchestras, called *pi phat, kruang sai,* and *mahori,* exist. The *pi phat,* which plays for court ceremonies and theatrical presentations, uses melodic percussion (gongs in a circle, xylophones, metallophones) and a blown reed. The *kruang sai* performs in popular village affairs and combines strings (monochords, lutes, and fiddles with two and three strings) and wind instruments (oboes and flutes); while the *mahori,* as accompaniment of solo and choral singing, mixes strings (floor zithers, three-stringed fiddles, and lutes) and melodic percussion (gongs and xylophones) with the winds (flutes and oboes). All three ensembles are provided with a rhythmic group of drums, cymbals, and a gong to punctuate the melody parts. Some of the above musical instruments and their functions may best be illustrated in the *pi phat* ensemble below.

A slow-moving theme is played by gongs arranged in a circle (*khong wong yai*) with variations in smaller gongs (*khong wong lek*), two wooden xylophones (*ranat ek, ranat thum*), and two box-shaped metallophones (*ranat thong ek, ranat thong thum*). The last three pairs of instruments vary the theme by playing twice as fast or by repeating, anticipating, and revolving around it. A double-reed oboe (*pi nai*) hovers above the melodic percussion, providing the only blown sound in the ensemble. Together with the punctuating gongs and drums, the whole orchestra displays a polyphonic (many-voiced) stratification of instrumental parts, using unisons and octaves mainly in the strong beats.

A melody may be broken down into phrase units consisting of two or four measures that may be joined by four other phrase units to make a phrase block, and a given number of blocks constitutes one musical composition. Three speeds of rendition—slow, medium, fast—in either duple or quadruple time are marked by two alternating strokes in a pair of cymbals; a dampened clap marks a strong beat, and a ringing vibration denotes a weak beat.

The tuning system is made up of seven tempered (approximately equidistant) tones to an octave. But the melodies constructed out of this system use only five tones out of seven—which sound close to a Chinese pentatonic scale. This scale may be constructed in any of seven levels or tones of the Thai tuning system. Further, through a process called *metabole,* melodies may move from one level to another.

Shadow-play music and dance

In the Cambodian shadow play (*nang sbek*) two narrators alternate in chanted recitative to explain the role of the leather puppets. Dancers parading these figures across the screen and simulating their actions are accompanied by an orchestra. A limited number of tunes is played to eight dance positions (walk, flight or military march, combat, meditation, sorrow or pain, promenade, reunion, and metamorphosis). In the play these poses are assumed by princes, princesses, monkeys, demons, peasants, or ascetics.

Among different ethnic groups, such as the Khmer Saoch, Pwo Karen, Bu Nuer, Kae Lisu, Kuoy, and Samre, a rural music related to that of the ancient Khmer peoples is played by aerophones (buffalo horns, mouth organs, vertical flutes), idiophones (flat gongs, gongs with boss, cymbals, jew's harps), chordophones (bamboo zithers), and membranophones (circle of drums). Other important instruments for solo performance or as accompaniment to songs are the three-stringed crocodile zither (*chakhe*), a four-stringed lute (*grajappi*), a plucked monochord with a gourd resonator (*phin nam tao*), and a bamboo whistle flute (*khlui*).

Traditiona court music

Vietnam. Although Vietnamese music belongs to the great Chinese musical tradition, which includes the music of Korea, Mongolia, and Japan, some of its musical elements are indigenous or come from other parts of Southeast Asia, and some derive from Champa, an ancient Hinduized kingdom of Vietnam. Archaeological finds in the village of Dong Son revealed that the ancient Vietnamese used kettle gongs, mouth organs, wooden clappers, and the conch trumpet. From the 10th to the 15th century a joint Indian and Chinese element left its musical imprint. The Chinese seven-stringed zither (*ch'in*) and a double-headed drum were played together, or a Champa melody was accompanied by a drum. It was at this time that two traditional Chinese ensembles—Great Music and Little Music—and an elementary Chinese theatrical art were introduced. From the 15th to the 18th century the Chinese influence reached its height. Court music (*nha nhac*) was played by two orchestras. One, located in the Upper Hall of the court, consisted of a chime of 12 stones, a series of 12 bells, a zither of 25 strings (Chinese *se*), a zither with seven strings (Chinese *ch'in*), flutes, panpipes, a scraper in the shape of a tiger, a double-headed drum, a mouth organ, and a globular whistle. The second orchestra in the Lower Hall used 16 iron chimes, a harp with 20 strings, a lute with four strings (Chinese *p'i-p'a*), a double flute, a double-headed drum, and a mouth organ. Ceremonial music, almost nonexistent in the 20th century, was patterned after court music.

In Buddhist ceremonies, prayers were recited in three ways: as recitation in a low voice, as a cantillation (sung, inflected recitation) following the six tones of the Vietnamese language, and as chant accompanied by an orchestra of two drums, bell, gong, cymbals, and fiddles.

Music as entertainment is mostly a vocal art played without ritual outside the court and still enjoyed by many people. The *hat a dao* found in the north is the oldest form. It is a woman's art song with different instrumental accompaniments, dances, a varied repertoire, and a long history of evolution.

From the 19th century to World War II, Vietnamese music reaffirmed its character. Although the playing of court music was restricted, popular music was encouraged, leading to northern and southern styles that were patronized by both the aristocracy and commoners. Western musical influence in this period was manifest in the use of the mandolin, the Spanish guitar, and the violin, as well as by the introduction of European classical music and composition following Occidental forms. In the late 20th century traditional Vietnamese music began to disappear, but attempts to revive it began in the early 1970s.

Vietnamese rural folk music is built on the same musical principles as court music. The main difference lies in its application to village activities—work, games, courting, marriage, cure for the sick, entertainment, feasts.

Elements of Vietnamese music

Common elements characterize and unify all Vietnamese music. It is based on an oral tradition, with written notation serving only as a reading guide. Melodies are generally built out of a pentatonic system (for example, C, D, F, G, A) to which two auxiliary tones (E, B) may be added to make other pentatonic melodies. A song, usually preceded by a prelude, may be sung in slow, moderate, or fast tempo divisible by two or four, with a simple contrapuntal (countermelody) accompaniment using unisons and octaves at beginning points of phrases. Outside of the first beats, intervals of fifths, fourths, thirds, and even seconds are allowed. An important aspect of melodies is the idea of mode (*dieu*), the elements of which do not essentially differ from those of Javanese and Burmese music.

ıstrument
roups
f the
ımelan

Indonesia and Malaysia. *Java.* A Javanese philosophical concept based on mysticism, refinement (*halus*), and the inner life as related to Hindu, Islāmic, and Indonesian thought may best be represented in music by the Javanese gamelan, an orchestra made up mostly of bronze instruments producing homogeneous blended sounds. The instruments in the ensemble may be divided into three groups of musical function. The first group comprises thick bronze slabs (*saron demung, saron barung, saron panerus*) on trough resonators playing the theme usually in regular note values without ornamentation. The second group consists of elaborating or panerusan instruments, which add ornaments to the main theme. In this group gongs in double rows (*bonang panembang, bonang barung, bonang panerus*) play variations with the same ratio of speed as the *saron* group. In softer sounding music for indoor performance, other panerusan instruments with very mellow sounds come in. These are three sizes of thin bronze slabs with bamboo resonators—*gender panembung* or *slentem, gender barung,* and *gender panerus.* Other elaborating instruments are the wooden xylophone (*gambang*), the zither (*tjelempung*) with 26 strings tuned in pairs, an end-blown flute (*suling*), and a two-stringed lute (called a *rebab* by the Javanese), which leads the orchestra. In loud-sounding music, the soft-sounding instruments are not played, and the drum (*kendang*) leads the orchestra. The third group provides "colotomic," or punctuating beats in four rhythmic patterns played separately by four types of heavy, suspended, or horizontally laid gongs.

:ales and
ıodes

Two tuning systems prevail. The *slendro* tends to have five equidistant but flexible (or varying) pitches in an octave, while the *pelog,* with seven equally flexible tones, has a more varied structure. One tuning with intervals expressed in cents (140, 143, 275, 127, 116, 204, 222) may roughly be represented by the following notes in a descending scale: C↑, A♯, G♯, G↓, F↑, D♯↓, C♯↑, and C. (Arrows up are tones slightly higher than Western tempered tuning [in which a semitone is equivalent to 100 cents] and vice versa for arrows down.) Melodies from these tunings are governed by a modal structure (*patet*) the elements of which are similar to those of Vietnamese and Burmese music.

In West Java the most popular ensembles use a vocal part, a two-stringed fiddle (*rebab*) or a bamboo flute (*suling*), and a zither (*kachapi*). In the gamelan, submodes (*surupan*) are formed by the use of vocal tones—sung or played on the *suling* or *rebab*—which amplify the number of scales in both the *pelog* and *slendro* systems.

ırieties
Balinese
melan

Bali. In contrast to the introspection of Javanese music, the Balinese gamelan exudes a music of brilliant sounds with syncopations (displaced accents) and sudden changes, as well as gradual increase and decrease in volume and speed and feats of fast, precise playing. The tuning system, musical instruments, and polyphonic stratification are similar to those of the Javanese gamelan, although in Bali the seven-tone *pelog* is not popular. Most gamelan are tuned to a five- or four-tone system, and the concept of modes is not as clearly developed as in Java. A variety of gamelan exists, each with a special function, instrumentation, repertoire, and tuning system. The *gamelan gong* orchestra is among the most extensive in its number of instruments. A modern version, *gong kebjar,* omits the *trompong* (gongs in a row) and *saron* (bronze slabs over a trough resonator) and replaces them with *gangsa gantung* (metallophone with bamboo resonators) and *rejong* of four gongs to produce exuberant outbursts of sound. The *gamelan gambuh,* now rare, comprises four end-blown flutes, one *rebab,* and a group of percussion. The *gamelan semar pegulingan,* played formerly in royal courts but now almost disappeared, emphasizes the *trompong* as a solo instrument. The *gamelan pelegongan* is a virtuoso orchestra that accompanies *legong* dances, while the *gamelan pedjogedan* is an orchestra of xylophones for dance (*djoged*) and entertainment in the marketplace. The *gender wayang* is a quartet of *slendro* tuned metallophones specially employed for shadow plays. The *gamelan angklung,* a village orchestra assembled during ceremonies, anniversaries, and cremations, originally consisted of rattling tubes that are now replaced by metallophones. The *gamelan ardja* is characterized by a soft timbre (tone colour) and the use of a one-stringed bamboo zither, the *guntang,* to accompany musical comedy and popular plays.

Other parts of Indonesia. In the islands of Flores, Nias, New Guinea, Celebes, and Borneo idiophones make up perhaps the most varied collection of musical instruments—gongs of various profiles, slit drums, jew's harps pulled with a string, clappers, bells, xylophones, percussion sticks, bull-roarers, and stamping tubes. Particularly interesting are idiophones made of bones, shells, skulls, fruits, seeds, planks, pellets, crab claws, clogs, coconut, and shark bones. Membranophones are represented by drums shaped like a cylinder, goblet, vase, round frame, hourglass, cone, cup, barrel, or a tube. Aerophones present an array of vertical and transverse (horizontally played) flutes, panpipes, ring flutes, shawms, clarinets, gourd trumpets, conch shells, ocarinas, and flutes with different mouthpieces. Chordophones include bamboo zithers, spike fiddles (in which the neck skewers the body), one- and two-stringed lutes, musical bows, monochords, guitars, *rebabs*, bar zithers, and sago zithers. In Flores, part singing with a sustained drone is frequent. Songs in Nias use diatonic (whole and half steps), chromatic (half steps), and gapped melodies largely less than an octave in range. In Sarawak descending melodies make up a tetrachord (four adjacent tones forming the interval of a fourth). In Indonesian New Guinea departures from songs with gapped scales include fanfare, stair descent, and tiled melodies (the last consisting of short phrases repeated at different pitch levels).

Malaysia. At least three principal cultural influences—Indonesian, Hindu, and Islāmic—left their musical marks in Malaysia. The Indonesian influence is seen principally in musical forms, participants, and paraphernalia of the Malaysian shadow play (*wayang kulit*). It is said that the Indian epics and, especially, the *Pandji* tales of Java came to Malaysia via Indonesia, but there are songs in certain plays and musical instruments (*e.g.,* the double-headed drum and oboe) that could have reached Malaysia from India through other routes. Islāmic traces are evident in melismatic songs among the Malay groups in songs connected with religious rituals and in choral singing in the *ma'yong* plays. Chinese music, a more recent development, is largely practiced among the Chinese communities, principally in Singapore.

Before Malaysian independence, the *nobat,* an old royal instrumental ensemble dating back to about the 16th century, played exclusively for important court ceremonies in the palaces of the sultans of Perak, Kedah, Selangor, and Trengganu. Today, in Kedah, the ensemble consists of five instruments: one big goblet drum (*negara*), two double-headed drums (*gendang*), one long oboe (*nafiri*), one small oboe (*nafiri*), and one gong. The music, which consists of ten surviving pieces, is broadcast today and performed live.

Three shadow plays exist, principally in the state of Kelantan. The *wayang gedek* is the Thai form; *wayang Djawa,* a Malay form, is almost extinct; and the *wayang Siam,* which is a combination of Thai and Malay influences, is the most popular form of puppet shadow play. The operator of the performance is the narrator (*dalang*), who manipulates the leather figures, introduces important characters, and describes different scenes with the accompaniment of the orchestra. The music is led by a two-stringed lute (*rebab*) in the *Rāmāyaṇa,* or an oboe (*serunai*) in *Mahābhārata* and *Pandji* cycles. The melodic instruments are supported by a percussion group consisting of pairs of goblet-shaped drums (*gedombak*), cylindrical drums (*gendang*), barrel drums (*geduk*), gongs lying on a support (*chanang*), suspended gongs (*gong*) or, sometimes, a row of gongs played by two or three men, and one pair of cymbals (*kesi*). The music usually begins with a prelude followed by a list of pieces the sequences of which are dictated by the narrator.

The *ma'yong,* a dance drama that probably dates back to more than 1,000 years, was introduced in Kelantan under the patronage of the royal courts. In the 20th century it exists as a folk theatre with an all-female cast. The music that accompanies 12 surviving stories is played by an orchestra of one bowed lute (*rebab*), two suspended gongs, and a pair of double-headed drums (*gendang*). A heterophony (simultaneous variation of the same melody) between a solo voice, a chorus, and the *rebab* creates a music with a Middle Eastern flavour.

Variety in instruments

A rich musical heritage in the rural sections of Malaya is shown in musical instruments used by Malay, Thai, Semang, and Sakai groups. Idiophones include shell and coconut rattles, the jew's harp (mostly pulled by a string, rather than plucked), bull-roarers, bamboo clappers, and the bamboo slit drum. Aerophones include the buffalo horn, wooden and clay whistles, nose flutes, end-blown flutes, and the oboe. Chordophones are two- and three-stringed fiddles with coconut resonators, monochords, and tube zithers. One membranophone is a double-headed cylindrical drum.

In Borneo among the Malay, Kadazan, and Iban groups, the principal instruments are gongs in a row (*gulintangan*) played with suspended gongs of different types (*chanang, gong, tawak-tawak*). Among the Murut, Kenyah, and Iban the mouth organ with a calabash resonator (*sompoton*) plays a melody with a drone accompaniment. The jew's harp (*ruding*), bamboo zither (*tongkungon*), nose flute (*tuali*), hourglass drum (*ketubong*), and vertical flute (*suling*) may be heard among different ethnic groups. Iban ceremonial songs are sung in connection with rice festivals and rituals to prevent sickness, while mourning songs make up a rich repertoire of solo and leader–chorus singing. The Kenyah are particularly adept at blending low voices of men singing a melody supported by a drone.

The Philippines. Two musical cultures—Western and Southeast Asian—prevail in the Philippines. Western music is practiced by about 90 percent of the population, while Southeast Asian examples are heard only in mountain and inland regions, among about 10 percent of the people.

The Western tradition dates back to the 17th century, when the first Spanish friars taught plainchant and musical theory and introduced such European musical instruments as the flute, oboe, guitar, and harp. There subsequently arose a new music related to Christian practices but not connected with the liturgy. Processional songs, hymns in honour of the Blessed Virgin, Easter songs, and songs for May (Mary's month) are still sung in different sections of the country. A secular music tradition also developed. Guitars, string ensembles (*rondalla*), flute, drum, harps, and brass bands flourished in the provinces among the principal linguistic groups and still appear during town fiestas and important gatherings. Competing bands played overtures of Italian operas, marches, and light music. Young men, like their counterparts throughout the Hispanic world, sang love songs (*kundiman*) in nightly serenades beneath the windows of their ladies. It was not uncommon in family gatherings for someone to be asked to sing an aria, play the harp, or declaim a poem. Orchestral music accompanied operas and operettas (*zarzuelas*), while solo recitals and concerts were organized in clubs or music associations. With the advent of formal music instruction in schools, performance and composition rose to professional levels. In the 20th century several symphony orchestras, choral groups, ballet companies, and instrumental ensembles performed with varying regularity.

Gong ensembles

A Southeast Asian musical tradition exists completely apart from the Western tradition. In the north, flat gongs are played in different instrumental combinations (six gongs; two gongs, two drums and a pair of sticks; three gongs). In the ensemble with six gongs, four are treated as "melody" instruments, one as ostinato, and another as a freer layer of improvisation. The melody consists of scattered tones produced by strokes, slaps, and slides of the hands against the flat side of the gong. Other musical instruments in the northern Philippines are bamboo. These are the nose flute (*kalleleng*), lip-valley or notched flute (*paldong*), whistle flute (*olimong*), panpipes (*diwdiwas*), buzzer (*balingbing*), half-tube percussion (*palangug*), stamping tube (*tongatong*), tube zither (*kolitong*), and jew's harp (*giwong*). Leader–chorus singing among the Ibaloi is smooth and sung freely without a metric beat, while the same form among the Bontoc is emphatic, loud, and metric. Scales in songs and musical instruments use from two to several tones within and beyond an octave and are arranged as gapped, diatonic, and pentatonic varieties.

In the southern Philippines (the islands of Mindanao and Sulu), the more developed ensemble is the *kulintang,* which consists of eight gongs in a row as melody instruments accompanied by three other gong types (a wide-rimmed pair; two narrow-rimmed pairs; one with turned-in rim) and a cylindrical drum. The *kulintang* scale is made up of flexible tones with combinations of wide and narrow gaps sometimes approaching a Chinese pentatonic variety and oftentimes not. Its melody is built on nuclear tones consisting of two, three, or more tones to form a phrase. Several phrases may be built, repeated, and elongated to complete one rendition lasting two to three minutes. Pieces of music are played continuously for a long period during the night.

In the central west Philippines on the island of Mindoro, love songs are sung that are based on reciting tones with interludes played by a miniature copy of the Western guitar or a small violin with three strings played like a cello. (Jé.Ma.)

The performing arts

In variety of dance and theatrical forms and in the number of performing groups, no area in the world except India and Pakistan compares to Southeast Asia. Some form of the performing arts is a normal part of life throughout the several nations. Sophisticated performing groups cluster in and around the present and former court cities—Jogjakarta and Surakarta in Java, Ubud and Gianjar in Bali, Bangkok in Thailand, Mandalay in Burma, Siem Reap near Angkor and Phnom Penh in Kampuchea (Cambodia), Hue in Vietnam—where drama, puppetry, dance, and music have been cultivated for ten centuries or more. Hundreds of commercial theatrical and dance groups perform in such newer centres as Rangoon, Saigon, and Jakarta and in scores of provincial cities and towns. Wandering troupes of actors, puppeteers, singers, and dancers travel from village to village in areas adjacent to these population centres. There are few communities in which some form of folk dance is not performed by local people.

Richness c form and experience

In the West, music, dance, and drama are usually separate arts, whereas in all areas of Southeast Asia, drama, dance, mime, music, song, and narrative are integrated into composite forms, often with masks or in the form of puppetry. The spectator's senses, emotions, and intellect are bombarded simultaneously with colour, movement, and sound. The result is a richness and a vividness in the theatre that is absent in most Western drama, so much of which rests on a literary basis.

More than 100 distinct forms or genres of performing arts can be distinguished in Southeast Asia. These can be grouped, according to which of the various stage arts is emphasized, into (1) masked dance and masked dance-mime, (2) unmasked dance and dance-drama, (3) drama with music and dance, (4) opera, (5) shadow-puppet plays, and (6) doll- or stick-puppet plays.

DIVERSE TRADITIONS IN THE PERFORMING ARTS

Four relatively distinct traditions exist in the performing arts: folk, court, popular, and Western.

The folk tradition. Dances in the folk tradition are exceptionally numerous and widespread. Some are performed as religious ritual, others, particularly on the Indonesian island of Bali, by highly trained and respected artists, and still another kind as entertainment in which the community participates. Folk theatre is more complex than folk dance and thus less widespread, but it has deep connections with religious ritual. Although the origins of most folk performing arts lie in remote times, later court forms exerted important influence on many of the folk forms. Conversely, folk forms have been a source of inspiration to court artists.

Shadow plays

The court tradition. The shadow play and masked and unmasked dance are court arts reflecting centuries of subtle refinement under the patronage of kings and princes. In Southeast Asia the shadow theatre is a major classic art. Leather puppets of mythological figures, the bodies intricately incised to allow light to pass through, are attached to sticks for manipulation. A lacy shadow is created by a flaming lamp as the puppet is pressed against the back of a vertical screen of white cloth. The flickering and insubstantial shadow seen from the other side creates for the understanding viewer a mystic world with deep symbolic meaning. In Java, Bali, Malaysia, Cambodia, and Thailand shadow plays and their techniques have been emulated by human actors and dancers and have been the models for marionette and doll-puppet theatre.

Dance troupes have been a part of court life at least since recorded history began. In the mainland courts of Cambodia, Thailand, Laos, and Burma, concubines of the ruler's harem who performed female dances were segregated from male performers, giving rise to separate forms of female unmasked dance and male masked dance-mime. Although certain dances traditionally are performed only by men or only by women in Indonesia and Vietnam, mixed casts have a long history, especially in dramatic pieces. Court dance on the mainland and in Indonesia has been influenced by Indian dance style, and Vietnamese dance by the dance styles of Chinese opera, but they have acquired a distinctly Southeast Asian character. Court dance reached its greatest development when applied to mythological and legendary themes, often taken from the shadow theatre. The resulting dance-dramas and masked dance-mimes of Thailand, Cambodia, and Java are world famous for their magnificent scale and elegance of execution. Some of these court arts are no longer performed, and others face increasing difficulty securing financial support, yet they remain important.

The popular and Western traditions. In the popular traditions are those 400 to 500 professional troupes who perform, except in the Philippines, in commercial theatre buildings of major cities for an urban ticket-buying audience. Some forms of popular theatre are directly modelled on court dance-drama, but most are spoken drama in which court-derived music, song, and dance movements have been inserted. Local legend and history provide the subject matter for many of these plays. As in much of Asia, the performer in the popular tradition is seldom accorded status and may be despised as a vagabond.

The spoken drama, the ballet, and the modern dances are known only superficially in Southeast Asia. The sole exception is the Philippines, where amateur performances of Western plays constitute the country's main theatrical tradition. Southeast Asian audiences generally find Western plays based mainly on dialogue to be uninteresting and deficient in artistic qualities. European and American films and television programs, however, are widely shown and appreciated, and popular Western dances are found in major urban areas. Undoubtedly the impact of these forms on local audiences will continue to increase, possibly to the detriment of the indigenous traditions.

CHARACTERISTICS OF DANCE

Dramatic and nondramatic forms. In the parts of Southeast Asia influenced by Indian forms—everywhere except for Vietnam and the Philippines—nondramatic and dramatic dance are both known. Nondramatic, or "pure," dances that do not express emotional states of characters are numerous in both folk and court traditions. Among court dances, the Javanese *bedaja* is typical. Nine dancers move in unison, without emotional expression, in precisely fixed choreographic patterns designed to demonstrate sheer grace of movement. The *maebot,* composed as a Thai "alphabet of dance," is used to train pupils in the basic movements of court dance. Other dances that include character impersonation yet are not explicitly storytelling dances lie between nondramatic and dramatic dance. In the Thai *praleng,* two performers wearing god masks and holding peacock feathers in both hands perform an offertory dance to the god before the main dance-play begins. The Balinese *legong,* danced by a pair of preadolescent girls, may have only the most tenuous dramatic content. Its interest lies in the girls' unison rapid foot movements and fluttering movements of eyes and hands. Dramatic dance is seen at its best in full dance-dramas and in the excerpts from them that are sometimes danced in concert form.

se of eyes
d hands

Styles and conventions of movement and costuming. General characteristics of both dramatic and nondramatic dance are (1) slowness of tempo except in battle scenes, (2) controlled and reserved movements rather than expansive ones, (3) little of the leaping typical of Western ballet but, instead, a feeling of closeness to the ground, and (4) extensive use of arm and hand gestures. From Indian dance has come an open and flexed position of the legs, a side-to-side sliding movement of the head and neck, and a rigidly codified vocabulary of hand and finger gestures known as *mudrās* or *hastās* in India. In most cases the Indian elements have been altered greatly over their 1,000-year period of assimilation. In Thai, Cambodian, and Lao dance, the 24 to 32 Indian *mudrās* have been reduced to nine; in Javanese dance seven can be recognized, and in Bali only one or two. They have also been altered in their shape, and the many specific meanings attached to each in India have become fewer, while in some cases a gesture has no specific meaning. Such hand gestures as shading the eyes and tying the sash, which appear in Javanese dances, are unknown in India. Foot movements in India typically follow the rhythm of a drum, often with vigorous stamping sounds that are emphasized by bells on the ankles, but such movements are virtually absent in Southeast Asia. The exaggerated eye, eyebrow, cheek, mouth, and chin movements through which the Indian dancer expresses a broad gamut of emotions are nowhere to be seen. Balinese dancers use darting eye movements, but the court dancer's face is composed into an almost unchanging expression of aloof gentility.

Close contact between neighbouring countries has led to the development of two regional Indian-influenced dance styles, one for Thailand, Cambodia, Laos, and Burma and one for Indonesia and Malaysia. Characteristics of the former style include the soft *pi phat* music of bamboo xylophones, drums, gongs, and oboe as accompaniment, bent-back finger positions not seen elsewhere in Asia, similar and often identical movements for male and female roles, courtship dances in which lovers touch each other and move in unison, and, in dance-drama, lengthy pure-dance pieces inserted solely for their beauty. In the latter style, the performance is accompanied by music of the gongs and metal bars of the gamelan orchestra. Scarves draped from the waist or neck are flicked for effect and manipulated to indicate strength or flying, and male and female dance are clearly distinguished by the powerful masculine lunges of the men and the tiny steps of the women, who also dexterously manipulate the train of the skirt with their feet. Visually, the mainland dance sparkles. Costumes of brilliant silk are covered with sequins and even jewels, and golden crowns and sparkling body ornaments glitter with reflected light. The male dancer in Indonesia wears a soft batik skirt of brown and white, the female a black velvet bodice. Arms and shoulders are bare and powdered golden brown, creating a subdued and warm effect.

Courtship
dances

The main style in Vietnam, apart from folk dance, is dramatic and highly pantomimic, like the movements of Chinese opera. In classical opera, the flowing white sleeves and the pheasant feathers bobbing from the general's headdress are twirled and flicked by the actor in many conventionalized movements derived from Chinese forms. Battle scenes are choreographed into precise dance patterns, but the acrobatic movements common in Chinese opera are seldom seen.

CHARACTERISTICS OF DRAMA

Thematic origins and materials. Most traditional plays and dramatic dances are derived from mythological and legendary sources. The tribal epics that relate the origin of the Ifugao and the Bicolano peoples in the Philippines and a number of animistic stories in Indonesian shadow theatre are indigenous myths of great age, while the widely used, romantic *Pandji* cycle from Java and the Thai *King Abhai Mani* and *Khun Chang Khun Phan* are more recent local legends. The most important dramatic sources, however, are borrowed from the Indian *Rāmāyaṇa* and *Mahābhārata* epics, from the *Jātaka* Buddhist birth stories, from Chinese novels (such as *The Romance of the Three Kingdoms*) and Chinese operas, and from a host of Islāmic stories, including the *Thousand and One Nights* and the Amīr Ḥamzah tales. These foreign stories are turned into local legends. For example, the Indian Prince Rāma becomes a Thai, a Balinese, or a Javanese prince, embodying the heroic traits admired in each of these countries.

Theme of the restoration of earthly harmony

Plays are invariably extensive and have many scenes. It is not unusual for a play to present action over several generations, an indication of the value placed on cultural continuity. A recurring theme concerns restoration of harmony on earth by a ruler acting in accord with divine law. A kingdom is restored, a prince unjustly exiled returns to assume his throne, a usurper is punished, or the prosperity of the land is assured by consummating a particularly desirable marriage. As in Western drama, the hero gains his ends through struggle. Because he acts as the human representative on earth of the known cosmic will, however, his actions exhibit a natural sweetness and serenity, even in the midst of violence, that is foreign to Western drama. Meditation is often the means whereby the hero gains the power to achieve his goal. In more recent plays based on local history and on contemporary events, the assumption of cosmic harmony has been muted, and emphasis has shifted to depicting human conflicts—nationalist versus Western colonialist, modern daughter versus conservative parents, for example—that may or may not resolve happily.

Characters. Gods, demigods, kings descended from the gods, and princes and princesses are the heroes and heroines of traditional drama and dance. Powerful religious seers advise them, allies and ministers serve them, crude foreign ogres oppose them, and grotesque, slapstick clown-servants are their attendants. The clowns have been the subject of much speculation. Like the *vidūṣaka* clown of Indian Sanskrit drama, they are gluttons, practical and even cynical, and confidants to their masters' passions and weaknesses. Scholars have theorized that the chief Javanese clown figure, Semar, is derived from an ancient Javanese god who was deposed from his supreme position by the introduction into the drama of the later Hindu gods. In the midst of mythological plays, the clowns comment irreverently on political or social issues of the day, seemingly as spokesmen for the common man in an otherwise aristocratic world. Comic and serious scenes alternate.

Improvisation by variation

Dramatic materials. A written script may be used as the starting point for performance, but usually actors, dancers, musicians, and stage crew improvise from a brief scenario. Specific musical selections are matched to certain kinds of scenes, characters, or actions, and standard movements for entrances and exits are known. Standard descriptive phrases of the kind common in all oral literature are used to introduce the hero and his kingdom, and more than a dozen types of recurring scenes are identifiable. A major interest in playgoing lies in perceiving the skill with which performers rearrange and subtly vary these familiar elements from play to play. Narrative commentary accompanying the dances often interprets a specific action in its broad context, thus helping to universalize the theatrical experience.

Costumes, makeup, and settings. Costume and makeup have great importance in plays and dances. By means of elaborate systems of changing the cut, colour, and ornamentation of costume, the shape of the hairdress, the configuration of the crown, or the facial delineation and colour of masks, at least 300 different dance and dramatic characters can be identified. Doll- and shadow-puppet figures are carved according to similarly elaborate means of identification. Persons familiar with a dance or theatrical form can identify most characters by name or by type. Costumes, masks, and puppets may be works of art highly prized in themselves. Court and folk performances once used no scenery at all. Canvas scenery depicting stock scenes is now used by most popular troupes, but unfortunately it is often as inartistic as it is inexpensive. Only the Thai National Theatre, major troupes performing the popular *cai luong* drama in Vietnam, and troupes performing in the Western tradition throughout Southeast Asia attempt to design three-dimensional scenery for each play.

ORIGINS AND DEVELOPMENT OF THE PERFORMING ARTS

Prehistory and links to the present. Knowledge of prehistoric performing arts is necessarily slight. That the performing arts were known and apparently widely practiced by the prehistoric peoples who had settled the mainland and the island archipelagoes is suggested by large bronze drums cast before the time of Christ, numerous pre-Hindu tribal myths in remote areas of the Philippines and elsewhere, masked dances of many types still performed by isolated tribes in Kalimantan (Borneo) and in New Guinea, and descriptions of music and dance by Chinese visitors beginning as early as the 1st century AD. Simple dances were almost certainly accompanied by rhythmic percussion sounds and probably by the tuned metal bars or gongs thought to be indigenous to Southeast Asia. Some scholars suggest that tribal ancestors, animistic spirits, and animals were represented, perhaps in shadow form. Whatever their nature, these were folk performances, in part religious rites connected with seasonal festivals and in part joyful entertainment.

The *trott*, *barong*, and *nat*

A number of existing dances and dramatic forms show prehistoric links. In the *trott*, a Cambodian deer-hunting dance, masked dancers representing hunter, demon, bull, girls, and deer enact the ritual of a deer hunt to ensure its success in real life. The Dayak of Kalimantan perform a dance to exorcise sickness. The *barong* dance-drama of Bali is staged by a village in which malicious spiritual forces are believed to have gained dominance over protective ones. By enacting the stand-off battle between the protective Barong lion figure and the destructive Rangda witch figure, the village ritually restores an equilibrium between the contending forces. A local *nat*, or animistic spirit, of which there are 37 in Burma, can be invoked by the dance of a professional "spirit wife," or *natkadaw*, through whom the *nat* communicates with the living. A disputed theory holds that the shadow play began as a ritual in which the spirits of magically powerful tribal ancestors were called to earth, in their natural form as shadows or shades, for advice.

Spreading of styles. Between *c.* AD 100 and 1000, dance and drama in Southeast Asia were profoundly affected by

Josephine Powell, Rome

Apsaras, heavenly dancing girls, bas-relief from Angkor Wat, Angkor, Cambodia, early 12th century.

the introduction of dance style and the vast Hindu historical epics of India. First in Cambodia, then in turn in Thailand, Laos, and Burma, the epic *Rāmāyaṇa* became the source of dance and shadow plays. In Java the *Mahābhārata* dominated, whereas in Bali and Malaysia both epics were popular. Indian influence, however, can be exaggerated. There is no evidence that Sanskrit play texts or written dramatic treatises such as the *Nātya-Śāstra* became known. Strong local performing traditions made it possible to assimilate elements of Indian dance and Hindu stories, and, in subsequent development, Southeast Asian dance and theatre grew ever farther away from Indian styles.

Copper inscriptions from Java identify clowns, actors, musicians, and possibly puppeteers in the 9th century, and epic literature of succeeding centuries contains numerous descriptions of shadow plays that were popular and emotionally gripping. By at least the 4th century, epic recitations were a part of the Brahmanic worship of ancient Cambodia. Carvings of the beautiful *apsaras*, or heavenly dancing girls, adorning the temples of Angkor attest to the importance of court dance in Cambodia between the 10th and 13th centuries.

igration
the
adow
ay

Accidents of history often carried the performing arts across national boundaries. It is believed King Jayavarman II brought dancers and musicians from Java when he left there in 802 to establish the Khmer dynasty in Cambodia, and shadow puppeteers may have accompanied him as well. Another theory suggests that Cambodia received the shadow play from India by way of Malaysia, through conquest by a Malay prince in 1002. Accidents of war took Khmer dance (and perhaps shadow theatre) first to Laos, when in 1353 a prince who had been raised at Angkor established an independent Lao court at Luang Prabang. Next, it reached the Thai capital at Ayutthaya in 1431, when Angkor fell to invading Thai armies. These returned to their court with the Cambodian court-dance troupe, thereby beginning the traditions of Thai court dance and dance-drama. In 1767 the Thai court was captured, in turn, by the Burmese, who brought to Burma the Thai-modified Khmer dance and created Burmese court drama. By this time, also, Javanese shadow theatre had been taken by colonists to Bali and to Malaysia, from whence it later entered southern Thailand.

When Indonesia was converted to Islām and Chinese influence became strong in the northern tier of mainland states beginning in the 13th and 14th centuries, existing court dance and dramatic forms were scarcely affected. Instead, new Islāmic plays were devised in Indonesia and Malaysia for shadow presentation and for the doll-puppet theatre. Islāmic influence was very strong in Malaysia, however, and even such pre-Islāmic forms as the shadow play absorbed Islāmic prayers, characters, and themes. Bali was never converted to Islām, and its performing arts are thought to reflect, even today, an older tradition than is seen in Java.

Chinese performing arts came to dominate Vietnam during the 1,000-year rule of northern Vietnam by the Chinese. Long after the Chinese were expelled, Vietnamese kings patterned their dances and opera on Chinese models. In time, however, local Vietnamese melodies and stories took their place alongside those of Chinese origin; and play scripts, at first filled with Chinese loan words, were rewritten in more colloquial Vietnamese.

Popular theatre and Western rule. From the 19th century onward, the incursion of Western culture brought about a variety of developments. A steady decline in the power of the royal courts precipitated the death of court drama in Burma; the shifting of support for dance and drama from the court to national bureaus of education and culture in Thailand, Cambodia, and Vietnam; and the movement of the court dance-drama into the popular theatre tradition in Java. In every country, new popular forms of theatre were created. These were based on historical events, on Islāmic and Chinese stories (but romances rather than Hindu and Buddhist myths), on national heroes fighting colonial rule, and on stories about contemporary events. It was not Western drama that sparked the burgeoning of popular theatre, though these plays were largely spoken dramas interspersed with music and dances. Rather, it was more of an indirect response to colonial rule, which caused an upsurge of nationalist feelings, and to the rapid growth of cities that created large populations without access to either folk or court theatre yet eager for some form of entertainment.

DIVERSE NATIONAL FORMS AND TRADITIONS

Although most of the dance and dramatic forms of Southeast Asia are related at least in the distant past, except in Vietnam and the Philippines, they acquired a very distinctive national and local character over the centuries. An examination of a few of these myriad forms will provide a more precise picture of the dense texture of the performing arts in Southeast Asia.

Cambodia. Court performing arts that had flourished during the Angkor period (802–1431) almost ceased in the centuries following the fall of the Khmer dynasty. Whether there was an organized court life or not is uncertain because of the scarcity of records, but in the 18th and 19th centuries performances in Thai form were produced by the Thai rulers of the western provinces of Cambodia. At Phnom Penh a classical ballet troupe was established by the royal family in the 19th century.

Puppet
tableaus

Court styles. The chief court forms are *nang sbek* shadow theatre, *lakon* female dance and dance-drama, and *lakon kawl* male masked pantomime. The puppets of *nang sbek* stand four to five feet in height, have no movable arms, and are manipulated from beneath by two fixed handles or sticks. The standing puppeteer either sways the puppet with his arms or he dances with it. In processional scenes, as many as ten puppeteers parade completely around the screen, front and back. An entire tableau may be carved on one puppet, including several figures, forest scenery, or palace buildings, as if to bring to life the epic scenes carved in relief on the temples of Angkor Wat. Two narrators alternate a slow chant with dialogue. During dance sections, the large *pi phat* ensemble, augmented by a large drum, is played. Only plays based on the *Rāmāyaṇa* are performed, and major puppet figures represent Rāma, his consort Sītā, the monkey Hanumān, and Rāvaṇa, a ten-headed demon king who kidnaps Sītā, Khmer peasant figures have been inserted as rustic clowns in every *nang sbek* play. Performance has religious significance, the gods being invoked and honoured, and a performance may be arranged to assure rain or to halt an epidemic. It is not certain when and how *nang sbek* originated, but it seems probable that it was taken to Thailand in the 15th century and then brought back. This would explain the details of costume and headdress of today's puppets that are in Thai style.

The lithe *apsaras* carved in Angkor's stone show details of the *lakon* style of female dance, but neither these nor other records are evidence that their lively dance was used in relating the epic stories. The 19th-century Thai rulers of western Cambodia reintroduced *lakon* dance and dance-drama, which was indigenous to Thailand as well. At the same time, Thailand's male masked pantomime was brought to Cambodia, as far as is known for the first time, and it became known as *lakon kawl.* Both male and female dance-plays were translated into Cambodian. Recently, costumes and headdresses have been redesigned in the style of the Angkor carvings. The stories, music, dance, and dramatic styles of *lakon* and *lakon kawl* are much like their Thai counterparts.

Popular forms. *Lakon bassac,* performed by some 20 professional troupes in Cambodia, is a highly eclectic form. Musical selections, dances for female characters, and costuming are borrowed from court *lakon.* The form was created by Khmers living in the Bassac River region of Vietnam. Villains wear Vietnamese costumes and move with Vietnamese opera movements, an evidence of the historical conflicts of the two peoples. Chinese, *Jātaka,* or Khmer stories may be performed. *Pi phat* music alternates with Chinese–Vietnamese instruments and with the Western saxophone and piano. Prince Sihanouk, chief of state between 1941 and 1970, encouraged a few French dramatic productions, but such drama is scarcely known outside the Western-educated elite.

Thailand. Folk *lakon jatri, lakon nai* female dance and dance-drama, *khon* masked pantomime, and *likay* popular theatre are Thailand's chief performing arts.

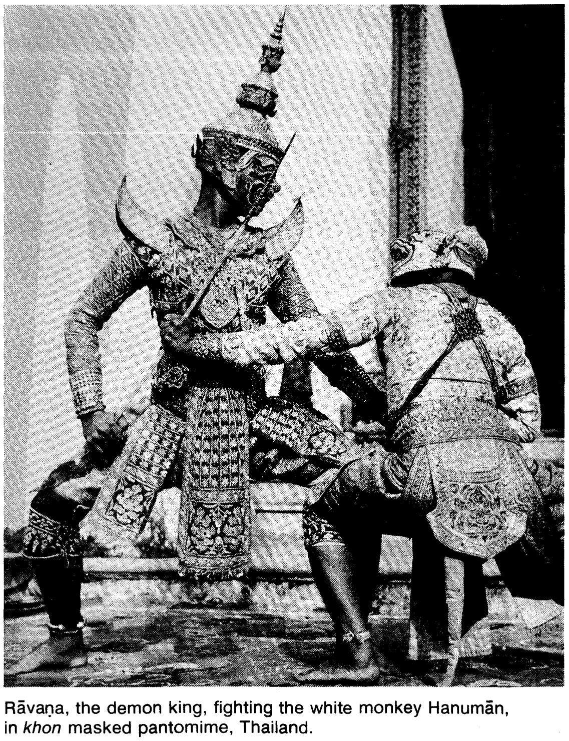
Folk performance. *Lakon jatri* began in the south, when male dancer-sorcerers performed, in simple folk style, the *Manora* Buddhist birth story as a dance-play. A troupe of three players was usual. One played the beautiful half-bird, half-human princess, Manora; a second played the hero, Prince Suton; and the third, often masked, played clown, ogre, or animal as needed. Flute, bell cymbal, and drums provided the music. The full *Manora* cycle of plays, staged in a village in the open, could last for two weeks. Probably after the 14th century, some *jatri* troupes moved to the Thai capital, where they established commercial theatres and staged a new all-male drama, *lakon nok* (*nok,* "outside" [the palace]), that emphasized plot and an often obscene humour. Advances in dramatic form were accomplished by court writers of *lakon nok* between 1800 and 1909. *Likay* troupes succeeded and completely supplanted *lakon nok* troupes in the early decades of the 20th century, but such popular *lakon nok* plays as *Sang Thong* ("The Prince of the Golden Conch") are presented today in modified form by the Thai National Theatre.

Bawdy humour

Female court dance-dramas. The *lakon nai* (*nai,* "inside" [the palace]) female dance-drama of the court was created in the mid-18th century from a confluence of three previously separate elements: female court dance, the *lakon nok* drama, and the Javanese *Pandji* stories as subject matter. Romantic episodes from the long *Pandji* tale were ideal for staging in the elegant and delicate style of female court dance, accompanied by songs and the music of a large *pi phat* ensemble. In the unhurried court atmosphere, dance scenes lasted an hour or more, and dance figures might be repeated many times. In time, other stories came to be staged in *lakon nai* and were given other names, but the *Pandji* plays composed by the daughters of King Boromokot (1733–58), by Rama I (1782–1809), and by Rama II (1809–24) remain favourites. In this form, *lakon nai* was introduced into Cambodia within the past two centuries.

Masked mime. Until recent years, a Thai version of the Khmer *nang sbek* shadow play, *nang yai,* occupied an important place in court as a Brahmanic-related ritual performance of the *Rāmāyaṇa.* Thai scholars describe it as the source of *khon* masked pantomime, citing celebrations for King Ramathibodi II in 1515 that included a *nang yai* performance without puppets. Wearing heavy makeup, the puppeteers themselves danced the usual *Rāmāyaṇa* episode as narrators told the story and spoke dialogue. Later, masks took the place of makeup, the screen was eliminated, and *khon* was born. In present-day Cambodia (Kampuchea), one troupe can perform both forms. A number of *lakon nai* elements entered *khon* in later years, so that today a *khon* performance mixes the vigorous, masculine *khon* with gentle *lakon nai* singing style and female dance. All of the Thai dance-drama traditions (*lakon jatri, lakon nok, lakon nai,* and *khon*) are taught at the Department of Fine Arts in Bangkok, and representative plays from them are staged, often mixing traditions, at the Thai National Theatre.

Marie Mattson—Black Star



Rāvaṇa, the demon king, fighting the white monkey Hanumān, in *khon* masked pantomime, Thailand.

Development of the *khon*

Popular plays and puppets. The major popular theatre form is *likay,* which evolved in part out of *lakon nok.* It is now performed by more than 100 troupes in most parts of Thailand. Actors are skilled in improvising not only the dialogue and lyrics but also the plot of a play as well, weaving romantic scenes and fragments of *lakon nai* dance, set to *pi phat* music, into a story from a well-known *Jātaka,* history, or court play. *Likay* plays are set to music of the Lao *khen,* a reed organ, in northeast Thailand. A type of shadow play called *nang talung,* in which a single, seated puppeteer moves small puppets of individual figures with movable arms, is very popular in southern Thailand. The performance technique undoubtedly came from Malaysia, while the plays and the identifying features of the puppet figures, mostly from the *Rāmāyaṇa,* are from Thai *khon* and *lakon nai.* A similar shadow play exists in Cambodia, suggesting that the form travelled from southern Thailand to Cambodia, perhaps in the 19th century.

Laos. From the time Laos became a kingdom in 1353, the performing arts at the relatively small Lao court at Luang Prabang followed those of the more illustrious courts to the south, Angkor in Cambodia and then Ayutthaya and Bangkok in Thailand. Today, Lao dancers study in Bangkok, and the style of dance, music, and drama of the Royal Lao Ballet, the only remaining court troupe in Southeast Asia, is almost identical with that of *lakon nai* in Thailand. It is usual to perform excerpts from the very long dance-plays, the staging of a full-length spectacle being beyond the means of the court at present. Male *khon* dance is known but seldom performed. A number of Lao folk dances are studied and performed by the royal ballet troupe.

Balladeers of Laos

Scores of popular troupes perform plays derived from Thai *likay* and set to the lively and melodic Lao folk song style known as *mohlam. Mohlam* balladeers, accompanied by the *khen* (a complex reed organ), have for centuries travelled the Lao-speaking countryside, which includes Laos and northeast Thailand, singing bawdy songs of physical love and weaving into their performance local gossip and bits from the epics and court plays. When *likay* troupes from Bangkok played in northeast Thailand, the *pi phat* music and court dancing were not popular, although the plays themselves were. Enterprising *mohlam* performers then set the *likay* plays to the familiar *mohlam* song style, thereby creating a new popular theatre form, *mohlam luong,* or "story *mohlam.*" Of the *mohlam* troupes, a few large ones are located in major cities in the two countries, but most are small and travel from village to village, performing for a few days or weeks in each.

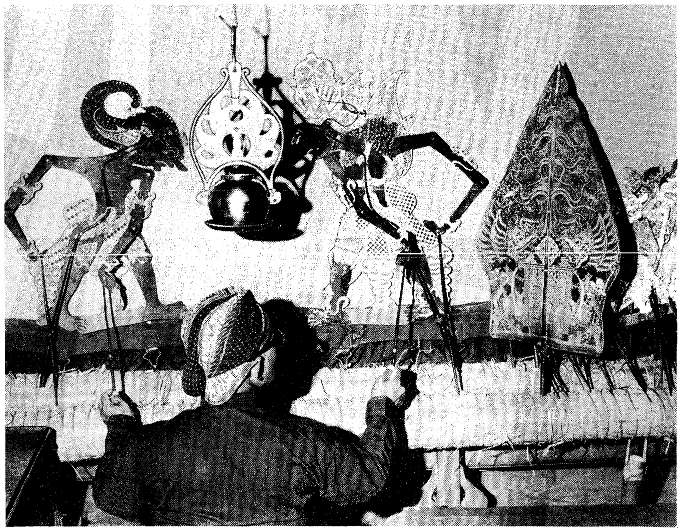
Burma. In spite of an old Burmese tradition of spirit dances stemming from animism and early contact with Indian culture, formal theatre did not begin until 1767, with the introduction of Thai *khon* and *lakon nai* to Burma following the capture and sack of Ayutthaya. Burmese courtiers and dancing girls immediately learned the two forms, and the plays were translated into Burmese. Because Rāma was viewed as a previous incarnation of Buddha, pious Burmese were reluctant to alter *khon* scripts. For a time *Jātaka* plays, including *Rāmāyaṇa* episodes, were forbidden to live actors. Instead, marionette troupes doing plays based on *khon* brought the Rāma stories to the Burmese countryside. But the *Pandji* plays were not considered *Jātakas*, and even the first Burmese version, by U Sa under the title *Inao,* departed from its Thai model, thus setting the stage for the creation of court drama, or *zat pwe,* based on myth and legend but capable of being independently developed. The three *zat* written by U Kyin

U portray the futility of political strife and urge a life of Buddhist renunciation. U Pon Nya created a freer form of dramatic verse, and his *Water Seller* is noted for its comparatively realistic treatment of court life.

Court drama ceased after 1866, when the British conquered Burma. Thereafter, drama was staged by professionals in public theatres, primarily in Rangoon. U Pok Ni in *Konmara* (*c.* 1875), U Ku in *The Orangoutan Brother and Sister* (1875), and others created a new type of drama, *pya zat,* that mixed royalty and commoners, emphasized humour, and added songs to appeal to a popular city audience. Hundreds of these works were published. Popular troupes in Burma today perform a long bill of attractions that lasts most of the night. It comprises songs and dances, a new contemporary play, and, as a final number, a classic *zat* in which remnants of old court music and dance are preserved. British touring companies in the late 19th and early 20th centuries brought examples of contemporary European melodrama and some classics to Burma. Subsequently a number of plays were written in Burmese and in English, following Western conventions and without songs or dance. Of these, *The People Win Through* (1950), by former prime minister U Nu, is among the most interesting examples.

Indonesia. The sober, majestic, and profound court arts of eastern and central Java, where Javanese is spoken, include *wayang kulit* shadow theatre, *wayang orang* unmasked dance, and *wayang topeng* masked dance.

By courtesy of the Indonesian Tourist Board



Wayang kulit (shadow-puppet theatre), Java.

Shadow-puppet theatre. It is uncertain whether the shadow theatre is indigenous to Java or was brought from India, but the *wayang kulit* technique of having a single seated puppeteer who manipulates puppets, sings, chants narration, and speaks dialogue seems to be an Indonesian invention. Unlike most court arts, *wayang kulit* has had centuries of performance in the folk tradition as well, so that today, with several thousand puppeteers active, it is the strongest traditional theatre form in Southeast Asia.

Plays are set in mythological times, some relating to indigenous animistic festivals and worship of local spirits, some directly dramatizing episodes from the *Rāmāyaṇa* and *Mahābhārata* epics, while the majority—the Pandawa (Pāṇḍav in Sanskrit) cycle of about 100 plays—are essentially Javanese creations in which the five heroic Pandawa brothers are placed in different situations. Three and sometimes four god-clown-servants and a set of ogre-antagonists who are not in the epics at all suggest how far removed the shadow plays are from the epics.

Conventions in puppet performances

The *wayang* puppeteer works within one of the world's most carefully organized performing arts, making possible a virtually solo performance without intermission, from around nine at night until the gray before dawn. Each play is in three parts, coordinated with three keys of music played by the gamelan ensemble. Certain standard scenes appear in a standard order, though some may be dropped. "Opening Audience" introduces the play's conflict, "Inner Palace" shows the king meeting his queen(s), and in "Outer Audience" the army is dispatched. In "Forest Clearing" the first battle scene occurs, and in "Foreign Audience" the antagonist kingdom, usually one of overseas ogres, is introduced. Concluding part one are "Foreign Outer Audience," in which the second army marches forth, and "Opening Skirmish," a battle scene between the two armies. The puppeteer chooses from among 150 musical selections, matched to scene type, character, mood, or action. The puppet figures are carved to indicate character type and status according to fixed patterns for nose, eyes, gaze, stance, body build, and costume. The puppeteer can choose one or another puppet of the same character, coloured gold or black or with a stern or relaxed countenance, to indicate the mood of the figure in a particular scene. In battle scenes, he develops individual encounters between opponents, drawing upon a repertory of 119 movements that are classified for use by god, female, refined hero, muscular hero, ogre, or monkey. Formula narrative phrases describe famous kingdoms and characters, and battles are preceded by challenges couched in standard phrases. Although the puppeteer works only from a brief scenario, he is able to extemporize each performance, adding contemporary jokes for the clowns and molding the performance to suit the occasion and the audience. He and his supporting musicians and female singers are improvising within completely known, although exceptionally complex and subtle, artistic conventions.

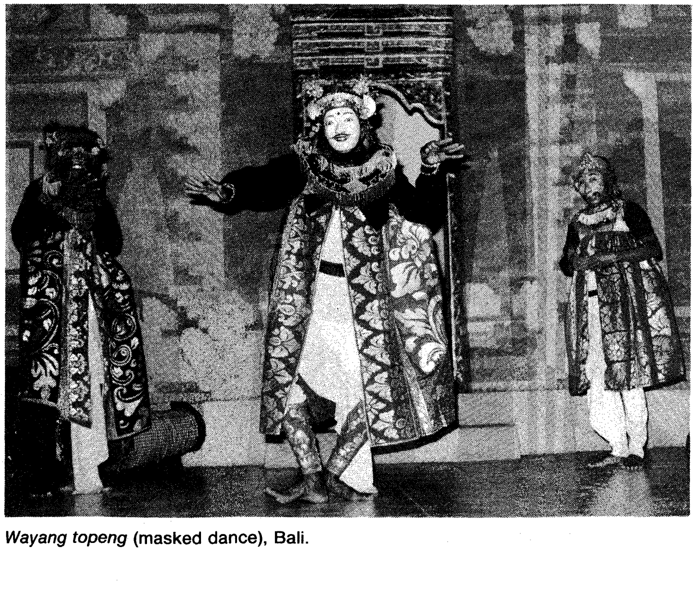
This artistic system, developed within the shadow theatre for performance of Pandawa plays, has proven to work so well that it has been widely imitated. The entire body of *wayang kulit* drama was adopted in Bali and in Malaysia. At least 25 other play cycles have been performed in Indonesia as shadow drama within this system, including the *Pandji* cycle (*wayang gedog*), Islāmic Amīr Ḥamzah plays (*wayang menak*), and plays dramatizing the revolutionary struggle against the Dutch (*wayang suluh*). The Pandawa *wayang kulit* repertory was transposed to the doll-puppet theatre (*wayang golek*) in Sunda, the western part of Java, and to dance-drama in eastern and central Java (*wayang orang*) and in Bali (*wayang wong*).

Performances are commissioned for special occasions and usually can be interpreted in religious or mystical fashion. There may be offertory plays at harvest time or animistic, ritualistic exorcisms protecting children from being devoured by the voracious god Kala. In *The Reincarnation of Rāma* the divine attributes of the god Wisnu (Vishnu in Sanskrit) reincarnate in Ardjuna (Arjuna), hero of the Pandawa cycle and ancestor of the Javanese race. The translucent screen can be interpreted as heaven, the banana-log stage as earth, the puppets as man, and the puppeteer as god, and the Pandawas can symbolize the manifold attributes of righteous behaviour.

Masked dance

Wayang topeng. Masked dance was also popular at the eastern Javanese courts (*c.* 1000–1400) and may be re-

© Pascal Grellety—Bosviel, Paris



Wayang topeng (masked dance), Bali.

lated to ancient animistic masked dance seen throughout the Pacific islands. Later, Indian dance style was assimilated, and sometime after the 15th century at the earliest, the *Pandji* story was dramatized. This is *wayang topeng,* widely performed as both a sophisticated and a folk art throughout Indonesia. Unlike the large-scale unmasked dance-drama, *topeng* dance focusses on interpreting character through solo dance.

Wayang orang. Java's spectacular dance-drama, *wayang orang,* grew out of the strong unmasked dance tradition that is illustrated in reliefs of female dancers carved on the 9th-century Borobuḍur and Prambanan temples in central Java and that produced the carefully cultivated female group dances of the Surakarta and Jogjakarta courts after their establishment in the 16th century. Of the latter dances, two stand out, the almost sacred *bedaja,* which even today is danced only in court surroundings, and the *srimpi,* in which two pairs of girls execute a delicate slow-motion duel with daggers and bows. In the middle of the 18th century, *wayang kulit's* Rāma and Pandawa plays were set to court dance to form *wayang orang,* or "human" *wayang.* The music, narrative, and dramatic organization of the shadow play was kept largely intact, many of the actors' movements mimicking the stiff actions of the puppets, while new dance sections were added. Court performances stopped with World War II, but *wayang orang* continues to be performed by some 20 to 30 professional troupes in major cities. In popular performances, attractive actresses play the roles of such refined heroes as Ardjuna and humour and spectacle take precedence over dance.

Actresses as refined heroes

Ketoprak and ludruk. Two other types of popular theatre, *ketoprak* and *ludruk,* were performed in Java by 150 to 200 professional troupes. *Ketoprak,* created by a Surakarta court official in 1914, evolved into a spoken drama of Javanese and Islāmic history in which the clown figure is a spokesman of the common man. Whereas *ketoprak* is performed primarily in central Java, *ludruk,* a spoken drama that handles mainly contemporary subject matter, is performed in eastern Java by both amateur and professional troupes. Though *ludruk* is relatively realistic, male actors play all roles. Songs and dances, accompanied by gamelan music, are performed between acts in both forms.

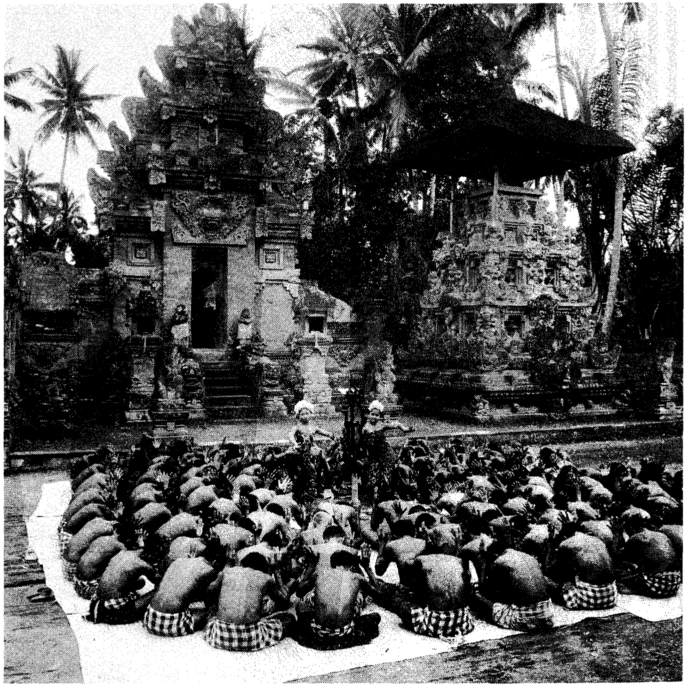
Sundanese performing arts. There are three main performing arts in the Sundanese area of western Java. *Reog,* a kind of urban folk performance, can be seen especially in the streets of Jakarta: two or three men improvise popular songs, dances, and dramatic sketches for a neighbourhood audience in this type of entertainment. *Wayang golek* is a performance based on *wayang kulit* but using doll puppets without a screen. Approximately 500 Sundanese puppeteers perform *wayang golek.* Female singers, who are almost as important as the puppeteer, respond to requests and gifts of money by singing song after song and virtually stopping the play. *Sandiwara* troupes in Jakarta, Bandung, and a score of other cities perform both *wayang* stories in the form of Sundanese dance-drama and spoken historical and contemporary dramas for popular audiences. Sundanese-style court dances and *topeng* masked dances are often performed solo at festivals and for circumcision or wedding celebrations in private homes. Sundanese dance is more sensuous than Javanese and broader in style.

Balinese dance-drama. Of the many factors that have contributed to the remarkable flourishing of dance and drama on the island of Bali for more than a millennium, three are of particular note. First, Bali remained isolated from both Islām and the West. Second, there was a merging of folk and court performance styles into a single communal tradition appreciated by all. Third, dances and plays are indissolubly linked to the recurring cycles of local festivals and rituals whereby the well-being of the community is maintained against constantly threatening malicious forces in the spirit world. From the verve and brilliance of Balinese performances it is clear not only that the people like to perform but also that there exists some culturally determined compulsion to do so.

Balinese dance and dramatic forms are so numerous that only a few can be noted. Balinese villagers playing in the *barong* exorcism dance-drama are not merely actors exercising theatrical skills. The actors' bodies, going into a trance, are believed to receive the spirits of Rangda and the Barong, and it is the spirits themselves that do battle. Thus the performance is actually more a ritual than a piece of theatre. The *sanghyang* dance is usually performed by two young girls who gradually go into a state of trance as women sing in chorus and incense is wafted about them. Supposedly entered by the spirit of the nymph Supraba, the girls rise and dance, often acrobatically, though they have been chosen from among girls untrained in dance. The dance's purpose is to entice Supraba to the village to gain her blessing when evil forces threaten. In the *ketjak,* or monkey dance, as many as 150 village men, sitting in concentric circles around a flaming lamp, chant and gesticulate in unison until, in trance, they appear to have become ecstatically possessed by the spirits of monkeys. This performance, however, has no ritual function of altering an earthly condition.

Hypnotic possessio of Balines dancers

Tor Eigeland—Black Star



Ketjak, or monkey dance, Bali.

That the Balinese *wayang kulit* may represent the older style of *wayang,* known on Java before the coming of Islām, is suggested by the less stylized shape of the puppets, by the shorter performing time of four to five hours, and by the simple music of only four *gender,* a bronze instrument similar to a xylophone with resonance chambers underneath, from the gamelan ensemble. In one type of shadow play having a special religious significance, the puppets perform before a screen during the daytime, and the puppeteer is seen in his role as a Brahman priest, bare to the waist. In the *redjang* processional dance, village women symbolically offer their bodies to their temple gods.

Because Balinese performing arts are vitally alive, they change from decade to decade, even from year to year. The *gambuh,* respected for its age, contains elements of dramatic dance, song, narrative, and characterization found in later forms. It is thought dull, however, and is seldom performed, though it is believed to have provided the model for the singing style of popular *ardja* opera troupes and the dance style of the lovely girls' *legong. Wayang wong* is analogous to the Javanese *wayang orang,* but masks are worn and the repertory is limited to Rāma plays. Pandawa plays are staged in identical style but are called *parwa.* It has been suggested that these forms also stem, at least in part, from *gambuh. Wayang topeng* masked-dance plays are ancient, being mentioned in a palm-leaf document of 1058. The Javanese chronicle of the Majapahit period (*c.* 1293–1520), the *Pararaton,* in which Ken Angrok is the hero, is a favourite *tapeng* story. This points to the strong influence exerted by Javanese on Balinese arts after the Majapahit court was transferred to Bali in the 16th century to escape Islāmic domination.

Malaysia. The Malay peninsula, in the geographical centre of Southeast Asia, has assimilated repeated intrusions of neighbouring cultures. The dances of the former princely states on the east coast show the influence of Indian nondramatic dance.

The multiform wayang. Rulers from Java in the 13th and 14th centuries and later large colonies of Javanese introduced their *wayang kulit* shadow theatre. The puppets of *wayang Djawa,* or "Javanese" *wayang,* are identical with the two-armed, long-nosed, highly stylized puppets of today's Javanese *wayang kulit.* Those of *wayang Melayu,* or "Malayan" *wayang,* have only a single movable arm and are less sophisticated in conception, which suggests that they are either descended from old Javanese puppets, before both arms were made movable, or are a degeneration of the more complex form. Rāma, Pandawa, and Pandji plays are staged. The puppets of *wayang Siam,* or "Siamese" *wayang,* though manipulated by a single seated puppeteer, represent a Thai conception of the figures from the *Rāmāyaṇa;* and costumes, headdresses, ornamentation, and facial features follow those of *khon.* The plays include Islāmic elements as well, while the chief clown figure, Pak Dogol, is thought to be a recent Malay creation that has supplanted Semar, the Javanese clown of *wayang kulit.*

ulti-
ıgual
rfor-
ances

In a performance, puppets of all types may appear together. Either such Thai instruments as the *lakon jatri* drum and small bell cymbals or gamelan instruments play the accompanying music. Song lyrics can be in ancient Javanese; animistic, Islāmic, and Hindu-derived invocations to the gods are offered in the Thai and Malay languages; and the play proper is in colloquial Malay. Puppeteers once performed throughout the peninsula, including the five Malay-speaking provinces of southern Thailand, but today puppeteers are found primarily in northeast Malaysia.

Chinese and popular entertainments. Chinese immigrants introduced various forms of opera during the 19th century. Troupes perform for Chinese Buddhist temple festivals, for local fairs, or on national holidays. In Singapore troupes occasionally perform in public theatres as well. Young people of Chinese descent in both Malaysia and Singapore have little interest in the opera, however, because their Chinese is limited. Occasionally troupes import star performers from Hong Kong or tour Chinese communities in Thailand.

Bangsawan was created by professional Malay-speaking actors in the 1920s as light, popular entertainment. Songs and contemporary dances were added to a repertory of dramatic pieces drawn from Islāmic romances and adventure stories. Troupes travelled to Sumatra, Kalimantan, Sunda, and Java, where their melodramatic plays found large audiences and influenced local performers of *sandiwara, ketoprak,* and *ludruk.* The cinema and television, however, have captured much of this audience.

Vietnam. An indication of the antiquity of the performing arts in Vietnam is a large bronze drum of the 3rd century BC found near Haiphong, in northern Vietnam, which is ornamented with instruments and musicians playing for dancers. Chinese performing arts presumably were a part of court life in northern Vietnam during the period of Chinese rule (111 BC–AD 939), and between the 10th and 13th centuries the dances and music of the Hinduized Cham peoples, living in what is now central Vietnam, were welcomed there. The melancholy Cham songs were particularly popular, and most authorities believe that the sad southern style of Vietnamese singing is derived from them.

Satirical drama. Hat cheo is a popular, satirical folk play of northern Vietnam that combines folk songs and dances with humorous sketches criticizing the people's rulers. Some scholars theorize that it is an indigenous folk art, whereas others, to show that it reached the people from the court, cite the legend of a Chinese actor who in 1005 was hired by the Vietnamese king to teach "Chinese satirical theatre" to his courtiers. *Hat cheo* is widely encouraged by the government.

The opera. The classic opera, known as *hat boi, hat bo,* or *hat tuong,* is a Vietnamese adaptation of the Chinese opera long supported by kings and provincial mandarins as a court art and performed for popular audiences as well, especially in central Vietnam. The introduction of Chinese opera is attributed to the capture of a troupe of performers attached to the Mongol army that invaded northern Vietnam in 1285. The actors' lives were spared in return for teaching their art to the Vietnamese. In 1350 another Chinese performer was engaged by the northern court as an instructor. Almost exclusively a court art in the north, *hat boi* was made a form of popular entertainment in central Vietnam by the playwright Dao Duy Tu in the 16th century. It was introduced to southern Vietnam under the Nguyen dynasty in the 18th and 19th centuries, but its future was jeopardized by the decades of war in the mid-20th century. The last large troupe of court musicians, dancers, and actors at Hue in southern Vietnam disbanded in 1945. The postwar government of the late 20th century did not provide *hat boi* with strong support, and the popular troupes lacked audiences.

Blighting
effects
of war

In form and content, *hat boi* is a blend of China and Vietnam. Direct imitation of Chinese costume and acting techniques was encouraged under the reign (1847–83) of Emperor Tu Duc and it is probable that the present form of *hat boi* dates from this period. At Tu Duc's court in Hue, the playwright and scholar Dao Tan gathered 300 actors and with them wrote out texts of the standard repertory that previously had been preserved orally. He then had the texts published and distributed them to actors and troupe managers. In the 20th century there has been a movement to loosen the rigid structure of *hat boi* and to reduce the high proportion of Chinese loan words that makes the operas difficult for the ordinary Vietnamese to appreciate.

Following Chinese practice, the operas are classified as military or domestic. The former, which may be derived from Chinese and Vietnamese legend or history or may be purely fictional, concern struggles for power between kings. The Chinese novel *The Romance of the Three Kingdoms* furnishes material for many military plays. The latter, dealing with the lives of commoners, contain humorous scenes alternating with scenes of suffering that are played to the accompaniment of sad southern-style songs. The Confucian ethic of obligation to one's superior—of wife to husband, of son to father, or of subject to king—underlies plays of both types.

Hat boi staging is modelled on conventions of Chinese opera. Actors perform on a stage that is bare except for a table and two chairs. These can serve as a castle, a cave, or a bed as well as for sitting and eating. A single embroidered drop at the rear has an entrance right and an exit left. Costume and makeup indicate character type: black for boldness, red for anger or rashness, white for treachery, and gold as the colour of the gods. Conventionalized mime may be used alone or in conjunction with symbolic properties. The actor mimes stepping over an imaginary threshold or sewing without needle and thread, but he indicates riding a horse by gestures with a riding crop and travels in a carriage when a stage assistant holds flags with wheels painted on them at each side of his body. Percussion instruments accompany stage action, and songs—which may be in falsetto Chinese style, in soft southern Vietnamese style, or in a form of prose recitative—are accompanied by stringed instruments.

The love
lament

The popular stage. Southern-style singing is the basis of another type of theatre, *cai luong,* begun in the 1920s by popular singers who performed plays in which they sang the love lament *Vong Co.* Today, regardless of whether a historical or contemporary play is being performed as *cai luong* or which of many troupes is staging it, this melody will be heard throughout the play many times, underlying different lyrics. *Cai luong* stars are lionized, and the best troupes maintain high artistic standards. Among popular theatre forms in Southeast Asia, only *cai luong* plays are fully scripted and directed as they would be in the Western theatre. In contrast to the operetta form of *cai luong,* modern spoken drama is known as *kich.* It is a young dramatic form performed mostly by amateurs who are trying to put Western dramatic conventions into practice.

The Philippines. Whatever indigenous theatrical forms may have existed in the Philippines, other than tribal epic recitations, were obliterated by the Spanish to facilitate the spread of Christianity.

The comedia. The earliest known form of organized theatre is the *comedia,* or *moro-moro,* created by Spanish priests. In 1637 a play was written to dramatize the recent capture by a Christian Filipino army of an Islāmic stronghold. It was so popular that other plays were written and staged as folk dramas in Christianized villages throughout the Philippines. All told similar stories of Christian armies defeating the hated Moors. With the decline of Spanish influence, the *comedia,* too, declined in popularity. Some professional troupes performed *comedia* in Manila and provincial capitals prior to World War II. Today it can still be seen at a number of church festivals in villages, where it remains a major social and religious event of the year. Much in the manner of the medieval European mystery-play performances, hundreds of local people donate time and money over several months to mount an impressive performance.

By courtesy of Philippine Embassy

Comedia, or *moro-moro,* folk drama based on the battles between the Christians and the Moros, the Philippines.

Styles from Europe. Dances and dramas from Spain were brought in, some of which took root. The "María Clara," a stately minuet, and the "Rigodón de Honor," a quadrille, were adopted by local European society for its formal balls. Spain's sprightly operetta, the *zarzuela,* became the favourite light entertainment in Manila and other cities. Professional *zarzuela* troupes continued to flourish in the early decades of the 20th century but had disappeared by World War II. New plays with original music were produced in profusion. A number of them based on topical themes and criticizing American colonial policies were banned.

Western drama is studied and widely performed in both English and Tagalog. There are no professional companies, but amateur university and community groups abound. Western classics and recent popular successes are staged, and in recent years many original plays have been written to celebrate the Filipino heritage.

(J.R.B.)

Visual arts

GENERAL CONSIDERATIONS

Religious–aesthetic traditions. The visual arts in Southeast Asia have followed two major traditions.

Indigenous magical and animist tradition. The first is a complex inheritance of magical and animist art shared by the different tribal peoples of the mainland, where it evolved from Paleolithic origins, and of the islands. Such art gave the peoples who made it a sense of their identity in relation to the forces of their natural environment, to the structure of their society, and to time. It consists of types of potent emblem, mask, and ancestral figures broadly similar to those that hunters and early farmers the world over have used in connection with seasonal ceremonies, life and death rituals, and ecstatic shamanism (belief in an unseen world of gods, demons, and ancestral spirits responsive only to the shamans, or priests). The spiritual powers that the arts name and invoke are local and vary from group to group of the population. The rich formal artistic languages have been subject to successive episodes of influence from inland Asia, but each of the tribal groupings has developed its own artistic language on the basis of a common fund of Southeast Asian thought forms.

Indian tradition. The second major tradition was received from India during the early centuries of the Christian Era, when seagoing merchants from that subcontinent so fertile in ideas were expanding their trading activity. Into many parts of Southeast Asia—especially Burma, Thailand, and the coasts of Cambodia and Indonesia, where Indian traders settled and married into the families of local chieftains—they brought with them a script and literature in the sophisticated Sanskrit language. They also brought a highly developed conceptual system dealing with kingship, statecraft, and hydraulic engineering, integrated and authenticated by profound metaphysical ideologies of Indian pattern, both Hindu and Buddhist. These ideologies claimed to be universal, embracing all human diversity within a cosmic frame of reference. And this explains why the culture was adopted. For there was no Indian conquest of terrain; instead, the Indian conceptions, along with the art that expressed them, were used by dynasties in the colonial kingdoms as a method of overcoming divisions in their population, of centralizing effort, and of uniting their religions into viable states based upon cities. Although the new religious conceptions must have offered deep personal satisfaction to the general population, the architecture and sculpture in stone and bronze in which they were artistically expressed were expensive in materials, labour, and skill and were thus available primarily to patrons who were claiming for themselves a royal (*i.e.,* divine) status and using the resources of art to demonstrate that status.

Why the Indian culture wa adopted ir Southeast Asia

The Indianizing traditions were continually refreshed by direct influences from India and Ceylon. There can be very little doubt that, during the early centuries after the birth of Christ, Indian artists and craftsmen travelled to work in the distant trading colonies of Southeast Asia, for they would have been needed to set up local traditions with proper formulae and methods. And there can be no doubt either that works of art made in India were continuously exported to the colonial kingdoms, thus keeping the local art styles in touch with developments "at home." It is also clear, however, that within a very short space of time the Southeast Asian kingdoms produced their own distinctive local versions of Indian styles; and some of their work shows skill, finesse, and invention on a colossal scale unrivalled even in India.

Although the art styles were to some extent sectarian, and sectarian partisanship played a part in political events, it was by no means unusual to find Hinduism and different forms of Buddhism flourishing side by side. In both Burma and Thailand, however, dynastic options were early exercised in favour of that particular form of Buddhism known as Theravāda (Hīnayāna), which adheres to the nontheistic ideal of purification of the self to Nirvāṇa. These countries follow the same form to the present day. It was also adopted in Cambodia and southern Vietnam after prolonged and successful periods of Hindu and Mahāyāna (a theistic branch teaching compassion and universal salvation) Buddhist dominance. The strongly Sinicized population of the region around the Gulf of Tonkin, which pushed gradually down the coast of Vietnam to become the modern plains Vietnamese, began to adopt Theravāda Buddhism with its artistic types by about the 13th century AD, partly because this form could be best adapted to its self-contained and antidynastic cellular social structure.

Theravādε and Mahāyānε Buddhism

Relations between the two traditions. Even in those regions where Indian influence became strongly entrenched, the older layers of more primitive religion and artistic consciousness remained very much alive. Indian deities were readily identified with local spirits. The tribal populations retained, as many still do, their old animist customs, especially those connected with fertility and practical magic, often with an art (in perishable materials) in which to express them. These arts were influenced by and exercised a reciprocal influence upon the styles of officially imposed Indianized arts. In many parts of Southeast Asia, where the official Indian styles were not completely established (most of Borneo) or where they died out (colonies in Celebes), in inaccessible areas beyond the reach of dynastic influence, or on isolated islands, primitive styles have survived unmodified. Even in Indianized regions where a strict formula, say, for a necessary building type, had not been imported, a native pattern was adopted into the official canon (*e.g.,* Laos). In the Indonesian island of Bali, which has remained nominally Hindu, the Indian and the folk elements were thoroughly assimilated to each other, producing a quite individual style of both religion and art. In Sumatra and Java, whose populations were gradually converted to Islām from India during the 13th–16th centuries, the primitive cult of the ancestors was revived and encouraged by Muslim rulers, with folk versions of denatured Hindu art adapted to it. Decorative styles based on this art have flourished there and have been officially revivified during recent years. In the Philippines, notably in and around Manila, Spanish Roman Catholic art flourished after the Spanish colonization of 1571.

he royal
mple

Artistic styles. The royal temple is the basis for the classic Indianizing styles of Southeast Asia. Each Hindu temple is centred on a shrine, symbolizing heaven upon earth, which is crowned by a roof tower representing the cosmic Indian mountain, Meru, conceived as the hub of creation. Since all the peoples of Southeast Asia already believed the natural habitat of spirits and gods to be a mountaintop, the Indian pattern was readily accepted. The temple usually stands upon a lofty terraced plinth (a block serving as a base), which itself also symbolized a mountain. Towered shrines could be multiplied on the terraces, though one of them remained the principal focus. Within the cell of this main shrine was a sacred image carved in stone or cast in bronze. The local Hindu ruler identified the subject of this image as his transcendent patron, or celestial alter ego. This was normally one of the Indian high gods, Śiva (represented perhaps by a phallic emblem, the *liṅga*) or Vishnu. In Mahāyāna Buddhist kingdoms a royal *bodhisattva* (a being that refrains from entering Nirvāṇa in order to save others) was sometimes adopted to fulfill the same role, a favourite form being known as Lokanātha, or Lokeśvara, Lord of the World. Subsidiary shrines, niches, or terraces sometimes contained subsidiary images, including goddesses representing at the same time wives of the god and queens of the king. These images were worked in smooth, deeply rounded, and sensuously emphatic styles derived from Indian art but with varying inflections characteristic of each region and time. The whole exterior of the shrine was usually adorned with rhythmic moldings, foliage, and scrollwork, with figures representing the inhabitants of the heavens. Ideally, the building was constructed and carved in stone; but, particularly where good stone was not readily available (for example, in Burmese Pagan), it could also be brick, coated and sculptured with stucco after northeast Indian patterns. Temple complexes tended to grow as successive kings strove to outdo their predecessors with the magnificence of their buildings. Hindu rulers, influenced perhaps by vestiges of tribal custom, would sometimes retain their own family's temples and images while destroying those of earlier dynasties.

Buddhism, however, is a religion based on a doctrine of transcendent merit and sustained by an order of monks who have, ultimately, no vested interest in kings and gods. They may, however, take a great interest in the world of spirits and the operations of astrology, just as the local population does, even though they regard such matters as subordinate to the ultimate Buddhist aim of universal Nirvāṇa. Buddhist monasteries, therefore, tended to expand around stupas (domed monuments emblematic of the Buddhist truth, also called pagodas or *dāgabas*) of ever-increasing size and number; the preaching halls, libraries, and living quarters for monks were continually enlarged and repeatedly rebuilt, often as a testimony to the piety of royal patrons. Although, strictly speaking, Theravāda Buddhism has no place for a "divine ruler" whose identity an actual king may adopt, provision was made in legend and in court and monastic ritual for the ruler of a Theravāda country to assume a magical role as the dominant sponsor and patron of the Buddhist truth. His legendary prototype was usually not identified, therefore, with an icon of the enlightened Buddha but with images such as the chief disciple at the knee of the enlightened Buddha, as Prince Siddhārtha (the Buddha-to-be), or figuring in scenes of the Buddha's life that lined the monastery halls and corridors.

Prescriptive
formulas
for art

Both Hindu and Buddhist art were produced according to prescriptive formulas. If the formulas were not followed, the art was believed not to fulfill its transcendent function. In practice, however, there has been room for styles and types of image to change and develop fairly quickly. Hindu and non-Theravāda art recognized what could be called aesthetic values as a component in religious expression. Theravāda Buddhism, however, which might be called fundamentalist, has always attempted to preserve the closest possible connections with the Buddha's recorded original deeds and sayings; its art, therefore, has concentrated on repeating in its main Buddha figures the most exact possible imitations of authentic ancient images. This had led to a relative monotony of style in Theravāda icons (see below *Burma; Thailand and Laos*). In the subsidiary sculptured and painted figures, however, which illustrate scenes from sacred history, Theravāda art has had greater freedom of invention. In the 20th century, Theravāda Buddhism is the only form of Indian religion to survive in Southeast Asia, save for the modified Hinduism of Bali. Its architecture in recent centuries has been decorated with a vigorous, sometimes coarse fantasy and made gaudy with gilt paint and coloured glass.

General development of Southeast Asian art. Most of the works made under the inspiration of the primitive, magical, and animist tradition are in perishable materials such as wood. Because the climate is so hostile, the works that survive are relatively recent; and any that is even 100 years old generally owes its preservation to Western interest. There are, however, a large number of Neolithic stone implements and prehistoric stone monuments (megaliths), as well as bronzes, which provide a solid archaeological basis for interpretation of Southeast Asian primitive art.

For the art of the classic Indianizing civilizations, French archaeology played the major role in clearing, excavating, and reconstructing major sites in Cambodia, Laos, and Vietnam; Dutch archaeology in Indonesia; British in Burma. Old bronzes have been found in fair quantities; apart from those of the early Dong Son culture (see below *Bronze Age: Dong Son culture*), all belong to one or other of the Indianizing traditions. Many old brick and stucco buildings survive, notably the medieval work at Pagan and in central Thailand, though an enormous number are known to have perished. Little very old painting is known, save a few Indianizing medieval rock and wall paintings on plaster. In spite of the fact that Buddhist monasteries are able to act as agents for preserving their own artworks, most of the surviving Buddhist pictorial art on wooden panels or other fragile material is less than 300 years old.

The stone of dynastic buildings, of course, has survived far the best. Scholars thus know much more about Indianizing stone architecture, with its sculpture, than about any other Southeast Asian visual art. But, where good relief sculpture flourished, one can legitimately assume that vanished pictorial arts also flourished; and from details carved in stone and incised on bronze, as well as from the scattered enthusiastic references in Chinese sources, one can be sure that throughout their history the Southeast Asian peoples have been intensely creative and have lived their lives surrounded by a wealth of imaginative art in many different mediums.

Unexcavated sites

There are many sites yet to be discovered and excavated. Twentieth-century knowledge of the history of art in many parts of Southeast Asia, especially of important episodes in Burma, Thailand, and Sumatra, is still scantily documented.

Neolithic period. The earliest works in Southeast Asia that can be called art are the rectangular polished axheads of a familiar late Neolithic type that have been found at many sites in Malaya, Indochina, and Indonesia. Some of the later Neolithic (*c.* 2000 BC to early centuries AD) implements are extremely beautiful and polished with the greatest care. They include practical adzes and axes; but some, made of semiprecious stone, were clearly intended for purely ritual purposes. Even in the 20th century a few such blades are preserved and revered as sacred objects in certain Indonesian farming communities and similar objects have continued to be made in some very remôte regions. These tools, with their fine edges, suggest that their owners were capable of very high quality woodworking and might well have decorated their wooden houses with designs of which we know nothing.

During the Neolithic Period, metal—both bronze and iron—came into use for implements but did not greatly alter the material culture. In many regions, notably Cambodia, Borneo, and Sumatra, numerous works of megalithic, or stone, art survive, including menhirs (single upright monoliths), dolmens (two or more upright monoliths supporting a horizontal slab), cist graves (Neolithic graves lined with stone slabs), and terraced burial mounds, all dating from the late Neolithic epoch. Some remarkable large stones are worked in relief with symbols and with images of animals and men, notably in the Pasemah region of Sumatra. Shaped stone sarcophagi and skull troughs (containers to hold the skulls of ancestors and of enemies at village shrines) are also known. These megalithic art objects suggest a highly developed cult of a spirit world connected with the remains of the dead (see below *Cambodia and Vietnam; Indonesia*).

Bronze age: Dong Son culture (c. 4th–1st centuries BC). By about 300 BC a civilization with elaborate arts based on bronzeworking existed, extending probably from the Tongking region into Laos, Vietnam, Cambodia, and Indonesia. This is called for convenience, after a major site, the Dong Son culture, though it may not have been a true cultural unity. A variety of bronze ritual works, many decorated with human and animal figures and with masks, were cast by the cire perdue method (metal casting using a wax model). The chief objects were ceremonial drums, large and small; the largest was found in Bali and is called "the Moon of Bali" (see below *Indonesia*). Extremely elaborate bronze ceremonial axes were made—probably as emblems of power. Certain relief patterns on the bronzes suggest that "ship of the dead" designs, like those still woven in textiles in both Borneo and Sumatra, may well have been woven even then. The spiral is a frequent Dong Son decorative motif; later Dong Son art was probably responsible for transmitting—especially into Vietnam, Cambodia, and Borneo—versions of the contemporary Chinese Chou dynasty's asymmetrical squared-hook patterns.

1st to 10th century. During the 1st century AD, Indian influence began to spread through Southeast Asia in the wake of trade, both overland, through Burma and Thailand, and by sea traders settled at especially favourable spots along the inland roads and along the sea routes around the coasts and into the islands. Buddhism, which was particularly popular among the Indian merchant classes, took root at a large number of trading cities, where monasteries were set up under the patronage of local kings. Many fragmentary Buddha images based upon Indian types of *c.* AD 300–400 have been found in Burma, Thailand, and Cambodia, produced in the kingdoms of the Mon people, the chief of which, in Thailand, was called Dvaravati. By the 5th century the first Hindu kingdoms had been established in western Java and Borneo. These kingdoms produced dynastic cult images, fragments of which have been found.

The first Hindu kingdoms

Perhaps the most splendid of the earlier Indianizing kingdoms, lasting till the 9th century AD, was that of the Pyu people in the upper Irrawaddy River Valley. The Pyu were the people most directly in touch with eastern India by land routes. Only one of their enormous cities has been explored archaeologically (see below *Burma*). The remains of Buddhist buildings, east Indian Buddhist images, and Hindu sculptures of Vishnu have been found there.

In the 1st century AD the predominantly Hindu kingdom known as Funan (the name given it by Chinese historians) was established in Cambodia. It seems to have controlled an empire that included kingdoms in Malaya and even parts of southern Burma. Its population was probably Mon and shared the culture of the Mon in the lower Irrawaddy Basin. (The Funan kingdom really represents the earliest phase of what became, in the 9th century, the great Cambodian Khmer Empire.) Between *c.* 550 and 680 the kingdom retreated from the coast up the Mekong River into Laos, where it was called by the Chinese Chenla. This joint Funan–Chenla tradition produced some of the world's most magnificent stone cult images. Though Buddhist icons are known, these images principally represent Hindu deities including Vishnu, his incarnation Krishna, Śiva, and a combined Śiva-Vishnu figure called Harihara. The images were housed in wooden or brick shrines, now vanished.

During the Chenla retreat the Theravāda Buddhist kingdom of Dvaravati flourished in southern Thailand, on the lower reaches of the Mae Nam Chao Phraya; the kingdom lasted until the 11th century, when it was captured by the Khmer. What little of its art is known is close to that of eastern India and provided the basis for later Buddhist art in the Khmer Empire, as well as for some of the later forms of Thai art.

The centra Javanese kingdom

Almost contemporary with Chenla was the rise of the central Javanese kingdom. Soon after AD 600 the earliest surviving Hindu temples were built. In *c.* 770 the Śailendra dynasty began its long series of superb stonecut monuments both Hindu and Buddhist, which culminate in two enormous symbolic architectural complexes: the Mahāyāna Buddhist Borobuḍur (*c.* 800) and the Hindu Lara Jonggrang, at Prambanam (*c.* 900–930). These monuments were decorated in an individual and exceptionally accomplished style of full-round and relief sculpture. Many small bronze religious images have survived. The art of the Śailendra dynasty testifies to the imperial and maritime power of the central Javanese kingdom, which seems to have influenced politics and art in Khmer Cambodia. It also took over the possessions of a major Theravāda Buddhist kingdom called Śrivijaya, which had flourished in Malaya and Sumatra and was centred at Palembang. The Javanese Śailendra ruled most of Malaya and Sumatra and installed themselves there in the mid-9th century, when their home terrain in Java was taken over by the Mataram dynasty, heralding the eastern Javanese period, which began in 927. Śrivijaya, under Śailendra rule, declined in the mid-11th century, and most of its remains still await discovery.

In Vietnam *c.* the 2nd century AD the predominantly Hindu kingdom of Champa was founded. Its capital was at My Son, where many temples have been found. This kingdom suffered much from attacks by the Chinese, and, after it began to lose the north to the Sinicized Vietnamese in 1069, the Cham capital moved in 1069 to Vijaya (Binh Dinh), in the south. There it was involved in continual warfare with the Khmer, who finally annexed southern Vietnam in 1203. The art of the northern Vietnamese as a whole has always been so strongly under the influence of China that it can best be characterized as a provincial Chinese style.

The Khmer Empire

10th century to the present. In Cambodia the Khmer Empire succeeded to the old territories of Funan–Chenla. About 790 the first major Khmer ruler, Jayavarman II, who was related to the old Funan royal family, came to Cambodia from the Śailendra court in Java. In 802 he set up a religious capital on a hill at Phnom Kulen; he seems to have called in artists from Champa and Java, thus giving to Khmer art a distinct new impetus. At another site, Sambhupura (Sambor), he built temples with sculpture based upon the old Funan–Chenla tradition. At Amarendrapura, about 800, he built a brick pyramid—an artificial mountain—to support a quincunx of temples.

It was Indravarman I (877–889) who laid the foundations of the fabulous temple complex known as Angkor. His plan was based on a rectangular grid of reservoirs, canals, and irrigation channels to control the waters of the river system. Later kings elaborated this original design to a colossal scale. Indravarman built the first great works of Khmer architecture: the Preah Ko, at Roluos, and at Angkor his temple mountain, the Bakong, ornamented with sculpture. Successive kings built their own temple mountains there, including the Bakheng (*c.* 893), Pre Rup (*c.* 961), the Ta Keo (*c.* 1000), and the Baphuon (*c.* 1050–66) and culminating in Angkor Wat, built in the first half of the 12th century by Suryavarman II. After a disastrous invasion by the Cham, Jayavarman VII undertook the most ambitious scheme of all, the Mahāyāna Buddhist Angkor Thom and Bayon (*c.* 1200). Thereafter, for a variety of reasons, including conquest by the Thai, no more large-scale work was done by Angkor and the country became Theravāda Buddhist. The modern dynasty has adapted remnants of traditional splendour, and the craftsmen of Cambodia have remained capable of work in the same vein as but often superior to the Thai.

Hindu Javanese art continued to be made under the eastern Javanese dynasties (1222–14th century), although their structures are not nearly as ambitious as the central Javanese works. There are many temple enclosures and volcanic bathing places with modest stonecut architecture. Some of the stone sculptures from these sites, however, are now world famous. In the 20th century the east Javanese tradition still survives, modified by folk elements, in Bali, to which the east Javanese Hindu kings retreated in the 16th century to maintain their religious independence in the face of Muslim expansion. Muslim monuments in the form of mosques and tombs are found in various parts of Indonesia. They adapt older forms of Indonesian art.

In 1056 the great Burmese king Anawrahta decreed Theravāda Buddhism to be the religion of his country, replacing earlier cults. He removed the Mon monks and artists from the capital of the old Mon kingdom in southern Burma, transporting them to his own northern capital, Pagan. There they built a city, with many large brick and stucco temples (pagodas) based on Indian patterns, that remains one of the most impressive sites in Asia. The Mongol invasion of 1287 put a stop to work there.

The Mon kingdom, Dvaravati, of southern Thailand, was annexed to the Khmer Empire in the 11th century and Khmer imperial shrines were built there. After the decline of the Khmer and the Mongol invasion of 1287, a powerful alliance of racially Thai kings established the first major Thai empire, retaining Theravāda Buddhism as the state religion. Thailand was divided into two principal regions, northern and southern, with capitals, respectively, at Chiang Mai and Ayutthaya, possession of the trade city of Sukhothai being an issue between them. In all the Thai cities, brick and stucco temples were built on variants of Indian and Burmese patterns. Many fine bronze Buddha figures, large and small, were cast in canonical Theravāda Buddhist styles. Most of these figures were accommodated in monastery halls built in impermanent materials.

In both Burma and Thailand a very large number of monasteries, usually surrounding one or two principal pagodas, were constructed during the later Middle Ages and into modern times. The major cities of Rangoon, Mandalay, and Bangkok contain the most elaborate examples, although there are many elsewhere. Because the pagodas were repeatedly enlarged and redecorated and the wooden monastic buildings and their many smaller stupas continuously reconstructed and renovated, no absolute chronology has been established for the arts of this epoch.

In Laos and Vietnam Theravāda monasteries, with brick stupas, were similarly built and rebuilt of wood. An outstanding stupa is the That Luang at Vientiane, in Laos, founded in 1566 but much restored in the 18th–19th centuries. In Vietnam local variants of Chinese styles were adapted during the Middle Ages to the planning and decoration of palaces and of Confucian, Taoist, and Buddhist temples.

The primitive styles that prevailed in the Philippines were modified by the conversion of various groups—the Moro people, especially—to Islām in the 15th–16th centuries. When, in 1571, the Spanish took control, Manila became the capital of a Spanish colony, and Roman Catholic Spanish art was adopted via Mexico. A local school of church architecture and figure sculpture flourished until the 20th century, when Manila became the centre of a modern commercial society, with its attendant architecture and art.

BURMA

One date is crucial in the art history of Burma: AD 1056. In that year King Anawrahta of Pagan decreed Theravāda Buddhism to be the state religion of all Burma. This signalled the unity of what had been a divided country, consummating tendencies apparent in earlier Burmese history.

6th to 11th century. The only major Burmese art known to scholars is based upon Indian and Ceylonese Buddhist art. In the period preceding Anawrahta's decree there had been three major historical eras in what is now the country of Burma, the first two of which produced Indianized art known to scholars only fragmentarily: the rule of the racially Mon kingdom of the lower Irrawaddy (9th–11th centuries), the contemporaneous dominion of the Pyu people in central and Upper Burma, and the subsequent decisive incursion of racially Burmese people from the northeast (11th century).

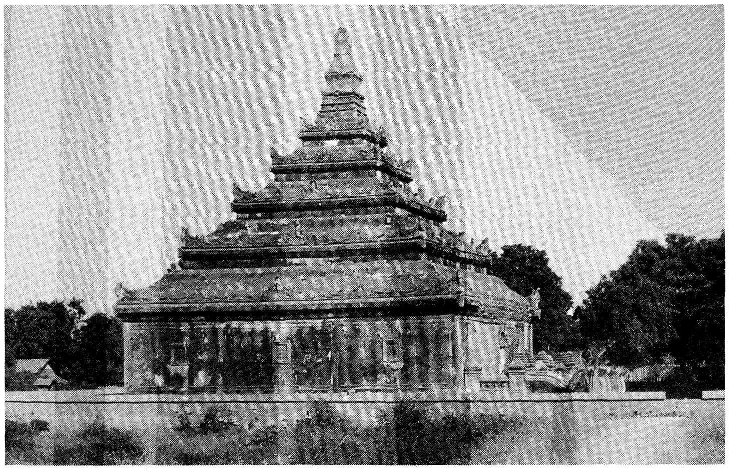
Major historical eras

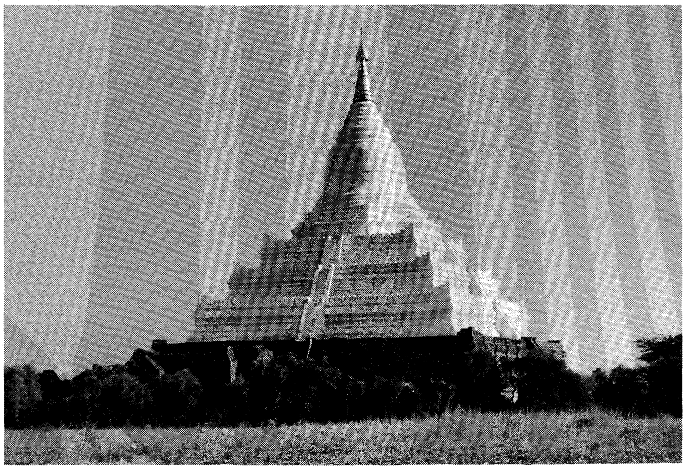
The earliest concrete evidence of Indian culture in Burma is a Buddhist inscription from Pye (Prome) dated *c.* AD 500. This and later inscriptions from the same area were cut probably in the western Mon kingdom, which followed Theravāda Buddhism and was confederated with the Theravāda Buddhist eastern Mon kingdom of Dvaravati (see below *Thailand and Laos*) in southern Thailand and part of Cambodia (AD 6th–12th centuries).

During this same period, in Upper Burma, the people called Pyu, speaking a Tibeto-Burman language and perhaps originating in Central Asia, built cities whose magnificence was known to contemporary compilers of the Chinese Tang dynasty history. In the 8th century one city was recorded as being some 50-odd miles (80 kilometres) in circumference, containing 100 Buddhist monasteries lavishly painted and decorated with gold and silver. The Pyu were in direct contact with northeast India, where various forms of Mahāyāna Buddhism, which embraced philosophies and rituals unacceptable to the Theravāda, flourished; their Ari priesthood was later proscribed by Anawrahta. Their capital city, Śrī Kṣetra (modern Hmawza, near Pye), which was once larger than even Pagan or Mandalay, has been partly excavated. Three huge Buddhist stupas—one 150 feet high—survive there. They illustrate the pattern from which all later Burmese stupas were developed. Enshrining revered relics of Buddhist saints, they consist of tall, solid brick cylinders mounted on shallow, circular, stepped plinths and crowned by what was probably a tapering bell-like pinnacle. Other excavated halls, one on a square plinth with four entrance doors, follow Indian examples. A few Hindu fragments survive as well.

The Pyu were conquered by a neighbouring kingdom, probably before AD 900. During the following century their terrain and cities were infiltrated by the racially Burmese people. These people were of common tribal stock with the Thai and northern Vietnamese and were probably on the move under pressure of the Chinese colonization of their home terrain around the Gulf of Tonkin. They were converted to Buddhism by the Pyu and later by the western Mon; but they never completely abandoned their own original cult of nature spirits, known today as the *nats*. The *nats* are a mixed collection of spirits that act supernaturally, each according to its character. The *nats* were worshipped with orgiastic ceremonies and trance rites of spiritual possession. Certain mountaintops were sacred to them. Even in the 20th century the *nats* exert a powerful influence on the lives of the ordinary people; every village has its own *nat* house—a fragile pavilion built into a tree, after the pattern of the tribal house, and adorned with shreds of coloured cloth, glass, and other offerings. The Buddhist temple in Burma is conceived essentially as an enormous *nat* house, a section of the domain of the

The *nats*





Burmese architecture.
(Left) Library, Pagan, *c.* 1058. (Right) Shwesandaw *cetiya,* Pagan, 11th century.
Louis Frederic—Rapho/Photo Researchers

spiritual located upon earth. And, since the Buddha was adopted as the last and greatest of the *nat*s, the same symbols of supernatural splendour that adorn the *nat*s adorn the Buddha's images, and a *nat*-like spirituality attaches to the ubiquitous monks in whom the presence of Buddhism is experienced as an everyday reality.

11th century to the present. When King Anawrahta came to the throne, he captured the Mon city Thaton in Lower Burma and carried off its royal family, many skilled craftsmen, and most of the Theravāda monks to his own northern city of Pagan. The king recognized the superior culture of the Mon captives; he established their main form of Buddhism by decree and gave them the task of organizing and civilizing the new united Burmese kingdom and producing for it a Buddhist art. Under Anawrahta's successor, links with the Buddhist homeland were forged. Embassies were sent to Buddh Gaya, in Indian Bihār, and the great Mahābodhi temple there—marking the spot where the Buddha achieved enlightenment—was restored with Burmese money and somewhat in Burmese taste. A smaller copy, with its large rectangular block crowned by the characteristic pyramidal, storied tower, was built at Anawrahta's Pagan. It is here that the greatest achievements of western Mon art—a splendid profusion of architecture and decorative work—are probably to be found. After 1287, when Burma was sacked and garrisoned by the Mongols, new construction at Pagan was virtually abandoned.

Pagan

In Pagan (founded *c.* 849), architecture is the dominant art; except for the big brick icons, mostly ruined, sculpture and painting play a subordinate role. Pagan contains the largest surviving group of buildings in brick and plaster of the many thousands that once stood in various parts of South Asia. The remains at the site, all religious buildings of one kind or another that must once have been surrounded by dense building in perishable materials, are in varying states of preservation. The inscriptions they bear indicate that royal devotees often turned their palaces over to religious use; so it is likely that palace and monastic architecture were very close in style. A few structures still stand that belong to the period before Anawrahta, some of them inspired by Mahāyāna Buddhism and one—the Nat Hlaung Gyaung (*c.* 931)—by Hinduism. Flanking the Sarabha Gate is a pair of small *nat* shrines with pointed, open windows—the earliest in Burma, perhaps 9th century.

The library, built about 1058 to house the books of one of the Buddhist monasteries, is one of the most important buildings in Pagan. It is rectangular, with a series of five stepped-in, sloping stone roofs crowned by a rectangular tower finial. The concave contours of the roofs are characteristic of much Burmese architecture. The eaves and corners of all the tiers are adorned with the typical Pagan flame ornament, or antefix.

The *cetiya*

There are other buildings of the same general type among the ruins of Pagan. Far the most numerous and important, however, are the buildings—called *cetiyas*—that combine the attributes of stupa and shrine. These have a history and a line of evolution of their own, which can be traced from the Pyu stupa to the huge structural temple. The normal stupa, derived from the early medieval Indian form, is a tall structure consisting of a solid dome set on a tiered square plinth (often with miniature stupas at the corners) around which the faithful may perambulate. The dome is surmounted by a *harmikā,* which resembles the small railed enclosure found on the oldest Indian stupas. In Burmese stupas, however, the *harmikā* becomes a decorated cubical die, above which is a circular pointed spire; in memory of its distant origin in India, the spire is horizontally flanged (rimmed) with moldings in a series of honorific umbrellas of decreasing size. In later practice *harmikā* and umbrella spire become a single architectural unit. The Burmese stupa dome, based on the tall, cylindrical Pyu prototype, has a spreading concave foot resembling a bell rim. The Lokananda and Shwesandaw at Pagan are two well-known examples. Because in recent times they have been coated with plaster, the finely detailed brick carving characteristic of early Pagan architecture has been obscured. Such carving is beautifully exemplified in the Seinnyet temple at Myinpagan (11th century).

Anawrahta's type of *cetiya* followed the general form of the early Pyu stupa. The main point of evolution was in the progressive elaboration of the terraced plinths on which the dome stands. The plinths became virtually sacred mountains, with a series of staircases running from terrace to terrace up each of the four sides. Perhaps inspired by vanished work in contemporary late-11th-century India, the Burmese began to open up the interior of the terraced base of the stupas with wide corridors and porticos, converting it into a roofed temple. The cylinder of the stupa dome was carried down through this temple space to its floor. Four large Buddha icons were added to the lower part of the dome, facing the four directions. Once this conception had evolved, it was possible to create around the central stupa a broad circuit of roofed enclosures, which from the outside would still suggest the traditional pattern of the stupa standing on its raised terraces, while the interior could be used for ceremonial, as in a true temple. Sculpture and painting, decorating the internal halls, corridors, and doorways, recounted the life of the Buddha and presented the example of his previous virtuous incarnations. The most famous example of this type of *cetiya* is the great Ananda temple at Pagan (dedicated 1090). It is still in use, unlike most of the old temples there, and so is kept in repair; it is painted a blazing white with lime stucco—which has, of course, obscured the finer detail of its old architecture. Its plan is square, with a broad, four-pillared porch hall added to all of the four doors in the four faces of the square. Its tower is a curvilinear pyramid resembling eastern Indian Hindu temple towers, and its enormous brick mass is pierced with two circuits of vaulted corridors. The sloping, curved terrace roofs have an elegant overall concave profile and flame antefixes along all the eaves.

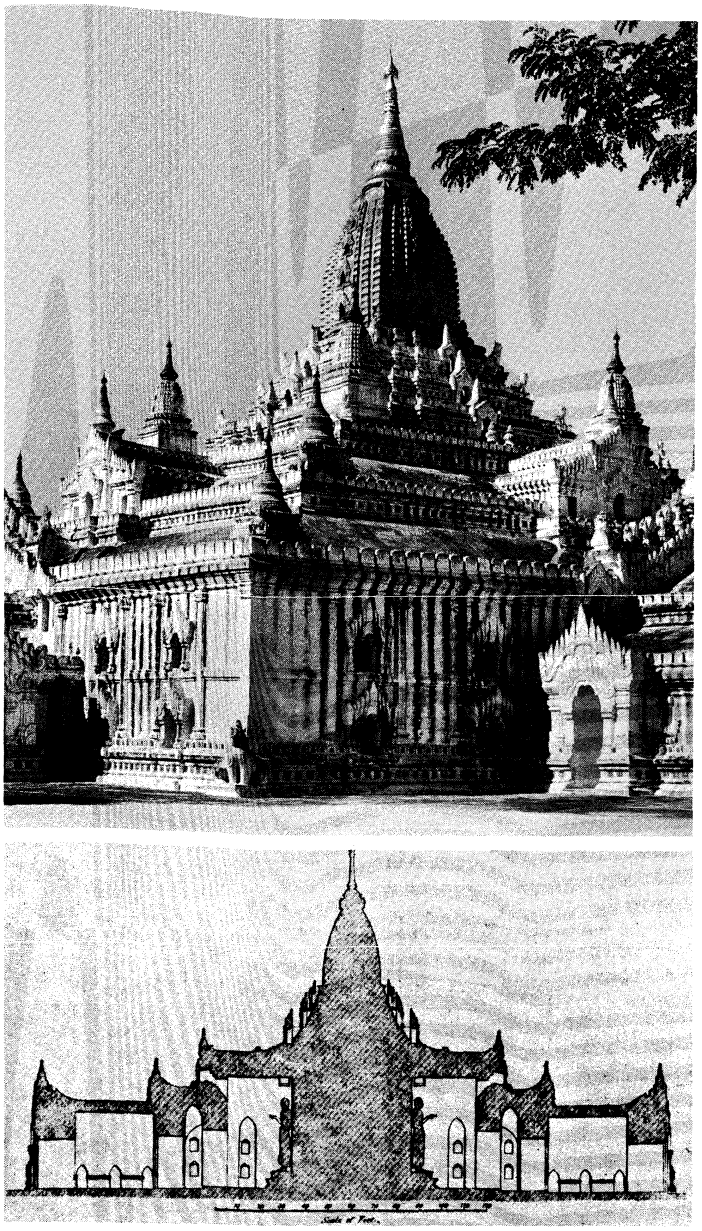
As time went on, Burmese brick and stucco architecture developed principally through the stiffening of masses into rectangular blocks and through the elaboration and often the coarsening of its ornament. The 13th-century Gawdawpalin temple at Pagan, for example, consists of a rectangular hall with a large closed entrance porch; the hall is surmounted by a tall but narrow second story whose decoration repeats that of the lower story; the whole building is crowned by a four-faced tower with a curved profile. Multiple moldings and decorative motifs are used as outlining elements and the doors are framed in elaborate upward-flaring hooded porches.

Until the Mongol conquest in 1287 much excellent work seems to have been done at Pagan. It is, however, impossible to form an adequate idea of the older styles of temple architecture at other sites in Burma, such as Rangoon or Mandalay. Whereas most of the temples of Pagan were abandoned early on, so that even though they may be ruined they show their original characteristics, temples in modern cities have been repeatedly and drastically restored. Old stupas may have as many as eight successive casings of brick and stucco; temple walls and doors are constantly torn down and rebuilt; and stucco surfaces may be renewed almost annually. Such attentions to a religious building are popularly regarded as acts of merit; thus, revered architectural monuments suffer continually from well-intentioned but disastrous renovation. At the big stupa sites huge numbers of pagodas are constantly falling into decay and new ones are being built at great speed. Among them are variants, whose evolution cannot at present be traced, on the basic pattern of the long tapering bell, with a variety of transverse moldings, standing perhaps on a recessed plinth. Many are covered quickly with extravagant and gross stucco ornament. Ornate flaring porches and flame finials are added to gates, wall ends, and eaves corners. A tapering slenderness is the outstanding characteristic of all the different types.

tupa sites
ı Ran-
oon and
Iandalay

The monastic architecture—patterned on the hall, with its elaborate doors—that surrounds the great stupa sites of Rangoon and Mandalay is mainly in wood, built by simple pillar and architrave construction. The roofs are steeply gabled, with multiple gables riding over each other on immense carved pillars in the larger halls. The angles between pillar and architrave and the edges of roof gables, tiers, and terraces are filled with flamboyant cartouches (scroll-shaped ornaments) of pierced work, often lacquered and gilt; thus, the whole building may be smothered in repetitive ornamental curlicues. All this ornament has an otherworldly or spiritual significance. Other stupa sites in Burma, where less money has been spent and less ornament added to the buildings, may be more beautiful to the modern eye, with only a few flamboyant antefixes pointing the gables and punctuating the eaves. All over Burma similar buildings can be found; but, while many have been listed, they have been scantily surveyed, and no real study of their complex history has yet been attempted. There may well be a substantial Chinese influence in the construction of some of the wooden halls and pavilions.

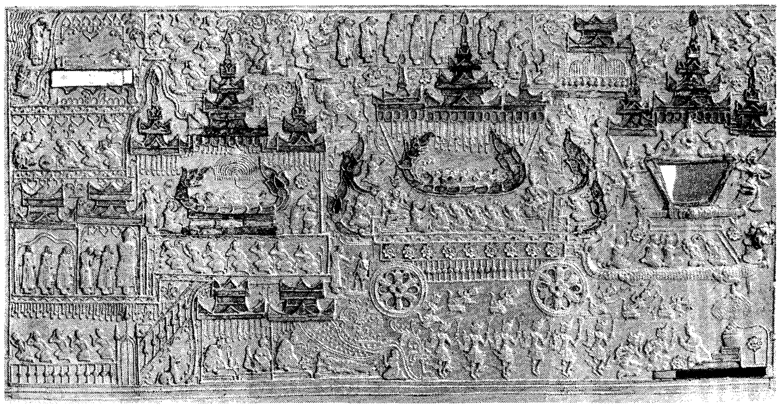
Paintings and sculpture in Theravāda Burma do not seem to have reached the same heights of achievement as in other countries of Southeast Asia. They do not show the same originality and sense of life. The temples of Pagan contain the best examples, although even these are highly schematic, reminiscent in design of eastern Indian Buddhist manuscript styles. A number of early buildings at Pagan contain fragmentary terra-cotta (fired clay) reliefs or scraps of wall painting whose individual figures display some of the sensuous charm of their Indian prototypes (it is quite likely that Indian artists worked there). The 12th-century terra-cotta panels from the elaborate facings of the Ananda temple, however, show the beginnings of the petrifaction that overtook later Burmese figurative art. Both in reliefs and in wall paintings, the figure compositions are reduced to schematic groups of the minimum number of standard human and celestial types needed to tell a moral story, without any infrastructure of significant form and execution. The colossal Buddha images enshrined in the temples were usually built of brick and finished in stucco, gilded and ornamented. Such work is still done at high speed today. The technique is not a flexible one, and the emphasis of Theravāda Buddhism on the exact imitation of ancient Buddha images gave the Burmese no aesthetic incentive to develop the expression of their figures or compositions. The repeated heavy gilding and repainting of older icons has almost entirely destroyed any formal vitality they may once have had.



(Top) Ananda temple, Pagan, Burma, dedicated 1090. (Bottom) Section of the Ananda temple.

(Top) Louis Frederic—Rapho/Photo Reseachers, (bottom) from P. Rawson, *The Art of Southeast Asia,* Thames and Hudson Ltd., London

From about 1700 to 1850 Burma excelled in the decorative arts, whose forms continually recall those of Theravāda Ceylon. Burmese woven silks and embroideries are well-known. The carved wooden screens, panels, and brackets used inside temple halls, many devoted to representing the *nat*s and the population of the spirit world, have benefitted from being outside the strict canon of Theravāda Buddhist orthodoxy. The figure types follow fluid, slightly "boneless" conventions derived from classical Indianizing dance postures. The decorative goldwares and silverwares, which use much stereotyped decorative scrollwork, are also based on standard Indianizing iconography. Perhaps the most aesthetically satisfying works of the Burmese sculptors are the reliefs ornamenting the *sū-*



Front view of a Burmese *sūtra* chest, wood and gesso, probably 18th century. In the Gulbenkian Museum of Oriental Art, the University of Durham, England. 63.5 × 120.6 cm.
By courtesy of the Gulbenkian Museum of Oriental Art, the University of Durham, England

tra chests that were used in monasteries to store the sacred texts of Buddhism. The gilt gesso (paste used for making reliefs) facings of these chests carry the schematic style of relief sculpture beyond its normal aesthetic limits. This is accomplished by the way they are compelled to set off their figures as an intelligible scheme of thin raised lines and inlay against a plain ground. The forms of the fine lacquer bowls and boxes used in monasteries, decorated only superficially with painted ornament, show the underlying formal sense of the Burmese at its clearest.

Regional Burmese art

Interesting regional types of Burmese art are those of the Shan and Karen peoples, who live in the relatively remote northern hills. These areas have often produced extremely beautiful types of domestic and religious architecture, made of wood, on stone bases. They are a simpler and more austere version of the ancient pattern that underlies the halls and pavilions of the more sophisticated recent southern temple buildings, with their steep, gabled roofs. The peoples of the north also produce a variety of decorative arts. Notable among them are the textiles, which are characterized by banding, checkering, and triangular counterchanging of brilliant colours set off against black. The woven shoulder bags, particularly, are well-known in the West.

THAILAND AND LAOS

Dvaravati Mon kingdom: 6th to 11th century. Archaeology has recovered in central Thailand substantial glimpses of the magnificent early layer of Indianized culture, which includes a religious art that was produced between the 6th and 11th centuries by the eastern Mon kingdom of Dvaravati. The art was created predominantly to serve Theravāda Buddhism. Remains of Dvaravati architecture so far excavated include stupa bases: notable examples include the Wat Phra Meru in Nagara Pathama (Nakhon Pathom) and others at Ku Bua and U Thong, some of which have elephants supporting their bases, following a pattern that originated in Ceylon. The plinths of Buddhist assembly halls, which existed near the solid monumental structures, have also been discovered. Many terra-cotta and stucco fragments of decorative surface designs and celestial figures have also been found. The Wat Pra Meru, on a plan similar to that of the Ananda temple at Pagan in Burma (see above *Burma*), probably antedates the latter's foundation (*c.* 1090). It is likely that many other ancient monuments are encased in later stupas that are still being used for religious purposes, for it was probably customary not to destroy an old sacred monument but to encase it in a new shell, maybe several times over, and perhaps to construct a small external replica of the encased original alongside.

At many sites, especially Lop Buri, Ayutthaya, and U Thong, fine Dvaravati sculptures have been found among the architectural remains. Particularly important are the seated and standing Buddha figures in stone and bronze. Many of the faces have characteristic Mon features, with lips turned outward (everted) and downward-curved eyelids marked by double channels. Some of these Dvaravati images may well have furnished models for later Khmer art in Cambodia.

Dvaravati sculpture

Dvaravati sculpture shows close relations with several Indian styles, notably those of Amarāvatī, Gupta, post-Gupta, and Pāla Bihār. It also was probably influenced strongly by the art of the enigmatic kingdom of Śrivijaya in Sumatra, as well as by central Javanese types (see below *Indonesia*). One outstanding masterpiece from Chaiya, of Dvaravati date, may well be a work produced in Śrivijaya. It is a bronze torso and head of a *bodhisattva,* for which a mid-8th-century date is suggested. The body and face are modelled with a plastic and delicate sensuousness; and the elaborate necklaces, crowns, earrings, and armlets are beautifully chased (decoratively indented by hammering). The Śrivijaya origin is made more likely by stylistic reminiscences of the sculpture of contemporary Indonesia, which was also under Sumatran inspiration.

Khmer conquest and Tai immigration: 11th to 13th century. In the 11th century Dvaravati was captured by the Khmer of Cambodia and became a province of their empire. A number of Khmer shrines, probably intended as focuses of the Khmer Hindu dynastic cult, were built in Siam (Thailand). At Phimai (Bimaya) was the most important full-fledged Khmer temple, where one of the personal cult statues of the Khmer king Jayavarman II (see below *Cambodia and Vietnam*) has been found, together with bronze images, some of Tantric Buddhist deities. At Lop Buri the Phra Prang Sam Yot is perhaps the best surviving example in brick and stucco of Khmer provincial art in Thailand, its tall towers having complex rebated (blunted) corners and its porticoes high, flamboyant pediments (the triangular space used as decoration over porticoes, doors, and windows). Wat Kukut, at Lamphun, built by a Dvaravati Mon king *c.* 1130, represents an adaptation of the Khmer stepped-pyramid temple base as pattern for the temple itself. The niches on its terraces are filled with images in a deliberately archaistic revival of the old Mon style.

Domination by the Tai people

During the period when the Khmer were taking over the southern Mon region of Thailand, the northern region was falling under the domination of immigrant racially Tai peoples. The Tais were a branch of the migrating population who invaded Burma as the Burmese and of the Sinicized Vietnamese who were then pushing southward into what is now Vietnam. The Tais seem to have professed an animist nature religion, resembling the early form of the Burmese cult of the *nat*s (see above *Burma*). This whole group of peoples originated most probably as a tribal population in the region of Tongking and Canton. In the course of their southward migrations they probably played an important role, as yet unclear, in a kingdom called Nanchao, in what is now the Chinese province of Yunnan. The rulers of this kingdom seem to have followed a Mahāyāna form of Buddhism, including the cult of a *bodhisattva* as personal patron of the king. Several smallish bronze icons of a *bodhisattva* with a nude torso and a strap round the upper belly are known from Nanchao, in a style reminiscent of the later Pallava art of the east coast of peninsular India. The date of these images is still uncertain. Tai kingdoms were gradually established further and further south. Some of their tribes gained experience of administrative techniques by living within the boundaries of the Khmer Empire, with their own chieftains under Khmer officials. When the Khmer power was broken in the 13th century, the Tai moved into central and southern Siam, intermarrying with the Mon.

The Tai people normally built in perishable materials, wood and bamboo in particular. Their animist religion, which has no canonical group resembling the Burmese *nats*, is still very much alive today. The spirits of trees need to be pacified, and the ancestors can be powerful helpers. Shamans, in a state of trance, make contact with the spirit world to perform good or evil magic. In the wooden high-gabled houses of the northern Tai (Chiengmai province), even today, ornate lintels are carved with floral relief designs to sanctify and potentiate the inner domestic part of the house where the domestic spirits live. The animist religion gave ground partially to Buddhism, which was gradually assimilated among the people, and at some date, as yet uncertain, was adopted by the greater Tai kings as a dynastic religion. With the spread of Buddhism a special religious architecture in brick and stucco was established.

)ivision of iam

The Thai kingdom: 13th to 17th century. During most of its history, Thailand has been divided into two fairly distinct regions, a northern and a southern, the capital of the north at Chiang Mai, the capital of the south at Ayutthaya. Between the two lies the great trade-route city of Sukhothai, possession of which fluctuated between the north and the south. Sukhothai seems to have been the principal focus and source of Buddhist culture in Siam, for it retained direct touch with Ceylon, which, after the decline of Buddhism in India in the 12th century, became the principal home of Theravāda Buddhism. By the 15th century the difficult art of casting large-scale Buddha figures in bronze had been mastered in the north of Siam, as well as in the south.

Sculpture. The Thai kings made repeated attempts to "purify" their conservative Theravāda strain of Buddhism, importing patterns of art along with texts and learned monks from Ceylon and trying to wean their people from worship of the spirits. To retain the greatest spiritual potency, Buddha icons in Thai temples had to be as close in type as possible to a great original prototype that Buddhist tradition erroneously believed had been made during the lifetime of the Buddha; in practice, this meant the types the local craftsmen knew as the oldest and most authentic. There were at least three major successive efforts by Thai kings to establish and distribute an "authentic" canon for the Buddha icons, which were their prime artistic concern. Each type that became canonical and was known to be magically effective was imitated repeatedly. For it was regarded as an act of merit simply to multiply images of the Buddha, whether they were to be installed in temples or not; hence, in addition to icons, enormous numbers of small images—made of many materials, from bronze, silver, stone, and wood to terra-cotta—were kept in temple storehouses. The images followed canonical patterns established for the major temple icons.

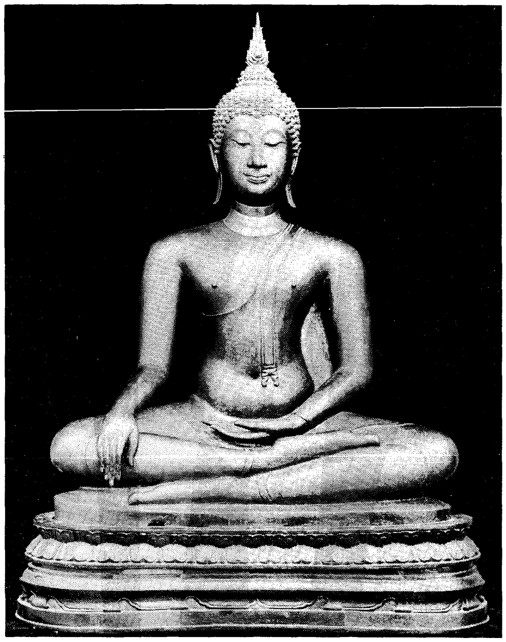
Since their work had to be as similar as possible to the oldest sacred images of which they knew, the Buddhist sculptors in Siam adhered to strict formulas and diagrams; artistic development was never a part of their purpose, though of course gradual change did occur. There is no tradition in Theravāda Siam in any way resembling the traditions of Mahāyāna art in, say, Cambodia or Indonesia, which encouraged artists to explore the possibilities of their mediums to express developing religious conceptions. Thus, Thai Buddhist sculpture consisted almost entirely of careful repetitions of the standardized types, which tended naturally, despite the artist's desire to capture an authentic sense of style, to lose their older vitality. It also happened that the three main canonical patterns often lost their individuality, blending into each other. Perhaps the best works were made in the 15th century, but work of high quality was still being done in the 16th and early 17th centuries.

Canonical types of Buddha image

The first canonical types were the Sukhothai, which seem to have been evolved in the trade-route city of Sukhothai as an attempt to capture the quality of early-medieval Ceylonese images and elements from Dvaravati sculpture. The developed versions of these types are marked by an extremely smooth, rounded modelling of the body and face, without any clearly defined planes. The outlines of hair, eyebrows, lips, and fingers are elegantly recurved, or S-curved, and the head is crowned by a tall, pointed flame finial. The entire figure gives an impression of great elegance. Full-fledged Sukhothai images of the full-round walking Buddha—an original Sukhothai invention—emphasize a kind of swaying, sinuous, boneless grace in the execution of the legs and arms. One of the most impressive colossal images of the type is the brick and stucco icon at the Wat Mahathat, Sawankhalok, another Sukhothai technical forte, dating probably to the 14th century. This type of image remained the most popular in Siam; an enormous number of imitations, of all dates, are preserved, many in Western collections.

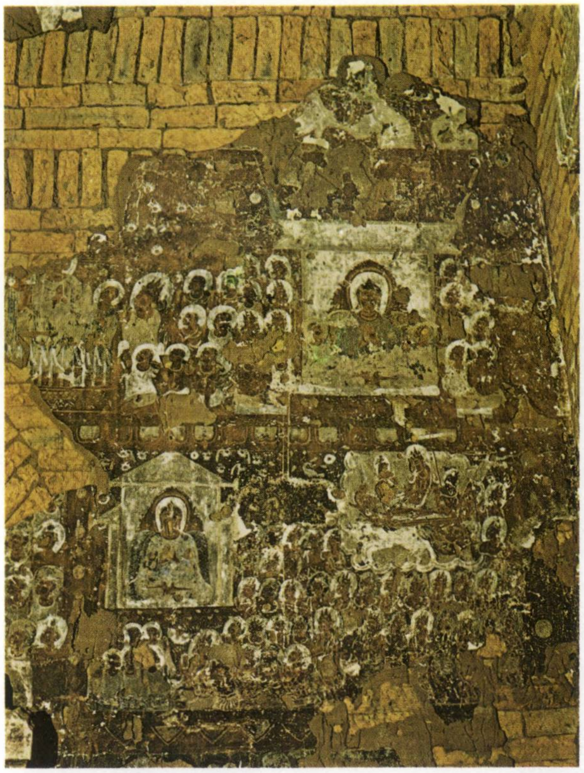
Perhaps the Buddha types most successful aesthetically were those called after U Thong. They were produced originally in the southern capital of Ayutthaya, which took over Sukhothai in 1349, and represent a fusion of the Sukhothai types with vestiges of Khmer and Theravāda Dvaravati traditions, whose Buddha types had been marked by a strong Mon sense of squared-off design and cubic volume. The latter may have been influential because they seemed to incorporate an older and more authentic tradition, since they were based upon patterns developed in eastern India, the true homeland of Buddhism. In the U Thong style the sinuous linear curves, loops, and dry ridges of the pure Sukhothai patterns are suppressed, and genuine modelling, with clearly defined planes and volumes, appears. In the northern kingdom a crude version of the Sukhothai type gained currency in the late 14th century. When, in the middle of the 15th century, King Tiloka of the northern kingdom reestablished contact with Ceylon, images seem to have been imported directly from that country. They must have shown clearly how far the Sukhothai types had departed from the type used in the Buddhist homeland, because the third Siamese icon pattern, known as the lion type, attempted to recapture the stern simplicity of the genuine Sinhalese images. Most of the best examples were made between 1470 and 1565. Limbs and bodies are given a massive cylindrical strength, and the Sukhothai elegance

By courtesy of the Breezewood Foundation, Monkton, Maryland

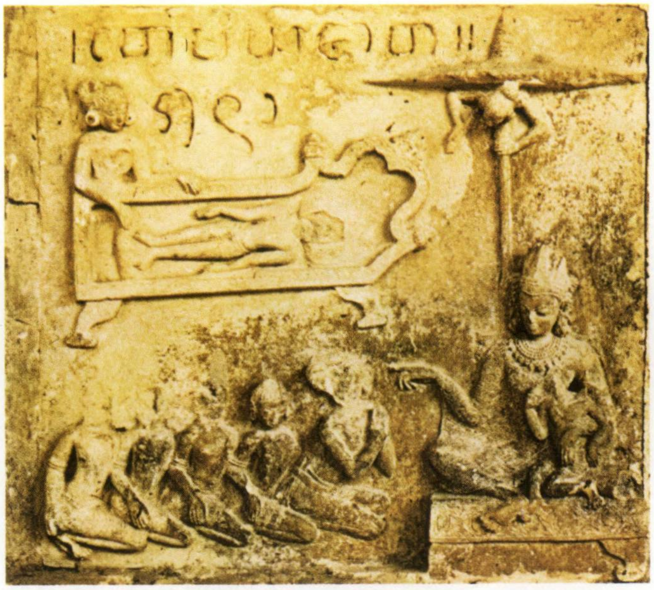


"Buddha Calling the Earth to Witness," Thai, bronze, Sukhothai high classical style, 14th century. In the collection of Prince Chalermbol Yugala, Bangkok. Height 94 cm.

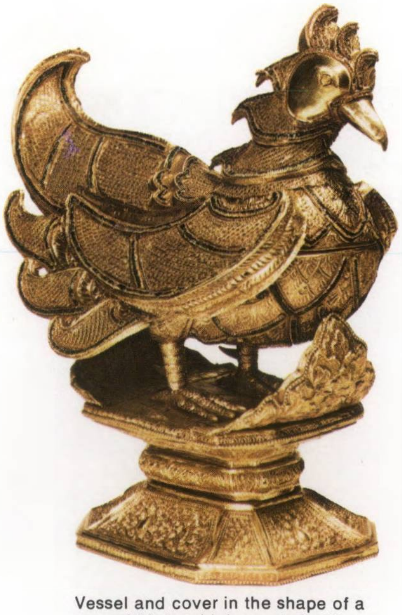
Burmese art



Fresco of the Preaching Buddha at the Wet-kyi-in, Ky-byauk-kyi, Pagan, *c.* 1113.



Terra cotta relief at the Petleik-Matangjakata temple, Pagan, late 11th century.



Vessel and cover in the shape of a sacred bird, gold decorated with filigree work and inlaid with rubies and imitation emeralds, 19th century. In the Victoria and Albert Museum, London. Height 41.5 cm.

Shwe Dagon pagoda, Rangoon, *c.* 15th century.

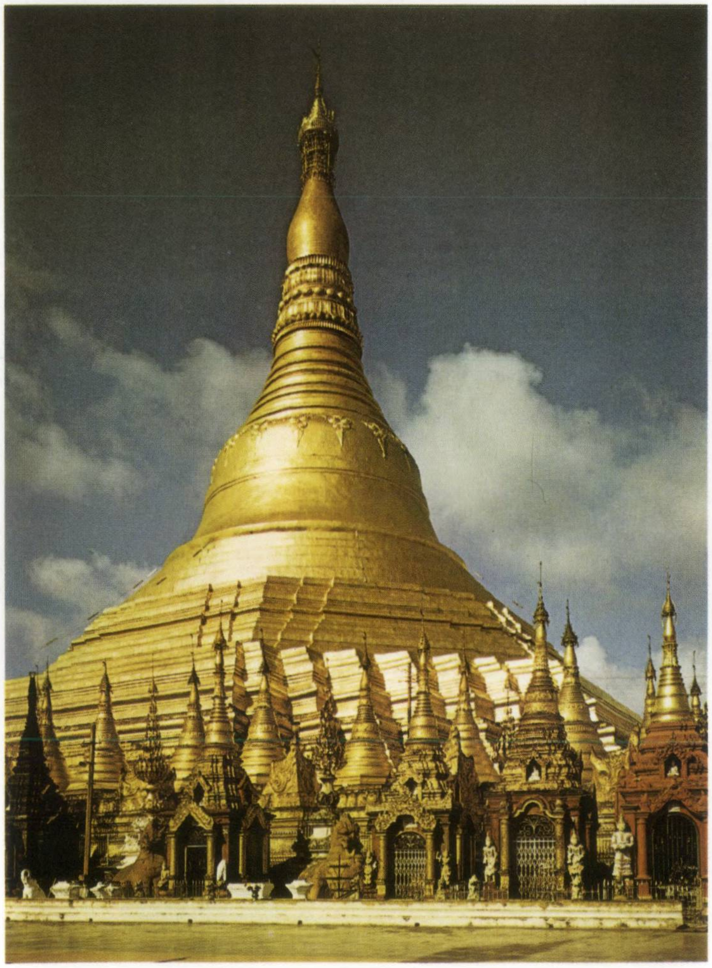
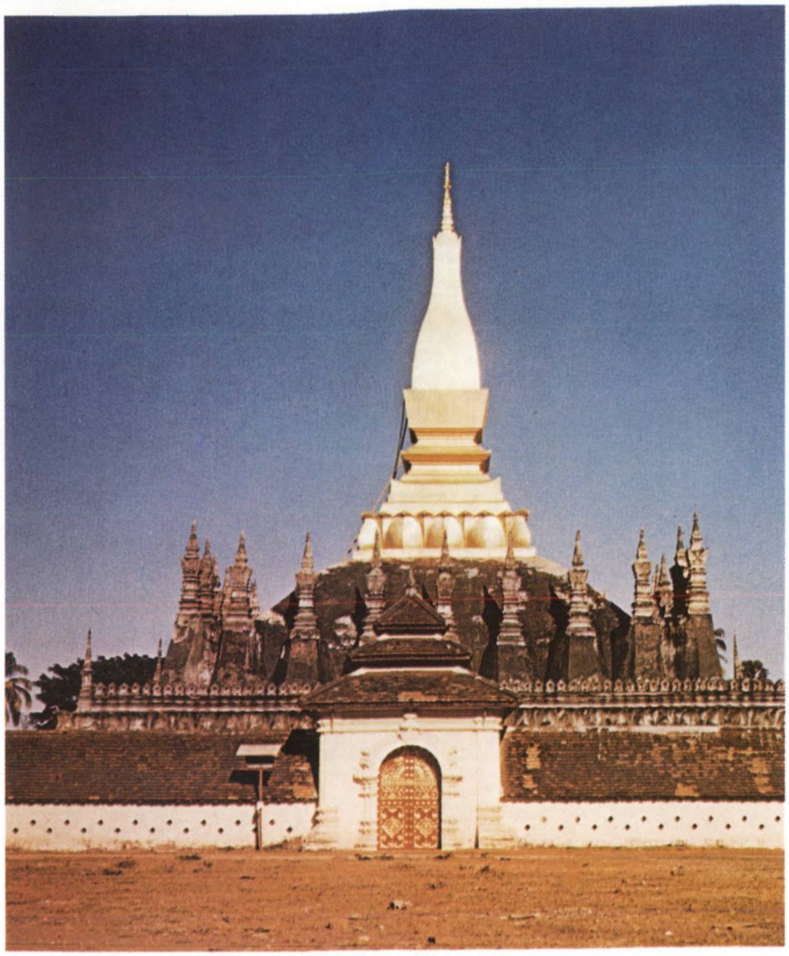
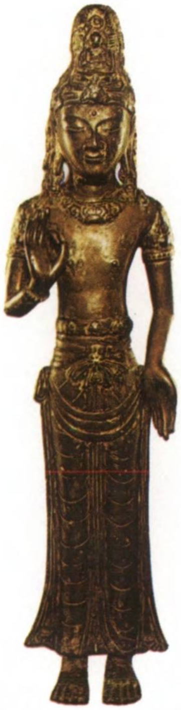


Plate 1: By courtesy of (bottom left) the Victoria and Albert Museum, London; photographs, (top left, top right) J.A. Lavaud, Paris, (bottom left) A.C. Cooper Ltd., (bottom right) R. Manley from Shostal—EB Inc.



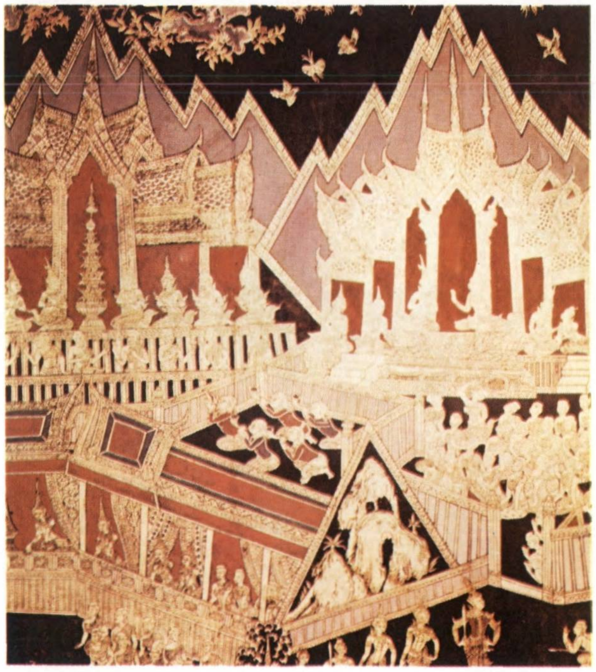
That Luang *stupa*, Vientiane, Laos, 1566, restored 18th and 19th centuries.

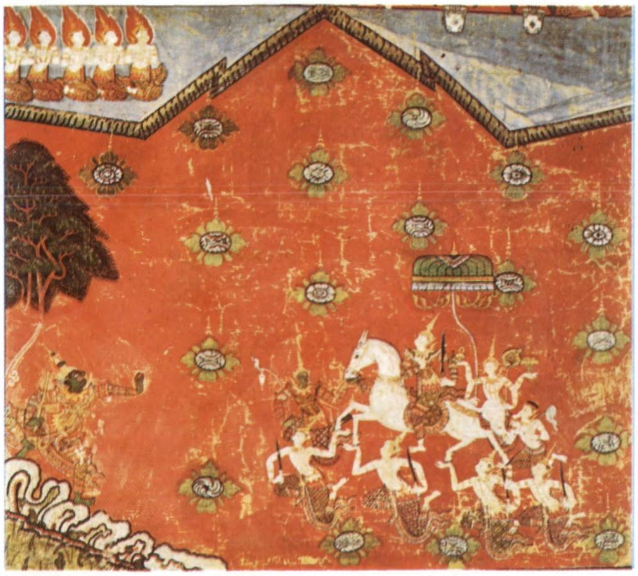


Bodhisattva from Nanchao, ancient Tai kingdom (now in the Chinese province of Yunnan), bronze, 13th century. In the British Museum. Height 44 cm.

Thai and Laotian art

Thai painted lacquer panel of a court scene, Bangkok style, middle of the 19th century. In the collection of Prince Piya Rangsit, Bangkok. Height 50 cm.

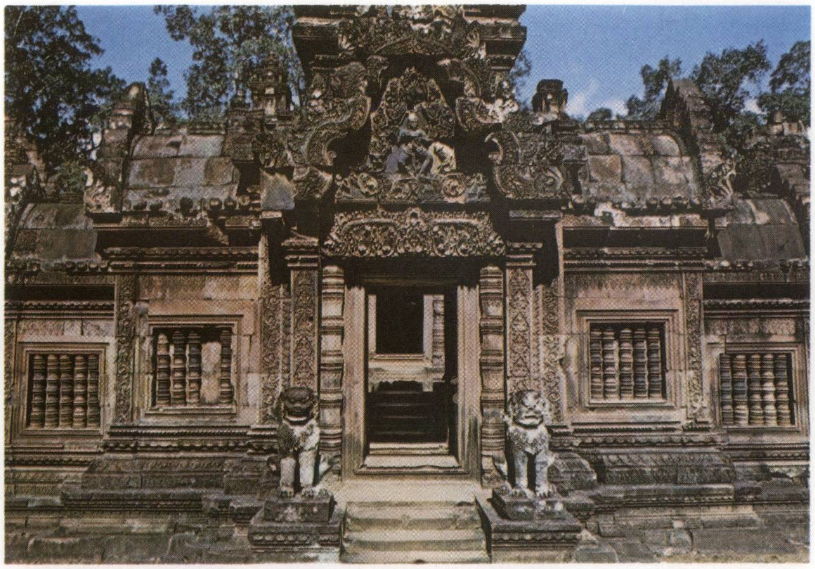




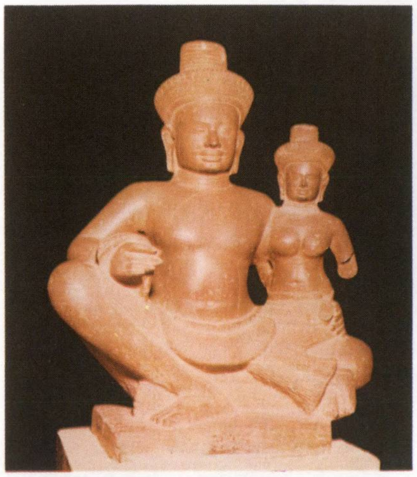
"The Great Departure of Bodhisattva," detail from "Episodes from the life of Buddha," Thai painting on silk panel, 17th-18th century. In the Musée Guimet, Paris. Detail 93 × 93 cm.

Plate 2: By courtesy of (top right) the trustees of the British Museum; photographs, (top left, bottom left) Holle Bildarchiv, Baden-Baden, (bottom right) Ciccione—Rapho/Photo Researchers

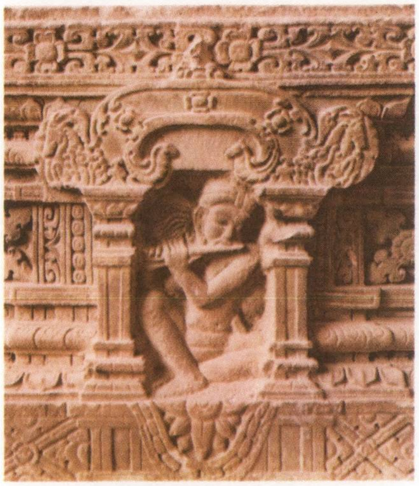
Plate 3: (Top left, centre) Holle Bildarchiv, Baden-Baden, (top right) © Pascal Grellety-Bosviel, Paris, (bottom) R. Manley from Shostal—EB Inc.



Library, Banteay Srei, Angkor, Cambodia, 10th century.



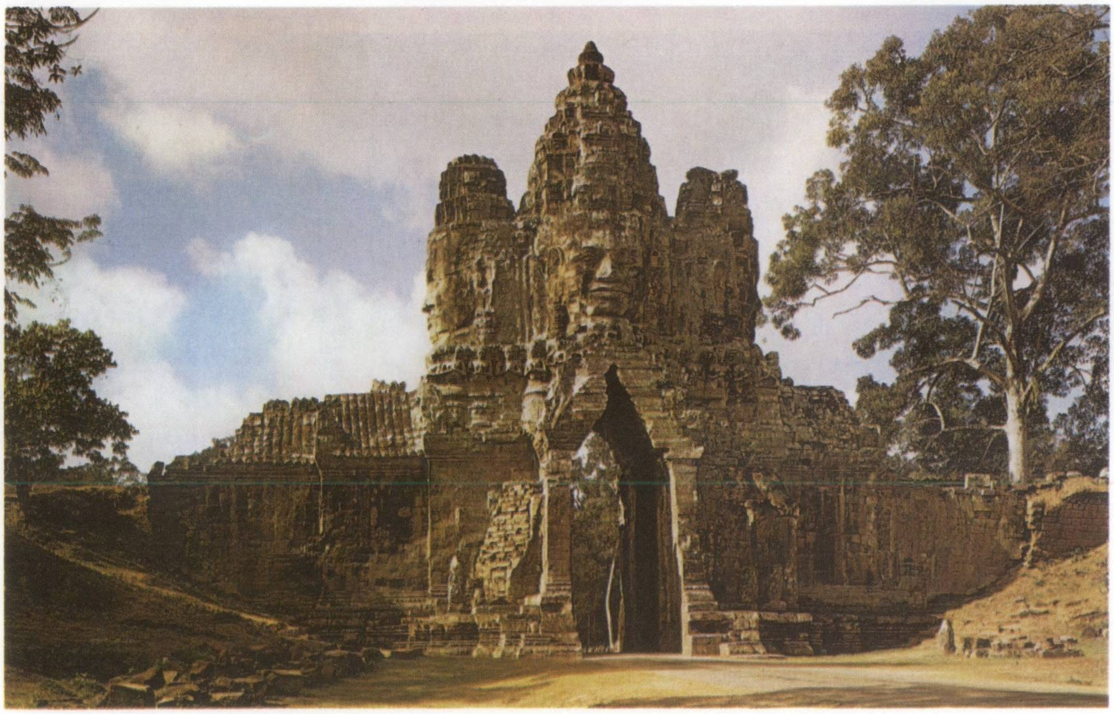
"Siva and Umā," sandstone sculpture from Banteay Srei, Angkor, Cambodia, late 10th century. In the Musée National, Phnom Penh, Cambodia. Height 60 cm.

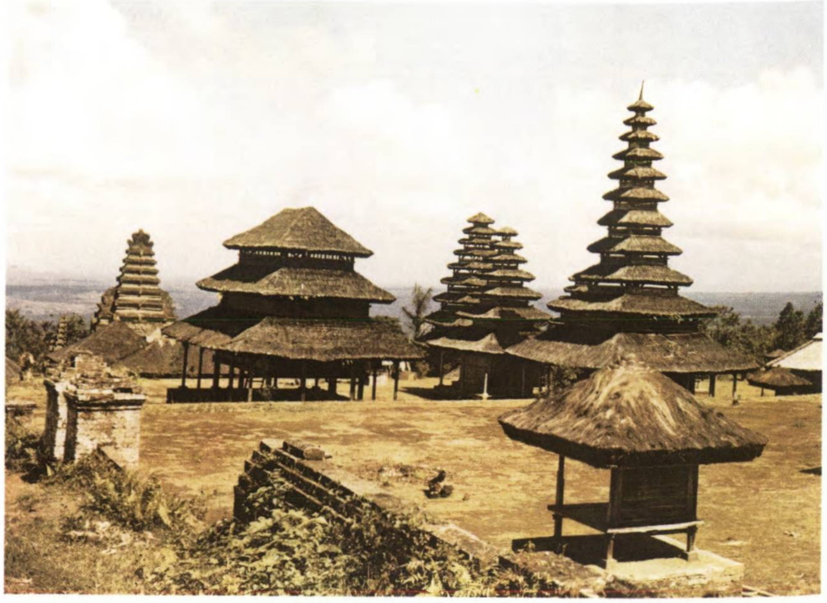


Cambodian and Vietnamese art

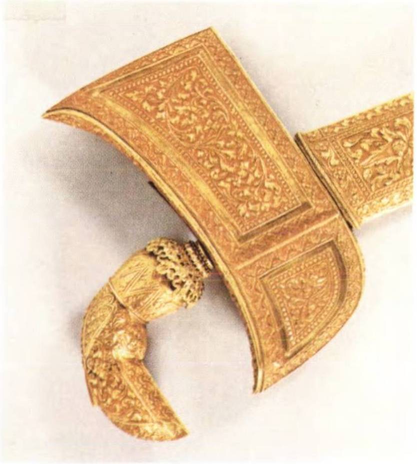
Cham sandstone panel of a pedestal altar showing an ascetic playing a flute, from Mi-son E 1, Vietnam, second half of the 7th century. In the Da Nang museum, Vietnam. Height 60 cm.

Gate at Angkor Thom, Angkor, Cambodia, *c.* 1200.



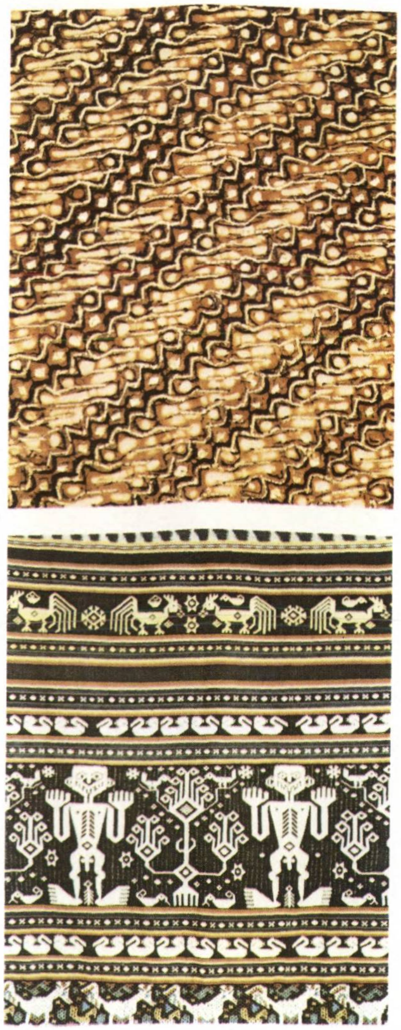


Pura Besakih, Gunung (mount) Agung, Bali, 14th century. In the centre is the eleven-storeyed *meru* dedicated to Siva.

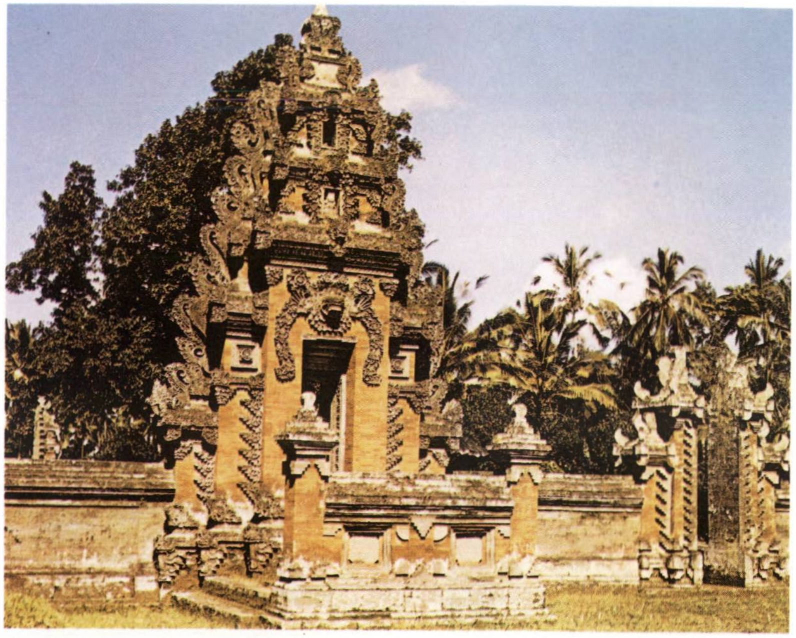


Gold kris, embossed scabbard and grip, from southern Celebes. In the Royal Tropical Institute Museum, Amsterdam. Overall length 40.5 cm.

Indonesian art



Indonesian textiles. (Top) Javanese batik textile accented with gilding. In the Royal Tropical Institute Museum, Amsterdam. (Bottom) *Ikat* cloth from Sumba Timur, Lesser Sunda Islands. In the J. and R. Langewis Collection, Castricum, The Netherlands.



Main entrance to the Pura-desa Pasaban, Bali.

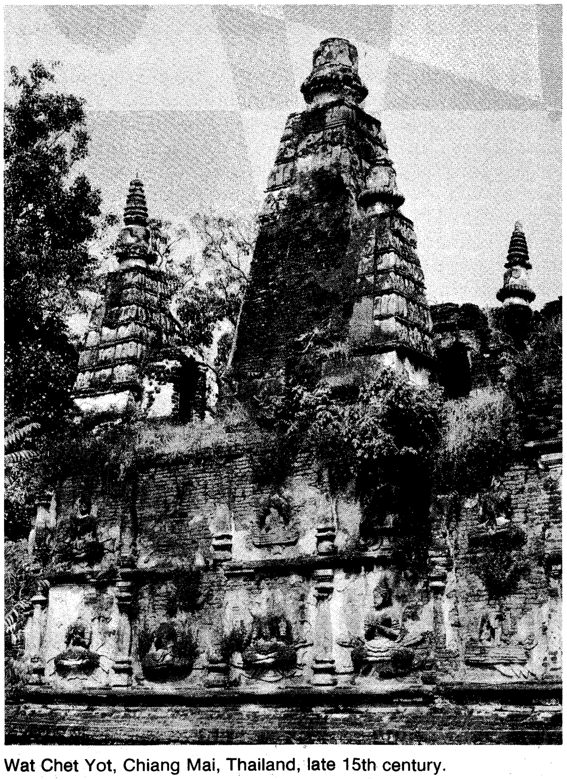
Plate 4: By courtesy of (centre left, centre right) the Royal Tropical Institute, Amsterdam; photographs, (top, bottom left, bottom right) Holle Bildarchiv, Baden-Baden

is eliminated. It seems, however, that the native Thai genius is for the sinuous and unplastic curve, which may have expressed for them the same spiritual unworldliness as it did in Burmese ornament. Thus, in later examples reminiscent of the lion type, the curvilinear patterns of the Sukhothai style reassert themselves with more or less emphasis; and by the end of the 16th century the lion type had lost its distinguishing features and merged into the run of Sukhothai patterns.

Architecture and painting. There are as yet few results of authenticated research available concerning the history of architecture during the early period of Thai supremacy. Many monasteries contain stupas, or *cheddis,* that either originated or were renewed in this period; but most of the monasteries themselves have been repeatedly overworked. Building complexes seem to have developed by accretion, rather than by the studied working out of space articulations. The oldest building in Ayutthaya, dating from the early 13th century, is the Wat Bhuddai Svarya, a towered shrine, approached by a columned hall. From the late 14th century onward, Sukhothai influence seems to have predominated everywhere. The architectural types included a bell-shaped reliquary stupa with a circular flanged base and onion finials, reminiscent of combined Ceylonese and Burmese patterns; a stupa raised upon a cylindrical shrine as its drum; and a shrine with a plinth faced with images (usually later additions) above which rise one or more pyramidal towers reminiscent of the tower of the Mahābodhi temple at Buddh Gaya in Bihār, India. An example of the third architectural type is King Tiloka's late-15th-century Wat Chet Yot at Chiang Mai, which has one large and four smaller pyramids mounted on a main block.

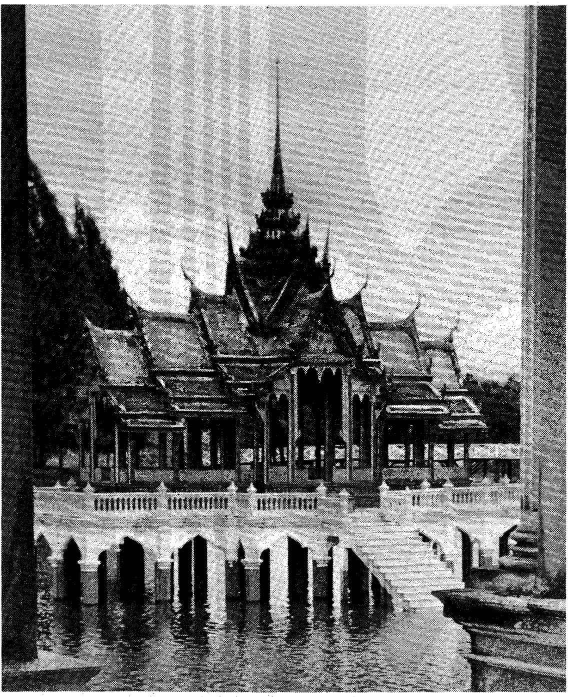
14th-century architectural types

Louis Frederic—Rapho/Photo Researchers



Wat Chet Yot, Chiang Mai, Thailand, late 15th century.

The Thai kings also adopted something of the personal funeral cult of Khmer Angkor (see below *Cambodia and Vietnam*), for a custom grew of building bell-shaped brick stupas—which had earlier been used only for the relics of Buddhist saints—as the kings' tombs, each approached by a colonnaded hall and surrounded by smaller stupas or shrines. In many of the brick and plaster or wooden monastic buildings of more recent centuries, such as the Wat Po in Bangkok, one can trace the distant influence of the Khmer styles of Angkor. Tall, gabled roofs, with steps and overlaps, the gables adorned with flame finials, are typical, exemplified by the Water Pavilion at Bang Pa-in.



Water Pavilion, Bang Pa-in, Thailand, 1294.
Luc Bouchage—Rapho/Photo Researchers

Thai painting of the early period (13th–16th centuries) demands a great deal more research and study than it has yet received. Although it is, of course, devoted to the canonical iconography of the Theravāda, its fluent and relatively unschematic outline shows that it retained much of the original inspiration visible in the earlier work at Burmese Pagan (see above *Burma*). The oldest examples of Thai painting are the much-ruined frescoes in the Silpa cave, Yala, and some engraved panels from Wat Si Chum, Sukhothai, dated to 1287. Later paintings (dating to the 1420s) in the inner chambers of the Wat Rat Burana and Wat Mahathat at Ayutthaya show strong Chinese and perhaps Khmer influence in their high perspectives and landscape backgrounds with animals, combined with the native Thai clear outlines and bright, flat colours. By the 17th century at, for example, the Wat Yai Suwannaram at Phet Buri, large mural compositions—such as an elaborate scene of demons worshipping the Buddha—were being undertaken. In this later painting, theatrical stereotypes from the Thai dance-drama exerted a strong influence in the rendering of figures.

18th century to the present. In the 18th century the Burmese invaded and conquered Siam. The Burmese king—in expiation, it is said, of his war guilt—ordered the construction of many Buddhist buildings in the current Burmese style (see above *Burma*). These made their impact on Thai art, and the gaudy gilding and inlay characteristic of late Burmese ornament were widely adopted. When the capital was moved to the present Bangkok, in 1782, no substantial artistic development took place, though large pagodas were built and filled with rows of images, many in gilt wood. A highly ornate interpretation of older, airily flamboyant Burmese decorative styles, featuring curved "oxhorn" projections, blunted the edge of architectural and sculptural quality. Without exception, the new large-scale icons were dull and inferior works of art; and the monstrous guardian figures of spiritual beings and lions decorating the major shrines are fantastic rather than aesthetically valuable. In the painting of wooden panels, some of them votive, and of historical manuscripts, the Thai retained a good deal of their older vigour. The figures illustrating legend and history are based upon the unworldly stereotypes of the court dance.

Burmese influence on Thai art

In addition to the incorporation of European motives, many buildings and their ornamentation in Bangkok have a strongly Chinese flavour. This is attributable partly to the influence of the large expatriate Chinese population

Sawankhalok, Sukhothai, and Chiang Mai pottery

living there and partly to the influence of earlier expatriate Chinese craftsmen. The early 20th-century Pathamacetiya at Nagara Pathama (Nakhon Pathom), which is entirely orange, is a fine example of the many *cheddis*. Some tiles were certainly imported from China, but others were descendants of the fine pottery (of basically Chinese inspiration) that was produced at the kilns of Sawankhalok during the 14th and 15th centuries by expatriate Chinese craftsmen. This pottery imitated in its own materials Chinese Yüan dynasty (1279–1368) Tzu-chou and celadon wares (stonewares and porcelain with a glaze developed by the Chinese) with underglaze ornament and blue or brown painted decoration. Similar wares were made in the 15th century at kilns at Sukhothai and at Chiang Mai. Some of these pieces are, in their own idiom, as fine as native Chinese work. Later, during the 18th and 19th centuries, somewhat garish, flamboyant Ayutthaya figure designs in polychrome were applied to rice bowls and other vessels.

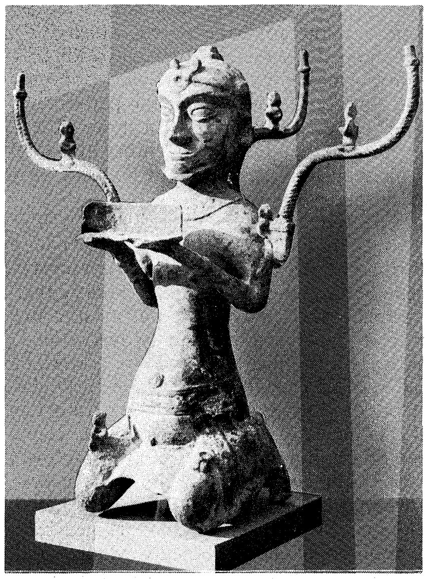
Laos. The kingdom of Lan Xang (Laos) was founded in the mid-14th century and ruled by Buddhist Thai. At the northern capital, Luang Prabang, the influence of the northern Thai city of Chiang Mai predominated; in the southern capital, Vientiane, a mixture of Ayutthaya and Khmer motives prevailed. In Laos there is no stone and little brick architecture. The most impressive single monument, the brick and stucco That Luang in Vientiane, founded in 1586 but much restored, is a stupa, shaped as a tall four-faced dome on a square plinth enclosed in a court; the dome is crowned with an ornate spire and encircled by a row of similarly shaped spires. The architecture of monastic halls also follows the Thai pattern; very steep multiple-gabled roofs, gently curved and overhung with long eaves, are carried on brick or wooden pillars and adorned with flame finials. Buddha figures, preserved in some of the monasteries, are based on northern Thai versions of Sukhothai types; some may be as early as the 17th century. The schematic paintings on monastery walls are in versions of the later Thai styles. In the northwest a strong influence from late Burmese art can be found in Buddhist images made to serve a religion that was far closer to the original Thai animism than to true Buddhism.

CAMBODIA AND VIETNAM

Paleolithic tools similar to types found in India have been found in Cambodia (Kampuchea) and Vietnam; and it is possible to trace the movement of population or culture groups, some of whom probably migrated onward by sea from Southeast Asia into the islands. The important group of speakers of Mon–Khmer languages may conceivably have been the people who produced the megalithic monuments in Cambodia and Laos, which include colossal stone burial urns, dolmens, and menhirs, perhaps associated with the many circular earth platforms as yet unexcavated (see above *General development of Southeast Asian art*). Probably contemporaneous, at least in part, with the Neolithic Mon–Khmer culture is the culture known by the name of its richest, most northerly site, Dong Son, on the coast of the Gulf of Tonkin in northern Vietnam. It seems probable that the chief influences on this culture came from southern China. Many sites, ranging in date from about the 4th to the 1st century BC, stretch southward from the coast of Vietnam, as far as northern New Guinea. The islands of Indonesia and parts of Malaya may have been the principal location of the Dong Son culture.

Bronze drums

The most impressive bronze objects produced by this culture are large drums, which seem sometimes to have been buried with the dead. Splendid examples have been found in Java and Bali (see below *Indonesia*). These and many other bronze objects, such as superb funeral urns with relief ornament based on squared hooks, lamp holders, dagger hilts in the form of human figures, and other weapons, are of extremely high quality. Their ornament was produced by the Chinese casting technique of incising the patterns into the negative mold that was to receive the molten bronze; much of it suggests a parallel version of contemporary Chinese ornament of the Ch'in period (221–206 BC). From the figures and objects represented in this bronze work, it seems that the Dong Son culture had much in common with that of some of the peoples of the Melanesian islands today. The culture knew large seagoing canoes, houses similar in structure to those still common among peoples of Melanesia, and ceremonies that the Melanesians might recognize. It is probable that one group of their descendants, which retained its identity, is known to the history of this region as the Cham (see below *Vietnam kingdom of Champa*).



Lamp holder from Lach Truong, Vietnam, bronze, Dong Son culture. In the National Museum, Hanoi. Height 33 cm.

By courtesy of the Fine Arts Conservation, Ministry of Culture, Hanoi

Although many peoples isolated in the densely forested uplands also retained a tribal identity, by far the most important art was produced in the two Indianizing empires: Khmer, in Cambodia, with its linear predecessors the kingdoms of Funan and of Chenla (names they were given by Chinese historians), and the Cham, in Vietnam.

Cambodian kingdoms of Funan and Chenla: 1st to 9th century. Funan, which was in existence by the 1st century AD, was the earliest of the kingdoms that arose along the lower reaches of the Mekong River in response to Indian ideas. Its influence probably extended over long stretches of the coast of the Gulf of Siam, even as far as southern Burma, and corresponded with the range of the Mon peoples. Lying on the natural focus of land and sea routes linking eastern India and southern China to the islands of the South Seas, its geographical situation was ideal for a kingdom whose wealth was based on trade. At Funan sites even Roman, Ptolemaic Egyptian, and Sassanian Persian objects have been found, giving an idea of the extent of its trading interests.

The founder was probably a Brahmin trader from western India; for a local legend describes how the first king, a Brahmin, married the daughter of a local serpent deity, so establishing the ruling family. Serpents (*nāgas*) in Indian mythology are the spiritual patrons of water; and the basis this kingdom laid for later kingdoms in the same area was an elaborate system of waterworks, canals, and irrigation channels controlling and distributing the waters of the Mekong River. Contemporary Chinese accounts refer to cities with splendid wooden buildings, carved, painted, and gilded. But nothing remains, save a few foundation piles. Probably during the 6th century AD the kingdom called Chenla was established in the upper-middle reaches of the Mekong River, in what is now Laos. The kings who ruled in Chenla were descended from the kings of Funan and took over much of the Funan domain. It seems that disastrous floods had finally ruined Funan, which had previously suffered from Indonesian aggression, and that the shift of power to Chenla represented a recognition of temporarily insuperable geographical difficulties.

Stone sculptures

Culturally, Funan and Chenla are continuous. Their artists produced some of the world's greatest stone sculptures, most of which are large, freestanding icons, carved in sandstone. Intended to be installed in brick-built shrines, none of which survive, they usually represent the two major deities of Hinduism, Śiva and Vishnu. Sometimes both deities are combined into a single figure called Harihara; the right half of the body is characterized as Śiva, the left as Vishnu. A few examples of other figures are known, including some magnificent images of goddesses. The style of these sculptures is marked by an extremely smooth, continuously undulating surface, given strength by a system of clear, broad frontal planes and side recessions related to the foursquare block. Such images were meant to demonstrate the power and charm of a heavenly prototype to whom an earthly king appealed for his authority. The earliest images belong to the 6th century, and the series continues into the 9th.

In later Khmer times each king and sometimes each member of a royal house had statues of himself or herself in the guise of a patron deity set up in the family temple precinct. That the same custom prevailed in 6th-century India, particularly in the southeast, suggests that some of the early Funan and Chenla sculptures may have served the same function. A number of figures are Indian in style—some more markedly than others, which is probably more than a matter of date; for it is quite likely that Indian craftsmen occasionally travelled into this region to work. The style of the greatest of these early sculptures, however, is not Indian at all.

Sandstone lintels

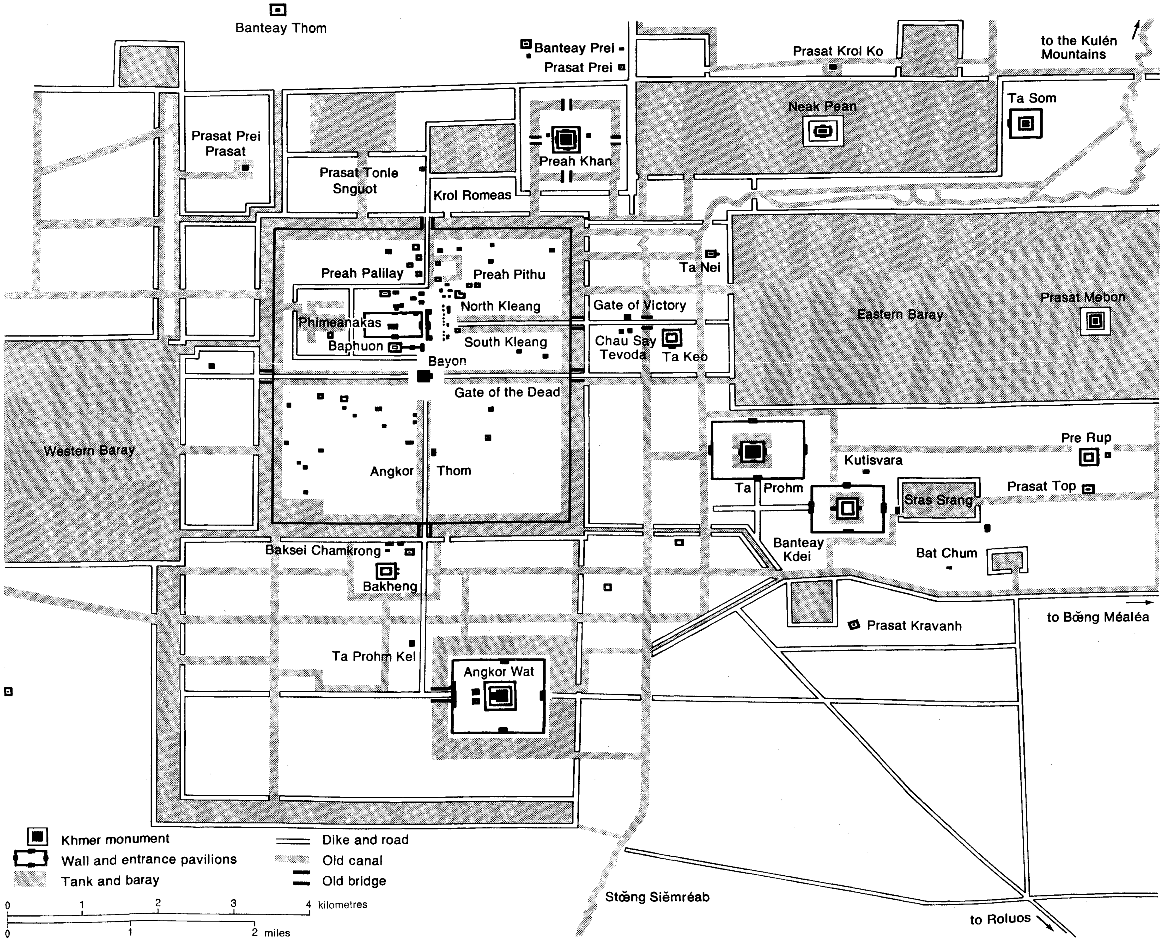
Similarly non-Indian are the magnificent sandstone lintels made for the doorways of the vanished brick shrines. Although distantly related to Indian prototypes of the 1st and 2nd centuries AD, they appear as full-fledged Indochinese inventions and may well have been developed in combination with a native conception of the lintel as a special attribute of the spirit shrine (see above *Thailand and Laos*). They are carved in relief with designs based on a pair of monsters, one at each end, which are linked by an ornate arched or lobed beam. The beam is adorned with figures inside foliate plaques, a long sequence of elaborately carved swags of jewels hanging beneath them.

Among the Funan–Chenla sculptures are a few Buddhist icons executed in sandstone, markedly less sensuous than the Hindu figures and close to the styles of Dvaravati (see above *Thailand and Laos*), though a number of small Buddhist bronzes representing *bodhisattva*s approach the delicacy of the Hindu work.

Kingdom of Khmer: 9th to 13th century. Late in the 8th century the kingdom of Chenla declined politically, perhaps because of dynastic disputes with the rising power of Indonesian kings, who were themselves also descended from the original royal dynasty of Funan. It seems that the Indonesians gave some assistance in establishing a new kingdom in the northern part of what had been the territory of Funan. In 802 a Khmer king, who took the title of Jayavarman II, established his capital near Phnom Kulen, about 20 miles (30 kilometres) from Angkor. It was a rather unsuitable place for an administrative capital, but it was a mountain, and the peoples of Southeast Asia have always believed that gods and spirits dwell on mountaintops. The image of the sacred mountain thereafter remained the inspiration for all the later architecture of the Khmer around Angkor. Jayavarman, who built other temples in the vicinity, seems to have revived the Chenla style. A distinctively Khmer art, however, began to emerge under Indravarman I (877–889), who expanded the boundaries of the Khmer kingdom and finally settled its administration. Most important of all, he developed the initial plan of the colossal city of Angkor, whose mysterious ruins, lost in dense jungle until very recently, have tantalized Western travellers for three centuries.

Angkor was not only a city; more important, it was an immense technological achievement, from which the agricultural prosperity of the whole Cambodian plain derived. This plain was well watered naturally, but its rivers were subject to strong seasonal fluctuations. Controlled,

Taken from *The Art of Indochina* by Bernard Philippe Groslier,© Holle & Co. Verlag, Baden-Baden, Germany. Used by permission of Crown Publishers, Inc.



Plan of Angkor.

chno-
gical
nificance
Angkor

they were capable of producing an enormous increase in fertility. Angkor was thus essentially an elaborate system of artificial lakes, canals, and radiating irrigation channels that watered a huge acreage of rice paddy; and it was the basis for the strength and prosperity of the Khmer Empire. Since Angkor itself was the technical source of the life-giving agricultural water controlled by the king, it was regarded by the Khmer with religious reverence. Its temples and palaces were an expression of that reverence and at the same time an essential part of its supernatural mechanism. Royal intercession by numerous ceremonies, some of which re-enacted the primal marriage of Hindu divinity and native earth spirit on the pattern of ancient folk cult, ensured the continuing gift of the waters of heaven. The king, an earthly image of his god, was the intermediary who ensured that his kingdom would continue to receive divine benevolence in the form of water in controlled quantities. Courtiers played roles at once religious and administrative for the king, who believed that after his death he would be united with his patron deity. Dedicatory statues were often set up in his chief temple to commemorate his divinization.

In order to conform with mountain mythology, the Khmer kings built themselves a series of artificial mountains on the Cambodian plain at Angkor, each crowned by shrines containing images of gods and of themselves, their family, and their ancestors. The huge platforms of earth on which these buildings were founded probably consist of the soil excavated in forming the lakes, moats, and channels that not only divided up the city but also provided an easy means of transport. The temple mountains, like the city itself, are oriented east to west, the main gates facing east. Each king strove to outdo his predecessor in the height, size, and splendour of his temple mountain. The earlier ones, therefore, are relatively small, though beautiful, while the later ones, such as Angkor Wat and the Bayon, are of stupendous size.

ie basic
ittern of
e Khmer
mple
ountain

In the basic pattern of the Khmer temple mountain the principal overall enclosure, which is square or rectangular, is at ground level. Within it the artificial mountain rises through a series of terraces and at least one further enclosure wall toward a flat summit. On the summit stands either a single shrine or a group of shrines, often a quincunx—five shrines, one at each corner and one in the middle of a square. Arranged along the terraces or within the enclosures there may be further shrines, whose arched doorway pediments refer to the rainbow bridge between heaven and earth. There may be other long buildings, perhaps used as libraries or administrative offices. A principal staircase runs directly up from the east gate to the summit, and sometimes subsidiary staircases run up from other gates at the cardinal directions.

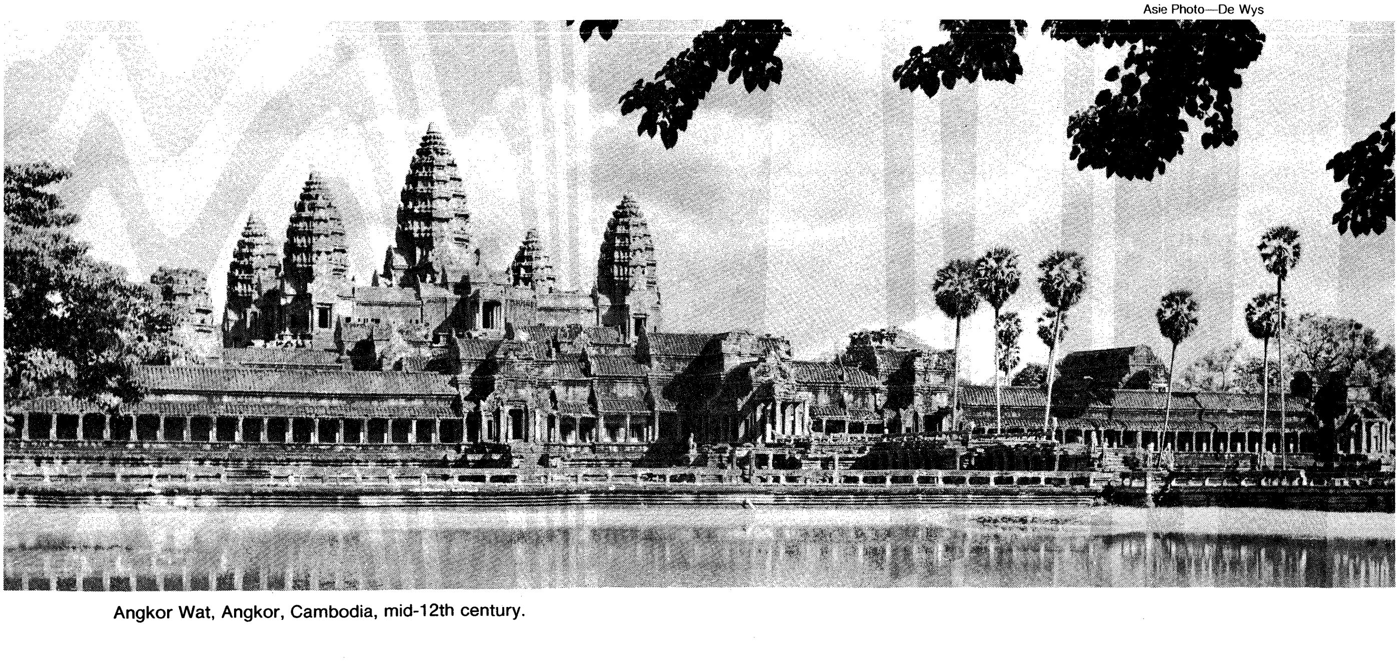
The architecture of the shrines themselves is relatively simple; it is based upon patterns invented in India, though the ornament of the shrines is often highly developed and characteristically Cambodian. Fundamentally, each shrine consists of a cell whose internal space is cubic and whose external walls are marked by moldings at top and bottom. The shrine is roofed by a pyramidal tower composed of a series of similar but diminishing tiers, each of them a compressed version of the exterior pattern of the main shrine volume. Depending on which Indian pattern is followed, the cell has one main door with an elaborately carved portal or, if the plan is cruciform, four entrances. The earlier shrines were built of brick, most commonly with stucco ornament and figures on the outside. The later shrines were built of stone, with all their ornament and figurative sculpture carved in relief. The moldings on the roofs of the shrines and the decoration of the roofs of many of the subsidiary buildings are extremely elaborate. There are long panels of dense foliate ornament, and the niches in which the sculptured relief figures of celestials are set are framed in flamboyant ogival (contoured like a pointed arch) moldings crowned by no less flamboyant foliate ornament; the smaller architectural features, such as niche pilasters, are elaborately carved and molded. The figures themselves wear gorgeous jewelry and chignons. The massive stone icons that survive in some of the shrines and on the terraces do not have the subtlety or strength of the Funan–Chenla sculptures. Instead, they have an inflated massiveness, intended, no doubt, to make them awe-inspiring. Among the lesser relief figures of celestials, which decorate the walls of the shrines, one finds a more sensuous touch; for many of these celestials represent *apsaras,* the celestial girls of Indian mythology.

Relief
panels

On some of the temple mountains there are also relief panels illustrating various aspects of the royal mythology. Episodic relief sculpture first appears on Banteay Srei (10th century). The relief revolves around a series of Indian legends dealing with the cosmic mountain Meru as the source of all creation and with the divine origin of water. The chief artistic achievement of its architecture is the way in which it conceives and coordinates the spaces between the walls of the enclosures, the faces of the terraces, and the volumes of the shrine buildings. A most sophisticated architecture of full and empty space, it seems to have been influenced by that of the Hindu Pallava dynasty in southeastern India.

The earliest more or less complete example of a shrine complex devoted to deifying the ancestors of a king is the Preah Ko at Roluos, near Angkor, completed in 879. The earliest surviving temple mountain at Angkor itself is the Bakong, probably finished in 881. In the central shrine at the summit was a *liṅga,* the phallic emblem sacred to

Asie Photo—De Wys



Angkor Wat, Angkor, Cambodia, mid-12th century.

Śiva. Around the base of the terraced pyramid stood eight large shrines inside the main enclosure, with a series of moats, causeways, and auxiliary sculptures guarding the approaches to the exterior. The Bakheng, begun in 893, had an enormous series of 108 tower shrines arranged on the terraces around the central pyramid, which was crowned by a quincunx of principal shrines. The whole was intended to illustrate a mystical conception of the cosmos, very much on the lines of the great temple mountain at Borobuḍur in Java (see below *Indonesia*). Pre Rup, dedicated in 961, was probably the first of the temple mountains intended as a permanent shrine for the divine spirit of a king after his death. It, too, has a quincunx of principal shrines, but it is distinguished by the large number of auxiliary pavilions arranged along both sides of the inner enclosure wall.

From about the same period is perhaps the most beautiful—and most beautifully preserved—of the early Khmer temples, Banteai Srei. This was actually a private foundation, built some 12 miles from Angkor by a Brahmin of royal descent. Its auxiliary buildings, all of sandstone, are adorned with a profusion of elaborate ornament and relief figure sculpture. The roof gables, in particular, are treated with antefixes of fantastic invention. Its principal icon, a huge sandstone sculpture of the god Śiva, seated with his wife Umā on his left knee, is perhaps the most impressive full-round sculpture from the whole Khmer epoch. It differs from 10th-century Khmer official sculpture, which began to take on a conventional and relatively insensitive massiveness.

The Baphuon temple mountain (1050–66) is unfortunately almost completely destroyed. It was a vast monument 480 yards (440 metres) long and 140 yards (130 metres) wide, approached by a 200-yard (180-metre) causeway raised on pillars. Its ground plan shows that it was no mere assemblage of buildings but a fully articulated structure. In this it must rank as the immediate prototype for the great Angkor Wat. Built by Suryavarman II in the early 12th century, Angkor Wat is the crowning work of Khmer architecture, the culmination of all the features of earlier styles.

Angkor Wat

The enormous structure of the Wat is some 1,700 yards (1,550 metres) long by 1,500 yards (1,400 metres) wide. Surrounded by a vast external cloister, it is approached from the west by a magnificent road, which is built on a causeway and lined with colossal balustrades carved in the likeness of the cosmic serpent, associated with the sources of life-giving water. The Wat rises in three concentric enclosures. The western gate complex itself is nearly as large as the complex of central shrines, and both are subdivided into smaller, beautifully decorated courts. Only five of the original nine towers still stand at the summit; although they follow the basic pattern of the Khmer roof tower composed of diminishing imitative stories, the contour of the towers is not rectilinear but curved, so as to suggest that the stories grow one out of another like a sprouting shoot. All the courtyards, with their molded plinths, staircases, porticoes, and eaves moldings, are perfectly articulated enclosed spaces. The symbolic meaning of the Wat is clear. Its central shrine indicates the hub of the universe, while its surroundings—the gate complex, the cloister, the city of Angkor itself, and, finally, the whole visible world—represent the successive outer envelopes of cosmic reality. That it is oriented toward the west—and not to the east, as was customary—indicates that its builder, Suryavarman II, intended it as his own mortuary shrine; for, according to Indochinese mythology, the west is the direction in which the dead depart.

The sculptors who worked at the Wat demonstrate little ability in carving in the full round. Such full-round figures as there are—the guardians on the terraces, for example—lack life. The relief sculpture, however, is magnificent and full of vitality. The open colonnaded gallery on the first story contains over a mile of relief carving six feet (two metres) high. Much of it was originally painted and gilded, which strongly suggests that there must have been a Khmer style of painting of which nothing is known. The subject matter of the carvings is taken principally from the Hindu epics, but there are also many scenes representing Suryavarman's earthly glory. Working in relief only about an inch deep, the sculptors were able to depict an extraordinary complex of scenes of figures in vigorous action, full of complex overlaps to suggest deep space. The solid bodies are created mainly out of groups of convex curves; and everywhere there is the typical regional feeling for decorative spirals. Perhaps the most interesting group of figures are the *apsaras*, carved in relief, either singly or in groups, on the plain walls of the courtyards. These celestial beauties, whom Indian tradition describes as rewarding with their charms the kings, heroes, and saints who attain heaven, are carved with sinuous sensuality; but the most important part of their charm is their elaborate clothing, jewelry, and hairdressing or ornate, towering, jewelled crowns. Apparently, deep, downward-drooping curves standing far out from the body represented the height of Khmer chic. Skirts, stoles, and the long sidelocks of the hair all follow these curves, laid out flat on the ground of the relief. Symbolizing the erotic joys that are essential attributes of heaven, the *apsaras* were natural possessions of the king.

Bas-relief of a battle scene, Angkor Wat, Angkor, Cambodia, early 12th century.

Holle Bildarchiv, Baden-Baden

The *apsaras*

In many senses the Wat was the end of the road for Khmer art. The effort demanded of the people in constructing this colossal stone monument, along with its four miles (six kilometres) of stone-lined moat 200 yards (180 metres) wide, appears to have been too great. The irrigation system itself may well have been neglected in favour not only of shifting the building stone—as much in quantity as there is in the Pyramid of Khafre in Egypt—but also of dressing, carving, and ornamenting it. After Suryavarman's death, the Cham, from the neighbouring kingdom of Champa (see below *Vietnam kingdom of Champa*), seized and sacked Angkor for the first time in its history (1177), thus shattering the confidence of the Khmer people in the protective powers of their Hindu deities. When Suryavarman's son, Jayavarman VII, came to the throne he inherited a ravaged kingdom. In 1181 he succeeded in driving out the Cham; he invaded their country and seized their capital, thereby making Champa a province of the Khmer. Then, over 60 years old, he embarked on a series of campaigns that extended the borders of the Khmer Empire further than ever before—into Malaya, Burma, and Annam.

The Bayon

The ruler of this empire naturally believed himself to be the greatest of the Khmers, and he set about demonstrating the truth of his belief by building his own city, Angkor Thom (*c.* 1200), and, at the centre of it, the biggest temple complex of them all—the Bayon (*c.* 1200). Breaking with all previous Khmer traditions, he took as his patron deity not one of the Hindu gods but one of the Buddhist *bodhisattvas*. Although Buddhism had flourished for several centuries in the whole of Indochina, it had not been adopted by the Khmer as an imperial cult. Now that the Hindu gods had been discredited by defeat, Jayavarman placed himself under the patronage of Mahāyāna Buddhism. The mythology according to which the Bayon was designed was thus another version of the old mythology of the celestial mountain and the divine origin of water. Only the central figure of his mythology, Lokeśvara, Lord of the World, was specifically Buddhist. The colossal masks that look out over the four directions of the world from the towers of the Bayon and from the gates of Angkor Thom are there to demonstrate the compassionate, all-seeing power of Lokeśvara and the king.

When Jayavarman VII set out to create Angkor Thom, he had to raze the fine older work of his predecessors, for the site at Angkor had become choked with nearly four centuries of grandiose temple building. Within Angkor Thom's ten miles (16 kilometres) of moats he constructed huge complexes of building and made his city the focus of a final system of canals and irrigation, with additional lakes.

Unfortunately, the innumerable new shrines that surround Angkor Thom, the towers that crowd the Bayon, and the vast stone terraces faced with relief that surround the royal palace are in general much inferior in execution to the work of the earlier kings. Thus, today Angkor is dominated by the overwhelming presence of Jayavarman's immense but relatively unrefined architecture. The King's ambition was satisfied by size and quantity rather than artistic quality. Because sculptors were obliged to produce such vast quantities of work so fast, their standards deteriorated, and the huge vistas of narrative relief show signs of haste and slipshod workmanship. The real achievement lay with Jayavarman's scholastic architects, who conceived and laid out a complex of mythical imagery in massive architectural symbols. Their stupendous overall plan illustrates the creation of the world, a cosmos spreading outward from the central mountain tower. The two roads leading from the tower are lined with mile-long rows of gigantic deities who are pulling on the body of the serpent *nāga*. According to Hindu legend, the gods use the magical mountain Meru, symbolized by the mountain tower, as a churning stick and the body of the cosmic serpent as a churning rope to churn the world out of the milk of nothingness. Lake-sized fountains represent the healing waters of the Buddhist paradise, and allegories of salvation are realized in carved architecture. Perhaps the most impressive works of art associated with this last period of Angkor are some stone icons, such as the famous "Leper King," in the Angkor Thom complex. Many excellent smaller bronze figures of deities have also been found among the ruins.

Decline of Angkor

13th century to the present. After the death of Jayavarman VII, *c.* 1215, possibly as late as 1219, Angkor declined. The Thai population of Siam gradually pushed the Khmer down toward the Mekong Delta. Theravāda Buddhism became the religion of the people, and the grandiose vision of a cultural unity based on sacred kingship disappeared. In the 15th century, Angkor was retaken from the Thai, and a few buildings were restored by the ancestors of the modern (now abdicated) Cambodian kings. Some of the buildings were used as monasteries, but the city, with its essential irrigation system, had fallen into ruin.

Vietnam kingdom of Champa: c. 2nd to 15th century. The kingdom of Champa existed alongside the Khmer kingdom, sometimes passing under its rule, sometimes maintaining a precarious independence. From the north it was continually subject to the pressure of the advancing Vietnamese, a people racially related to the Burmese and Thai, who were themselves under pressure from the Chinese. The Hinduizing dynasties who ruled Champa from the 6th century were obliged to pay heavy tribute to the Chinese Empire. After 980 they were forced by the Vietnamese to abandon their northern sacred capital, My Son; thereafter, except for a brief return to My Son in the 11th century, their southern capital at Vijaya (Binh Dinh) became their centre. Under such disruptive circumstances, it is perhaps surprising that the Cham succeeded in creating and maintaining a dynastic art of their own. It was, however, always on a relatively modest scale, devoted to a conception of divine kingship similar to but far less ambitious than the Khmer.

The evolution of Cham art falls naturally into two epochs, the first when the capital was in the north, the second when it was removed to the south.

Art of the northern capital: 4th to 11th century. The form of the earliest temple at My Son, built by King Bhadravarman in the late 4th century, is not known. The earliest surviving fragments of art come from the second half of the 7th century, when the king was a descendant of the royal house at Chenla. The remains of the many dynastic temples built in My Son up until 980 follow a common pattern with only minor variations. It is a relatively simple one, with no attempt at the elaborate architecture of space evolved by the Khmer. Each tower shrine is based upon the central rectangular volume of the cell. The faces are marked by central porticoes that are blind on all but the western face, where the entrance door is situated. The blind porticoes seem to have contained figures of deities—perhaps armed guardians standing in a threatening posture. The porticoes are set in a tall, narrow frame of pilasters (columns projecting a third of their width or less from the wall), crowned with horizontally molded capitals that step out upward. They support a tall, double-ogival blind arch, crowned by another stepped in behind it. The arches are based on an Indian pattern and are carved with a design of slowly undulating foliage springing from the mouth of a monster whose head forms the apex of the arch. The faces of the walls are formed of pilasters framing tall recesses. The pilasters are carved with foliate relief, and elaborate recessed and stepped-out horizontal moldings mark their bases. The height of the pilasters and recesses gives a strong vertical accent to the body of the shrine. The principal architrave is carried on stepped-out false capitals to the pilasters. The roof of the tower is composed of three diminishing, compressed stories, each marked by little pavilions on the faces above the main porticoes. Inside the tower is a high space created by a simple corbel vault with its stepped courses of masonry. The chief portico was extended to include a porch, and the whole structure stood upon a plinth whose faces bore molded dwarfed columns (small columns) and recesses.

Pedestal altars

These temples have one distinguishing internal feature: a pedestal altar within the cell, upon which statues were set, sometimes, it seems, in groups. The pedestals themselves are often beautifully adorned with reliefs, and some of the best Cham sculpture appears upon them. The subjects are usually based on Indian imagery of the celestial court. The fact that the pedestal altars carried their sculptures in the space of the cell, away from the wall, meant that the Cham sculptors could think in terms of three-dimensional plasticity as well as relief.

The glory of Cham art is the sculpture of the whole of the first period. Much of what survives consists of lesser figures that formed part of an architectural decor: heads of monsters, for example, which decorated the corners of architraves, and figures of lions, which supported bases and plinths. These figures reflect the heavy ornateness of the Cham decorative style at its most aggressive; and many of them effloresce into the solid, wormlike ornament that is the Cham version of Indo-Khmer foliage carving and carries strong reminiscences of Dong Son work. The remaining fragments of the large icons suggest a double origin for Cham art traditions. On many of the capitals and altar pedestals are series of figures carved in relief in a sensuous style, which is nevertheless strictly conceptualized. This sophisticated work is reminiscent both of late Chenla art (see above *Cambodia and Vietnam*) and of Indonesian decoration, especially during the 11th-century return. Other figures are more coarsely emphatic in style, with the crudely defined cubic volumes and the heavy faces of Melanesian sculpture. It is thus probable that artists trained in the sophisticated Cambodian tradition worked for the Cham kings at one time or another, while Champa's own native craftsmen emulated the work of the foreigners in their own fashion.

Apart from My Son there are one or two other sites in north and central Vietnam where Cham art was made in quantity. The most important of these is Dong Duong, in

Quang Nam. It is a ruined Buddhist monastery complex of the late 9th century, conceived on the most beautifully elaborated plan of structured space in Champa. The architectural detail is distinguished from the My Son work by its greater emphasis upon the plasticity of architectural elements such as angle pilasters and porticoes. The circuit wall was about half a mile (one kilometre) long and once contained many shrines dedicated to Buddhist deities. It is possible that, when this complex of brick courts, halls, and gate pavilions was intact, it may have resembled very closely the contemporary Buddhist monasteries of northeastern India.

Art of the southern capital: 11th to 15th century. After 980, when the northern provinces were taken over by the Vietnamese and the Cham capital established at Binh Dinh in 1069, the kings maintained a gradually diminishing splendour. After the Khmer attack of 1145 they could claim little in the way of royal glory.

Decline of shrine architecture

Although the Cham kings made a brief return to My Son from 1074 to 1080, most of their artistic effort was spent on shrines at Vijaya (Binh Dinh) and a few other sites in the south. The early 12th-century Silver Towers at Binh Dinh are simplified versions of the older northern towers, with corner pavilions added to the roofing stories and arches of pointed horseshoe shape. Throughout the 13th and early 14th centuries the architecture of successive shrines gradually declined. The plasticity of the old pilasters and architraves was suppressed into simple moldings, and the beauty of the buildings became largely a matter of proportion. By the mid-14th century even the temples erected at Binh Dinh amounted to little more than piles of crudely cut stones articulated only by reminiscences of the classic Cham style.

Sculpture shows a parallel decline. One or two reliefs at the Silver Towers do convey a sense of tranquillity and splendour, but an indigenous style of rigid cubical emphasis came progressively to dominate the iconic Hindu figures at southern sites. The curlicued design of earlier figures was gradually converted into a style of massive, scarcely carved blocks that convey, at their best, an impression of barbarous strength but without the refinement of first-class primitive art.

As was the case in Cambodia, this decline in art by the mid-14th century may be attributed to the people's loss of confidence in the concept—and, with it, the imagery—of divine kingship. Theravāda Buddhism, as a popular religion based upon numerous small, local monasteries, adopted probably from the Tai, was spreading all over the region. The northern Vietnamese, who had originally been organized in self-contained kingdoms without any concept of royal divinity, owing an intermittent administrative allegiance only to the distant Chinese emperor, found this ultimately suitable as a state religion after the final eclipse of Confucianism in the 17th century. They did incorporate echoes of older Hindu architecture, however, in details of the flamboyant ornament used on eaves and gables of their wooden monastery buildings.

Vietnam: 2nd–19th century. The great achievement of Vietnamese art, at least during the Le period (15th–18th centuries), seems to have been in architectural planning, incorporating Confucian, Taoist, or Buddhist temples into the landscape environment. The plans themselves include halls for a multitude of images in South Chinese vein and provision for a variety of rituals. There are no intact monuments of early Vietnamese architecture that are unrestored. Numerous fragments exist, however—either isolated stone bases, columns, stairways, and bridges or carved wooden members incorporated into later buildings—all of which are influenced to some degree by Chinese styles.

Tombs of the Chinese type

Tombs of generically Chinese type from the 2nd to 7th century contain bronze furnishings, in many of which, such as lampstands, the influence of the Dong Son style is clearly visible. There are no spirit images so typical of Six Dynasties (3rd–6th centuries) and T'ang (7th–10th centuries) Chinese tombs. The Chua Mot-cot, Hanoi, has vestiges of a stone shrine probably dated 1049. The only old paintings, on rock, at Tuyen Quang, (9th century) represent the Buddha, *bodhisattvas*, and donors. The Van-mieu at Hanoi (built 1070 but frequently restored) contains ritual bronzes in "barbaric" Chinese style.

Perhaps the most interesting early sculptures to survive are the stone fragments from the Van-phuc temple (9th–11th centuries), which are based on Chinese Buddhist imagery but in a style strongly Indianized, perhaps by Cham influence. The most important piece of old work still virtually intact is the portable octagonal wooden stupa kept in the hall of the But-thap, at Bac Ninh, east of Hanoi. It has wooden panels carved in a flamboyant 14th-century Chinese style; part of it bears a representation of the Buddhist paradise of Amitābha. Incorporated in many Buddhist temples of the Le period (15th–18th centuries), as well as in stone terraces, bridges, and gateways, is extremely elaborate carved and coloured woodwork in a style based upon the coiling dragon-and-cloud decoration of Ming (1368–1644) and Ch'ing (1644–1911) China, but with a characteristically Vietnamese exaggeration of weight and curve.

At Tho Ha there was a potters' village, where the glazed ceramic figures used on many types of Chinse temple were manufactured. The remains of many tombs, palaces, bridges, and Confucian and Taoist temples decorated in similar vein are known everywhere.

19th and 20th centuries. The beautiful imperial palace of Hue (final plan before 1810) contained vestiges of older architecture and many works of Sinicized art before its devastation in 1968. It consisted of a series of simple, rectangular, one-story pavilions, laid out among trees inside a group of courts. These buildings and their decoration were southern Chinese in basic conception.

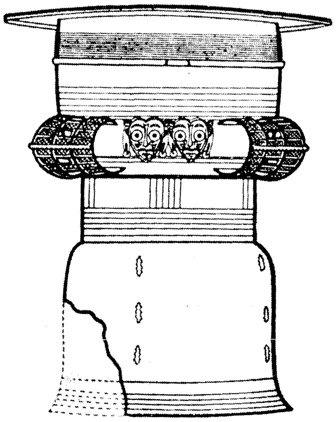
Elsewhere in Vietnam, both religious and secular buildings have been constructed in the 20th century in provincial versions of Chinese styles. There was little demand for the sculptor's art beyond the carving of stereotyped Buddha icons, monsters, and guardians. In modern times southern Vietnam has adopted a decorative style partly derived from the active traditions of Bangkok. Religious sects abound, with hybrid native European and Chinese elements used in their iconic and decorative art. Because of political turmoil, no clear and individual modern Vietnamese artistic tradition has been able to emerge.

INDONESIA

Megalithic monuments

The islands that at the present day compose Indonesia probably once shared in the complex Neolithic heritage of artistic tradition, which also spread further, into the islands of Melanesia and Micronesia. Beautifully ground Neolithic axes of semiprecious stone are treasured still in some countries. In many parts of Indonesia there are quantities of megalithic monuments—menhirs, dolmens, terraced burial mounds, stone skull troughs, and other objects. Some of these are undoubtedly of Neolithic date, but megaliths continued to be made in much more recent times. One stone sarcophagus in eastern Java, for example, is dated post-9th century. On Nias island, megaliths are still revered, and they are still being erected on Sumba and Flores islands. Thus, in Indonesia especially, different layers of Southeast Asian culture have existed side by side. The most impressive and important collection of megaliths is in the Pasemah region, in south Sumatra, where there are also many large stones roughly carved into the shape of animals, such as the buffalo and elephant, and human figures—some with swords, helmets, and ornaments and some apparently carrying drums.

These drums immediately suggest the drums characteristic of the mainland Southeast Asian Dong Son culture, which flourished *c.* 4th–1st centuries BC (see above *General development of Southeast Asian art*). This culture may well have helped to diffuse throughout the region styles related to Chinese Chou and pre-Han ornamental work. Certainly, the Dong Son influence is clear in many of the ceremonial axes, as well as many of the ornamented bronze drums, that have been found in the islands. The bronzes were cast by a cire perdue process resembling that used in parts of the Asian mainland. The largest and most famous drum is "the Moon of Bali," found on that island near Pedjeng. It has molded flanges, and cast onto its faces is extremely elaborate relief ornament consisting of stylized masks with ears pierced and lengthened by large earrings. Such drums were probably originally used in ritual—by the rainmaker, perhaps—and they may have been buried with the distinguished dead. No one knows the exact age of these bronzes; "the Moon of Bali," for example, could be anywhere between 1,000 and 2,000 years old. Similar small drums were used quite recently as bride prices; and many of the islands today produce textile designs and ceremonial bronzes that are strikingly reminiscent of Dong Son ornament.



"The Moon of Bali," bronze drum found near Pedjeng, Bali. In the Panataram Sasih temple, Pedjeng, Bali. Height 1.87 m, diameter 1.60 m.

From P. Rawson, *The Art of Southeast Asia*, Thames and Hudson Ltd., London

Central Javanese period: 7th to 13th century. Sometime between the 3rd and 6th centuries AD, Indianized principalities existed in Java. The chieftains who lived in their *kratons* (fortified villages) seem to have derived great inspiration, prestige, and practical assistance from the skills and ideas imported from India. In Sumatra there was the important but so far enigmatic Indianized kingdom of Śrivijaya, which, from its strategic position on the Strait of Malacca, exercised a powerful artistic influence in the whole region. Its great Buddhist centre, Palembang, might have had direct connections with the monasteries of southeastern India; for fine bronze Buddhas and *bodhisattvas* in a style reminiscent of Amarāvatī (2nd century AD) have been found in many regions where the influence of Śrivijaya might have been felt, including Mon Dvaravati (see above *Thailand and Laos*) and distant Celebes. Elsewhere among the islands were Indianized kingdoms still unknown to history.

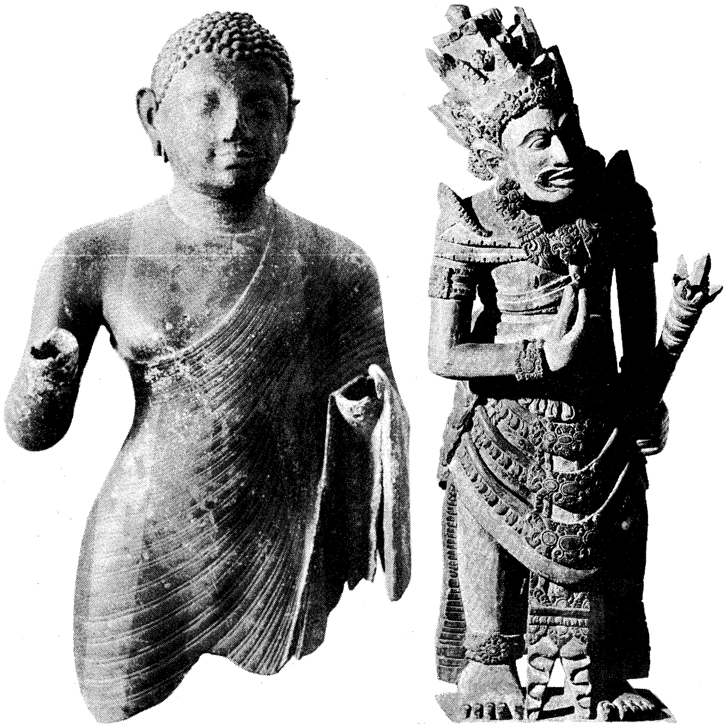
The
Indianized
art of
Śrivijaya

The local dynasties of the *kratons* competed among themselves for power, and eventually the principal dynasties known to history came to the fore. The earliest major cultural assimilations from India took place probably during the 7th century, when the Hindu Pallava form of southeast Indian script was adopted for inscriptions in west Java. Thereafter, a central Javanese dynasty that worshipped Śiva made the oldest surviving artworks in stone. The last king of this dynasty retreated to east Java in the face of the rising power of another central Javanese dynasty, the Śailendra (AD 775–864.) The Śailendra were followers of Mahāyāna and Tantric forms of Buddhism, although Hinduism, as manifested in the worship of Śiva and Vishnu, was by no means eliminated. This dynasty created far the larger part of the immense wealth of first-class art known today in Java.

In Indonesia, the word *tjandi* refers to any religious structure based on an Indianized shrine with a pyramidal tower. This was the essential form on which virtually all the stone Indianizing architecture of Southeast Asia was originally based. The Javanese, like the Khmer, evolved an elaborate architecture of their own around the basic Indian prototype.

Central Javanese stone architecture did not use structural pillars, nor did its major stone monuments conceptualize hollow space in the way Khmer architecture did. Like Indian stonework, central Javanese stonework is fundamentally conceived as a solid mass, serving as a vehicle for figurative and symbolic sculpture. Its temples are centralized, with enclosures radiating around the central shrine. In eastern Java and Bali, however, the pattern of the shrine was influenced by older traditions and was usually conceived as an enclosure, the walled area of ground be-

By courtesy of (centre, right) the Royal Tropical Institute, Amsterdam; photograph, (left) Louis Frederic—Rapho/ Photo Researchers



Indonesian sculpture.
(Left) "Buddha" from Celebes, bronze, Amarāvatī style, 3rd–5th century. In the Jakarta Museum. Height 75 cm. (Centre) "Vishnu" from Bali, stone. In the Royal Tropical Institute Museum, Amsterdam. Height 80 cm. (Right) Ancestor figure from the Tanimbar Islands, Indonesia. In the Royal Tropical Institute Museum, Amsterdam. Height 38 cm.

ing the sacred element, while the buildings in it were of secondary importance. Old wooden buildings do not survive; but representations of wooden architecture in stone reliefs and the recent architecture of Bali show that eastern Indonesia was influenced by the ancient Southeast Asian tradition of constructing wooden pillared halls with tiered, sloping, and gabled roofs.

Hindu *tjandi*s

Because there are no inscriptions to supply dating points, the exact dates of the earliest Indonesian architectural monuments are not certain. The group of shrines generally believed to be the earliest is situated on the Dijeng Plateau. This is a high volcanic region, about 6,000 feet (2,000 metres) above sea level, where there are sulfur springs and lakes. The whole mountain seems to have been sacred to the Hindu deity Śiva, for all temples on the Dijeng are dedicated to him. There can be little doubt that during the 8th and 9th centuries the Javanese, who traditionally had interpreted the volcanic turbulence of their landscape as a manifestation of divine power, identified this power with the terrifying Śiva. On other Javanese volcanic mountains, also, groups of shrines are dedicated to him.

The temples on the Dijeng are single-cell shrines, roofed with diminishing stories. The exteriors of the temples are relatively plain; only around door frames and window frames are there distinctive passages of central Javanese ornament. Around the niches of Tjandi Puntadewa are perhaps the earliest surviving examples of the characteristic Javanese doorframe: across its lintel is carved a mask of the Indian Kāla monster, which represents time; and down the jambs, as if vomited from his open mouth, run string panels of foliage. The foot of each jamb terminates in an elaborately carved scrollwork cartouche, which is itself a *makara* (water monster) head seen in profile. This *tjandi,* like others on the Dijeng, has a single approach stairway rising between curved balusters. A few stone images of Śiva from these temples have been found. In broad, vigorous forms they express the dangerous power of the god.

Two of the very finest early Javanese sculptures—virtually in the full round—come from yet another Śiva temple, Chandi Banon, near Borobuḍur (see below). One, representing the god Vishnu (no stranger in syncretic Javanese temples of Śiva), has the extremely smooth, faintly amorphous suavity, the absolute convexity, and the lack of definition between planes characteristic of the classical central Javanese sculptural style; while the garment he wears, with its assortment of girdles, is closely reminiscent of late Pallava–early Cōḻa Hindu styles of southeast India. Another icon, sometimes called Agastya but more likely the third deity of the Hindu trinity, Brahmā, represents the god in the form of a bearded Brahmin sage. He has a large and, to Oriental eyes, splendid potbelly. This icon was indigenous to southeast India. The great depth of the side recessions of these figures, although perhaps not so clearly defined as in the great Funan–Chenla style (see above *Cambodia and Vietnam*), gives them a bland massiveness. The lack of movement in the figures and the regularity of the designs, the impassive faces, and the slowness of the lines must have been part of the central Javanese conception of transcendent glory.

Buddhist *tjandi*s

The Hindu temples of central Java are conceived simply as shrines to contain icons of deities for worship. The Mahāyāna and especially the Tantric Buddhist *tjandi*s, however, were called upon to do far more. They were designed to express complex metaphysical theories. The challenge this presented to the central Javanese architects was met in a series of splendid monuments, completely original in conception. The culminating work of the series, Borobuḍur, is a highly evolved architectural image, whose subtlety and refinement were never matched, even at Angkor in Cambodia.

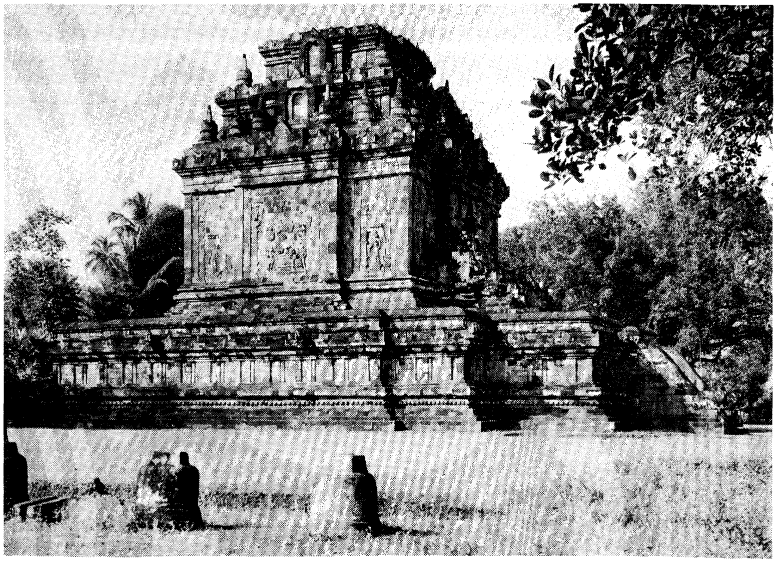
The first work of this Buddhist series is Tjandi Ngawen, near Muntilan. This *tjandi* consists of five shrines facing east, 12 feet (four metres) apart in a row from north to south. Each shrine contained one of the five Buddhas who, according to Tantric Buddhist theory, presides over one of the five major psychological categories under which ultimate reality reveals itself. The shrines themselves are based on but more developed than those used for Hindu deities elsewhere in Java. Roughly square in plan and roofed with diminishing stories, they have pilastered projections on three faces and a portico on the east. Along the architrave are small triangular antefixes and reliefs of Kāla monsters vomiting floral scrolls hood the niches and portals.

The group of five Buddhas is familiar in the art of Tibet, Japan, and northeast India. Among them they compose what is called the *vajra-dhātu,* which means, roughly speaking, "the realm of total reality." According to the old Javanese theology, above this group is another, called the deities of the *garbha-dhātu. Garbha* means "womb" or "innermost secret," and its three deities personify the most esoteric realms of Buddhist speculation. At the centre of the group is the image of the single, undivided Buddha nature, which symbolizes the ultimate reality of the entire universe. From his right side emanates the *bodhisattva* Lokeśvara (Lord of the World), who is both compassionate and possessed of all power. From the left emanates the *bodhisattva* Vajrapāṇi, who is the personification of the most secret doctrines and practices of Tantric Buddhism. One of Java's greatest monuments, Tjandi Mendut, is a shrine expressly created to illustrate the combined doctrine of *garbha-dhātu* and *vajra-dhātu.*

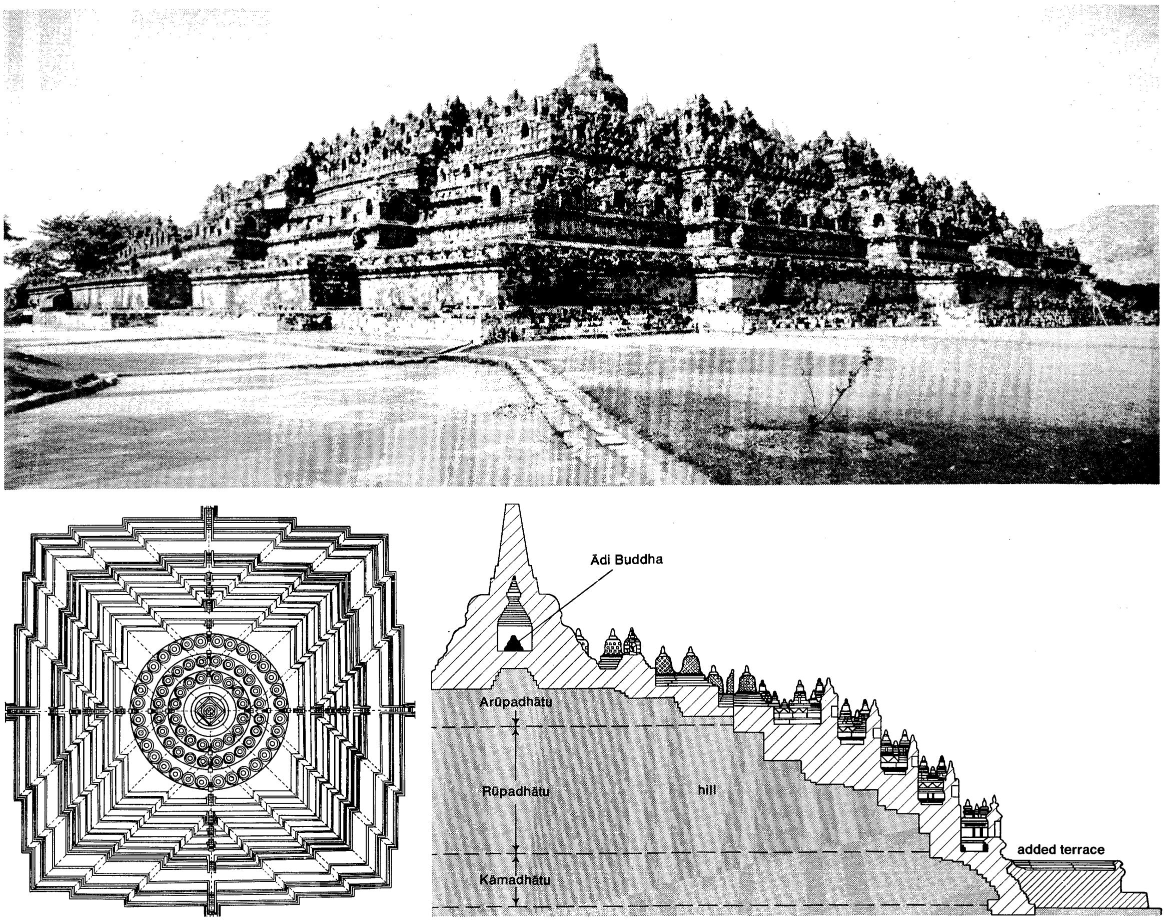
Tjandi Mendut

Mendut dates from about 800 and is thus, generally speaking, contemporary with Borobuḍur. It is formed as a single large, square chamber, roofed with the usual diminishing stories, and mounted on a high, broad plinth,

Louis Frederic—Rapho/Photo Researchers



Tjandi Mendut, near Borobuḍur, Java, *c.* 800.



(Top) Borobuḍur, Java, *c.* 800. (Bottom left) Plan of Borobuḍur and (bottom right) section of Borobuḍur.

(Top) Holle Bildarchiv, Baden-Baden, (bottom left, bottom right) from P. Rawson, *The Art of Southeast Asia,* Thames and Hudson Ltd., London

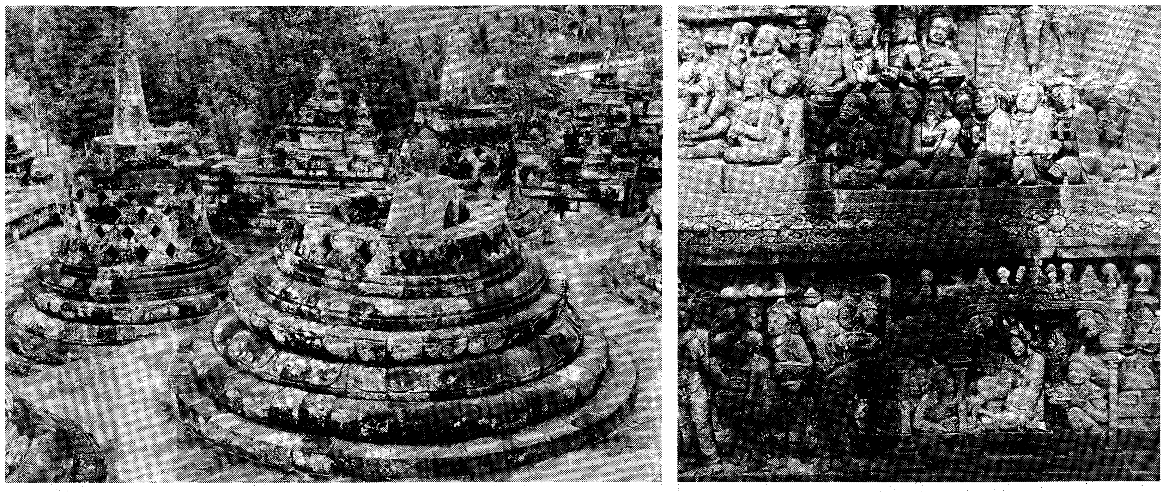
which is approached on its northwestern face by a staircase with recurved balustrades. The exterior is in every way more ornate than that of any shrine so far discussed. In addition to floral diaper (an allover pattern consisting of one or more small repeated units of design connecting with or growing out of one another) and scrolls, there are numerous figures in relief representing male and female deities, the subsidiary principles of the combined doctrine of *garbha-dhātu* and Vajrapāṇi. Cut into the fine ashlar (squared-stone) masonry are many relief panels with scenes from Buddhist literature, each panel self-contained and placed with consummate aesthetic judgment. Some represent mythical ideas, such as the wish-granting tree, others narratives from Buddhist legend.

The principal images were placed inside the cell chamber. Apparently, there were originally seven huge stone icons, but only three remain: the central Buddha, who also represented the ultimate Buddha nature of the *garbha-dhātu,* and his two emanations in the *garbha-dhātu,* Lokeśvara, and Vajrapāṇi. When complete, the interior of Mendut must have been an even more awe-inspiring and spiritually moving place than it is now. The three great statues are seated on elaborate thrones, backed against walls, but the figures are carved virtually in the full round. The inflated, gently inflected forms of the figures give them a majestic presence. The types and carving technique, as well as the monumental scale of the figures, are reminiscent of contemporary work in the cave temples of the western Deccan in India.

On the west–east road from Tjandi Mendut to Borobuḍur stands a small, relatively plain temple called Tjandi Pawon, dedicated to the god of wealth. Pawon was probably a kind of anteroom to Borobuḍur, catering to the more worldly interest of pilgrims. The outside has fine reliefs of female figures, and the roof bears towers of small stupas. On the reliefs are wish-granting trees surrounded by pots of money, and bearded dwarfs over the entrance pour out jewels from sacks.

Borobuḍur

Borobuḍur is one of the most impressive monuments ever created by man. It is both a temple and a complete exposition of doctrine, designed as a whole, and completed as it was designed, with only one major afterthought. It seems to have provided a pattern for Hindu temple mountains at Angkor (see above *Cambodia and Vietnam*), and in its own day it must have been one of the wonders of the Asian world. Built about 800, it probably fell into neglect by about 1000 and was overgrown. It was excavated and restored by the Dutch between 1907 and 1911. It now appears as a large, square plinth (the processional path) upon which stand five terraces gradually diminishing in size. The plans of the squares are stepped out twice to a central projection. Above the fifth terrace stands a series of three diminishing circular terraces carrying small stupas, crowned at the centre of the summit by a large, circular, bell-shaped stupa. Running up the centre of each face is a long staircase; all four are given equal importance. There are no internal cell shrines, and the terraces are solid; Borobuḍur is thus a Buddhist stupa in the Indian sense. Each of the square terraces is enclosed in a high wall with pavilions and niches along the whole perimeter, which prevents the visitor on one level from seeing into any of the other levels. All of these terraces are lined with relief sculptures, and the niches contain Buddha figures. The top three circular terraces are open and unwalled, and the 72 lesser, bell-shaped stupas they support are of open stone latticework; inside each was a huge stone Buddha figure. The convex contour of the whole monument is steepest near the ground, flattening as it reaches the summit. The bottom plinth, the processional path, was the major afterthought. It consists of a massive heap of stone pressed up against the original bottom story of the designed structure, so that it obscures an entire series of reliefs—a few of which have been uncovered in modern



Sculpture at Borobuḍur.
(Left) Stupas on one of the circular terraces. The figure in the stupa in the foreground is a Buddha. (Right) Bas-relief on the exterior wall of one of the circular terraces.
By courtesy of (right) the Indonesian Tourist Board; photograph, (left) Josephine Powell, Rome

times. It was probably added to hold together the bottom story, which began to spread under the pressure of the immense weight of earth and stone accumulated above.

Symbolism of Borobuḍur

The whole building symbolizes a Buddhist transition from the lowest manifestations of reality at the base, through a series of regions representing psychological states, toward the ultimate condition of spiritual enlightenment at the summit. The unity of the monument effectively proclaims the unity of the cosmos permeated by the light of truth. The visitor was meant to be transformed as he climbed through the levels of Borobuḍur, encountering illustrations of progressively more profound doctrines the nearer he came to the summit. The topmost terrace, whose main stupa contained an unfinished image of Buddha that was hidden from the spectator's view, symbolized the indefinable ultimate spiritual state. The 72 openwork stupas on the circular terraces, with their barely visible internal Buddhas, symbolize incomplete states of enlightenment on the borders of manifestation. The usual way for a pilgrim to pay reverence to a Buddhist stupa is to walk around it, keeping it on his right hand. The vast series of reliefs about three feet (one metre) high on the exterior walls of the terraces would thus be read by the visitor in series from right to left. Between the reliefs are decorative scroll panels, and a hundred monster-head waterspouts carry off the tropical rainwater. The gates on the stairways between terraces are of the standard Indonesian type, with the face of the Kāla monster at the apex, vomiting his scrolls.

The reliefs of the lowest level illustrate scenes that show the causal workings of good and bad deeds through successive reincarnations. They show, for example, how those who hunt, kill, and cook living creatures such as tortoises and fish are themselves cooked in hells or die as children in their next life. They show how foolish people waste their time at entertainments. From these scenes of everyday life, one moves to the terraces above, where the subject matter becomes more profound and metaphysical. It illustrates important Mahāyāna texts dealing with the self-discovery and education of the *bodhisattva,* conceived as being possessed by compassion for and devoted wholly to the salvation of all creatures. The reliefs on the uppermost terraces gradually become more static. The sensuous roundness of the forms of the figures is not abated; but, in the design, great emphasis is laid upon horizontals and verticals and upon static, formal enclosures of repeated figures and gestures. At the summit all movement disappears, and the design is entirely subordinated to the circle enclosing the stupa.

The iconography of Borobuḍur suggests that the legend of the royal *bodhisattva* recounted in many of the reliefs was meant to "authenticate" some king or dynasty. Yet it hardly seems possible that Borobuḍur was the focus of a specific royal cult, as there is no provision at all for the performance of royal ritual. It must have been, then, in some sense a monument for the whole people, the focus for their religion and life, and a perpetual reminder of the doctrines of their religion.

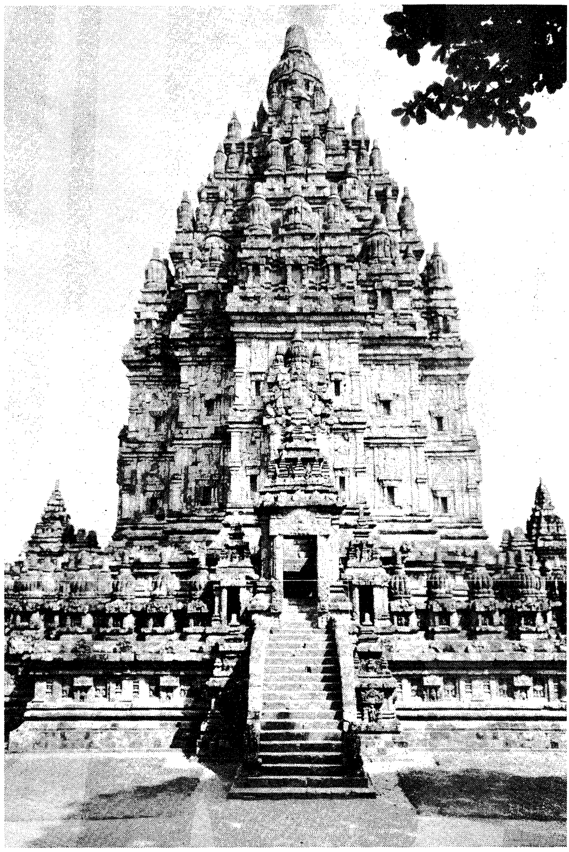
A considerable number of bronzes, some small, some large, have been found in Indonesia in a style close to that of the sculptures of Borobuḍur and Mendut. One fine, large standing image comes from Kotabangun in Borneo; but some come from Java. Many small cult images of the Buddha and Buddhist deities exist. Some are close in type to the early Pāla images of Indian Bihār, the homeland of Buddhism, with which the Javanese must have maintained close touch. A few small but extremely fine gold figurines of undoubted Javanese workmanship have also turned up. For all their small size they must rate as first-class works of art. As well as images there are many beautiful bronze ceremonial objects, such as lamps, trays, and bells. These objects are decorated with the same kinds of ornament, although on a miniature scale, as the architectural monuments: scrolled leaves, swags, and bands of jewels.

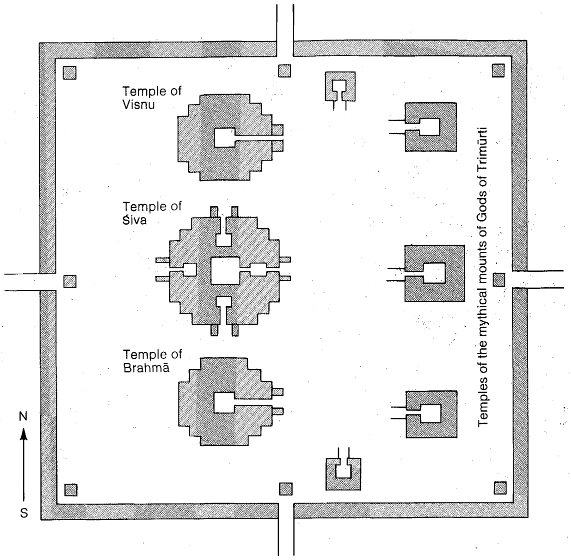
Post-Borobuḍur *tjandi*s illustrate the Buddhist doctrine in different ways. Kalasan, for example, built in the second half of the 8th century, was a large, square shrine on a plinth, with projecting porticoes at the centre of each face. The roof was surmounted by a high circular stupa mounted on an octagonal drum, the faces of which bear reliefs of divinities. Topping each portico was a group of five small stupas, and another large stupa stood at each disengaged corner of the main shrine. The moldings were restrained and elegantly profiled. Each section of the exterior wall contains a niche meant for a figure sculpture. The decorative scroll carving is especially fine.

Another shrine from this period, Tjandi Sewu, consisted of a large cruciform shrine surrounded by smaller temples, only one of which has been restored. All of the temples seem to have had roofs in the form of tiered stupas, compressing the overall Borobuḍur scheme into the scope of a storied shrine tower. From Tjandi Plaosan came many beautiful sculptures, donor figures, and iconic images of *bodhisattva*s.

Tjandi Sa

Perhaps the most interesting of the post-Borobuḍur Buddhist shrines of the 9th century is Tjandi Sari. It is an outstanding architectural invention. From the outside it appears as a large, rectangular, three-storied block, with the main entrance piercing the centre of one of the longer sides. The third story stands above a substantial architrave with horizontal moldings and antefixes. Two windows on each short side, three on each long, open into each story, though at the rear they are blind. The windows are crowned by large antefix-like cartouches of ornamental carving based on curvilinear pavilions hung with strings of gems. The uppermost windows are hooded with the Kāla-monster motif. The roof bears rows of small stupas, and perhaps there was once a large central stupa. Inside, Tjandi Sari contains a processional corridor around three





(Left) Temple of Śiva, the central temple of the Lara Jonggrang complex, Prambanan, Java, *c.* 900. (Right) Plan of the Lara Jonggrang complex.

(Left) Holle Bildarchiv, Baden-Baden, (right) from F. Wagner, *Indonesia,* Holle and Co. Verlag, Baden-Baden

interior shrines that were possibly intended for images of the *garbha-dhātu* deities, as at Tjandi Mendut.

The last great monument of the central Javanese period, Lara Jonggrang at Prambanan, is indeed a colossal work, rivalling Borobuḍur. It was probably built soon after 900. Not Buddhist but Hindu, the shrine represents the cosmic mountain. There were originally 232 temples incorporated into the design. The plan was centred on a square court with four gates containing the eight principal temples. Facing east, the central and largest temple, some 120 feet (40 metres) high, was devoted to the image of Śiva. To the north and south it is flanked by slightly smaller temples devoted to the two other members of the Hindu trinity, Vishnu and Brahmā. The smaller shrines contained many subsidiary images. The whole complex was enclosed, far off-centre, in an extremely large walled courtyard.

Although these are Hindu buildings, their high-terraced shrine roofs bear tiers of elongated and gadrooned stupas. The reliefs on these structures are especially beautiful. One series, representing the guardians of the directions, integrates the ornamental motifs with the plastic forms of the bodies in a most original way. The balustrades and inset panels abound with lively reliefs portraying various deities or scenes taken from the great Hindu classics, especially the *Rāmāyaṇa.*

East Javanese period: 927–16th century. During the east Javanese period a very large number of monuments were produced at the eastern end of the island (after 1222) and in Bali (after *c.* 1050). Few single structures, however, are as impressive and as comprehensively planned as are the monuments of Borobuḍur or Lara Jonggrang.

Around the strange natural mountain with tiered peaks cut and built in stone called Mount Penanggungan there are 81 structures (10th-century) of different kinds (now mostly in ruins). Prominent among these structures are bathing places. This mountain was identified by the people with the sacred Mount Meru, and its natural springs were believed to have a magical healing power and a mystical purifying capacity. Another such bathing place is Belahan (11th century). Made of brick, it, too, has extensive ruined temples. Belahan is supposed to have been the burial place of King Airlangga, who probably died about 1049. One of the greatest east Javanese icons formed the central figure against the back wall of the tank. Carved of red tufa (a porous rock), it shows the god Vishnu seated at peace on the back of his violently dramatic bird-vehicle, Garuḍa. It is said that the image represents the King himself in divine guise. Beside this image was a sculpture of a type associated with many of these sacred bathing sites. It is a relief of a four-armed goddess of abundance, her two lower hands holding jars pierced with holes, her two upper hands squeezing her breasts, which are also pierced; through the holes the sacred water flowed into the basin. There are many variants of this idea at the springs of Mount Penanggungan. On Bali the same kind of fountain sculpture appears at the Goa Gadjah, at Bedulu, in a spring-fed tank below a cave.

Four-armed goddess of abundance

In both Java and Bali there are many rock-face relief carvings from this period (there are no secure dates). Some represent legendary scenes; others represent *tjandis*; the shallow chambers of others are thought to be royal tombs.

The structure that gives the best ideas of what the typical east Javanese shrine of the mid-13th century was like is Tjandi Kidal. The nucleus of the building is a square cell, with slightly projecting porticoes each hooded by an enormous Kāla-monster head. But the cell itself is dwarfed both by the massive molded plinth upon which it stands and by the huge tower with which it is surmounted. The tower stands above an architrave stepped far out on tiered moldings. It is no longer composed of diminishing stories, as earlier towers were, but is conceived as a massive pyramidal obelisk made up of double bands of ornament spaced by stumpy pilasters and bands of recessed panels. The architectural projections and moldings distinguish Tjandi Kidal from earlier Javanese architecture, with its plain wall surfaces.

Many masterpieces of sculpture belong to the east Javanese period. Among them are some superb icons of Śiva and of a goddess of Buddhist wisdom from Singhasāri and a splendidly "primitivist" image of the elephant-headed god of wealth from Bara, Blitar.

From the late 13th century onward a whole series of *tjandi*s was created in eastern Java. As time went on the *tjandi*s lost their monumental scale and became simply shrines within a series of courtyards on a pre-Indian pattern. From Tjandi Djago through Tjandi Panataran at Blitar (14th century) and Tjandi Surawana it is possible to trace the line of descent of the modern Balinese temple enclosures.

Influence of shadow puppets on relief figures

By the end of the 14th century, the figures in the relief sculpture at these shrines had come more and more to resemble the shadow puppets of the popular *wayang* drama. They adopt the stiff profile stance that presents both shoulders, while the trees and houses resemble the stereotype silhouette leather and wood cutouts used as properties in the shadow plays. The art of carving in the near-full round, however, did not follow the same course of evolution as the reliefs. Such work did become softer and more delicate in style, with accretions of broad floral forms, but well into the 15th century the icons retain something of the strength of older sculptural conceptions. Another plastic tradition that seems to have escaped domination by the *wayang* formula resulted in the production of beautiful small terra-cotta figures as part of the revetment (stone facing sustaining the embankment) of the east Javanese capital city of Majapahit. Like the reliefs, the many small excavated bronzes of Hindu scenes are under the *wayang* influence, three-dimensional though they may be. Curlicues proliferate, and the plasticity of bodies is virtually ignored.

16th century to the present. When Islām arrived in Indonesia, it used the repertoire of traditional ornament for its mosques and tombs; but, in conformity to a puritan Muslim custom, the representation of living creatures was excluded on religious buildings. The gates of the 16th-century mosque at Sendangduwur, Badjanegara, show a splendid example of this adaptation. The wings of the old Hindu Garuḍa, a colossal bird-vehicle of the high god Vishnu, frame the gate; the body and head are suppressed. Above the lintel are abstract tree-clad mountain forms recalling the imagery of the cosmic Meru; and legendary snakes hood the jambs. The 16th-century mosque at Kudus even has a gate based on the split-*tjandi* pattern used in Bali (see below). Tombs such as that of Ratu Ibu at Airmata, on the island of Madura, add to their simple volumes elaborate but abstract variants of the scroll-filled antefixes of older architecture and of the petal-shaped aureoles of the larger east Javanese icons. In Sumatra the Muslim rulers encouraged a revival of the pre-Indian ancestor cult, along with its ancient and characteristic arts.

Bali. The rajas of eastern Java finally retreated before the Muslim invaders during the 16th century and departed to the island of Bali, where they remained. The old Javanese Indianized culture they brought with them survived and combined with animist folk elements. In Bali today that culture has bred a widespread popular art. There are now many hundreds of temples in Bali of varying age. Each family group has its own temple, dedicated to the ancestors; each village, too, has its temple, in which special attention is paid to a rich fertility goddess identified with the ancient Indian goddess of bounty, Śrī. Special temples dedicated to the goddess of death stand near the cremation ground. There are numerous major temples—many associated with volcanic peaks—dedicated to different deities and spirits; they range in size and importance from Besakih on Mount Agung (where a megalith is incorporated as a phallic Śiva-emblem) to Panataram Sasih of Pedjeng (where the bronze drum called "the Moon of Bali" is preserved).

Balinese temples

Balinese temples are conceived as multiple courts raised on terraces. The tall stone or brick and plaster gates are shaped like a *tjandi*-tower split down the centre; they are usually encrusted with ornament based upon deep multiple curlicues interspersed with simplified, two-dimensional relief figure sculpture. Fantastic three-dimensional guardians sometimes stand at the foot of the access staircase. Beyond the gates are one or two courts within which various ceremonies (including sacrifices and cockfights) may take place. The rearmost court backs onto the mountain, whence spirits descend temporarily when invoked. The court has no icons; at most, there is a seat for invisible deities. The structures in the court, mostly of wood and thatch, may be of many stories. (Such structures are called *merus*.) Sometimes the treasuries are ornamented with carving; and a few older stone *meru* towers in local shrines are carved with mythological figures.

Temple ceremonials, especially the cremation of distinguished people, evoked elaborate ritual art objects in precious metals, as well as in wood or fabric. All were characterized by exuberant and repetitive curvilinear floral ornament and by figures based on Indian legend, especially the *Rāmāyaṇa* and parts of the *Mahābhārata.* In the villages today, music, dance, sculpture, and painting are focussed on the shrines and are practiced with an intensity unknown elsewhere in the world. Art is woven intimately into the life of the people. The masks carved of wood for the dances are specially refined, sometimes ornate versions of the masks used in the animist rituals of other Southeast Asian peoples. In the 20th century there are numerous village sculptors and painters, who sell to tourists work based upon the old ceremonial arts. During the 1930s an outside impetus to develop their traditional legendary imagery in Western formats came from a German painter, Walter Spies, who lived on Bali. A landscape tradition was evolved, and the painters have been able to communicate something of the extravagant visual charm of their island, giving glimpses of luminous village and mountain landscape. The style of both sculptors and painters, however, is based upon gently undulating curves and is often highly ornamental, with repeated patterns. A repertoire of posture and gesture has been abstracted from the *wayang.* The work thus tends to prettiness rather than vigour; the sculptors create no truly intelligible volumes, and the painters fill their surfaces with naïvely structured shapes.

Indonesia painting

Java: 20th century. A conscious revival of traditional art has been attempted in the 20th century, especially in Java, the main territory of modern Indonesia. There has been government support for the resuscitation of old crafts—silverwork, for example. A number of artists have adapted Westernized figure drawing to their own decorative compositions. The best known painter of Indonesia is the Javanese Affandi. He has used oil paint to execute pictures of Indonesian subjects in a vividly coloured Expressionist impasto (thick application of pigment to the canvas). This European brushwork technique, however, contains a strong element of the sinuosity of Javanese tradition. As yet, Affandi is the only artist from Southeast Asia to have attained a personal worldwide reputation.

THE PHILIPPINES

The population of this island group contains a number of different ethnic strata, the oldest of which shares in the general folk culture and its associated folk arts of the islands of Southeast Asia (see above *Indonesia*), with an emphasis on geometric simplification. An element in the Tagalog (a people of central Luzon) is perhaps descended from the oldest level of immigrants with a Paleolithic background. The Moro are Muslims, converted to Islām during the 15th and 16th centuries. Today they produce a decorative art in which old Muslim geometric motifs are combined with strong Chinese decorative influences (from Sung times, Chinese ceramics and textiles were imported). The decoration is applied primarily to textiles, weapons, and containers to hold the betel nuts that are chewed throughout Southeast Asia.

Influence of Spain upon Philippine art

The most important departure in Philippine art was the result of the Spanish conquest of 1571. Thereafter, the bishopric of Manila and all of Luzon became the focus for an elaborate development of Spanish colonial art, primarily devoted to the construction and decoration of Roman Catholic churches in the current flamboyant, highly ornate, and colourful colonial style. There is good colonial architecture in other islands, including Bohol and Cebu. A large quantity of religious sculpture of the canonical Christian subjects was imported from Mexico and from Spain itself. Sculptors and missionary painters also immigrated, and a powerful local school developed under the direct influence of the 17th-century Spanish artists Murillo and Alonso Cano. Local arts were encouraged in 1785 by the remission of taxes for religious artists. Because of the close colonial ties, the stylistic developments corresponded

San Agustin church, Manila, the Philippines, 1599–1614.
Bruce Coleman

substantially with those elsewhere in the Spanish Empire, and European prints served as models for local artists. Of the major early churches for which this sculpture and painting was executed, only San Agustin (1599–1614), in Manila, still stands; it was designed by Fray Antonio de Herrera, son or nephew of the great Spanish architect Juan de Herrera. During the 19th century the neo-Gothic style was imported, mainly through the Philippine architect Felipe Roxas, who had travelled in Europe and England. San Sebastian in Manila is a notable example of this style. The Spaniard Hervas, Manila's municipal architect from 1887 to 1893, favoured neo-Byzantine forms; *e.g.,* Manila Cathedral (1878–79).

ɔdern ilippine hitects

It was only in the later 19th century that any secular art flourished at all. Schools of fine art modelled on the European schools were set up between 1815 and 1820, and a number of painters began to work in versions of European academic styles, painting landscapes, portraits, and classical subjects. The best known among them are Juan Luna, Felix Resurrección Hidalgo, Antonio Malantic, and the genre painter Fabian de la Rosa. After the transfer of rule to the United States in 1898, industrialization began in earnest; the methods of the art schools were adapted, as in Europe, to the needs of modern commercial society. In the 1930s a substantial modern experimental school of Philippine architects began the remodelling of the industrial environment in terms of 20th-century architectural and design conceptions. Prominent names are Pablo Antonio, Carlos Arguelles, and Cesar Concio. Beginning in the 1930s but especially after World War II, artists in Manila adopted the Abstract and Expressionist styles current in the United States. After the devastation wrought by that war, Manila and other cities and towns were rebuilt, virtually anew, in local variants of the international business style.

FOLK ARTS

The arts of many regions in Southeast Asia remained either untouched or only slightly influenced by the Indianized arts of other regions. Such influence is found especially in regions where the gold trade flourished. In Sarawak (Bonkisam), for example, the remains of buildings similar to late Tantric east Javanese *tjandi*s have been discovered. Among a few people (*e.g.,* the Meo of highland Vietnam), vestiges of Indian erotic temple imagery have been adapted to local fertility ceremonies, and most of the religious ideas of the region show at least faint traces of Indian influence.

Save for the megaliths and Dong Son bronzes, most of the known folk-art objects are relatively recent, although their inspiration and types belong to traditions far older and geographically more far-reaching than the Indianized traditions.

The two main non-Indian art styles in the whole region have been provisionally named the "monumental" and the "ornamental–fanciful." They coexist virtually everywhere, though they probably represent two evolutionary phases. The principal manifestations of the monumental style are the megalithic monuments, although there is great variety among the megalithic customs of the many different populations in Sumatra, Laos, Indonesia, Borneo, and the Philippines. The influence of the ornamental–fanciful style, which is characterized especially by the scrolled spiral, insinuates itself even into many of the decorative arts, particularly in the curvilinear and flamboyant inflection given to ornamental motives in the major Indianizing styles.

The squatting ancestor figure

The link between the two styles is probably the ubiquitous squatting ancestor figure, cocked knees supporting elbows, carved in soft wood or woven in cane or fibre. These figures may be either male or female. Under special social circumstances in recent times, very large wooden versions of the figure have been used as substitutes for more conventional, standing megalithic ancestral monuments (Sumatra and Sabah). The custom is probably an old one. There can be little doubt, for example, that the Theravāda Buddhist images of Burma, Thailand, and Laos were accepted as special modifications of the ancestor image. The transition from revering numinous ancestor images whose identity had been forgotten to worshipping an Indianizing icon was easy for the tribal populations.

The complex significance of the original squatting ancestor figure enabled it to be used in a variety of contexts. It might have combined associations of the fetus, the fetal burial position, and female birth and intercourse positions, as well as a ceremonial posture assumed by the living. It came to be used primarily in wooden sculpture on all scales, but also in woven textiles (*e.g.,* Iban), to represent the continuing power informing human existence, both in the purely ancestral sense of family continuity and identity and in the sense of the fertility of the land. Its earliest recorded appearance may be on Chinese Yang-shao painted pottery (*c.* 2000 BC); but it appears in essentially the same form over a range of territory including Sumatra, Nias and Sunda islands, Java, Borneo, New Guinea, Taiwan, the Philippines, and out into northern Australia and Melanesia. It may be used purely as an ancestral image in a family shrine house or as a motif added to any one of a variety of implements to potentiate them; for example, large bowls (Sumatra), kris or sword handles (Java, Sumatra, Borneo), spoons (Timor, the Philippines), musical instruments (Borneo), and magicians' staves (Borneo).

The treatment of such figures may be invested with more or less of the characteristics of the ornamental–fanciful style, in those regions where this style prevails (*e.g.,* Batak, Dayak). There are also special versions of the squatting figure that seem to belong especially to important magical crafts, such as the Javanese kris handle, on which miniature carvings can give an extraordinarily monumental effect. Sumatran Dayak hereditary magical staves may be carved with a "tower" or "tree" of such ancestor figures. On Nias, for example, along with the squatting figure, a standing figure in the bent-knee posture common in Polynesia also appears as a variant. In the Philippines similar variants are sometimes interpreted as vestiges from a remote Indian mythology, adopted probably for the sake of their cultural prestige. In southern Borneo the figure appears carved in the full round and as a pattern for woven textiles; it often has a protruding tongue and, sometimes, antlers—a combined motif known in the Ch'ang-sha art of southern China (*c.* 300 BC). Antlers also appear on certain Sumatran knifehilt figures. A variety of designs, some of them "abstract," are based on this figure. Among the Jarai of Vietnam, for example, a pattern of lozenges represents an abstraction from a group of these figures. Especially in the textiles of Sumba and other Indonesian islands, similar patterns, often referred to as decorated triangles, represent the same phenomenon. When, as in textiles, the anthropomorphic reference of the abstract pattern is lost, the male genitals may remain to assert the ancestor significance.

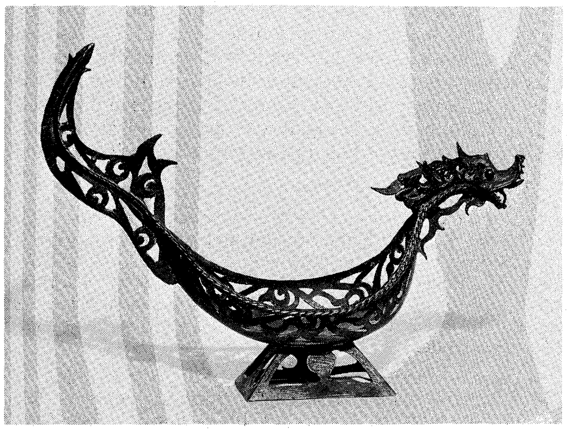
Cult of the skull

The association between the squatting figure and the widely practiced cult of the skull is manifested in the combined cult of ancestors, headhunting, and head worship. Among the Wa of Upper Burma, for example, the squatting figure in a lozenge abstraction decorates the chests in which the severed heads of enemies are stored. Virtually everywhere among the early farmers of Southeast Asia, such heads were regarded as repositories of great spiritual power. The cult of the skull has produced a version of the squatting figure that is commonly known by the Indonesian word *korvar;* it is a figure with an ancestral skull in place of a carved head. Such figures are especially common in the more easterly island cultures. The ghostly power of the deceased ancestor can thus become present and available to his descendants—to give oracular advice, for example. A related idea is incorporated in the masks used in a wide variety of rituals and dance-dramas throughout Southeast Asia; for example, among the Batak of Sumatra and the Dayak of Borneo, where especially fine examples are made. There can be little doubt that the same idea (blended with imagery from the imported Hindu epics) underlies the range of elaborate masks that were once used in the Javanese and now can be seen in the Balinese *wayang* dances. It is possible that the flamboyant flame skull protuberances and winglike flanges ornamenting the head in so much of the Buddhist art produced in Burma and Thailand reflect a persistent but submerged interest in the cult of the skull.

Another major motif is the snake, which (even in areas where direct Indianizing influence was not strong) is frequently combined with imagery derived from the cult of the powerful, magical Hindu *nāga;* often many-headed, this serpent is the patron and guardian of water and treasure, both material and spiritual. The snake motif has also been blended with images of the Chinese dragon, going back perhaps to Chinese Han ornamental designs. Outstanding examples are found on the elaborate relief-carved doors of Sumatran Batak houses; "flying" roof finials in many parts of Indonesia; and in much Borneo Dayak ornament, from tattoos to carved bamboos and bronze body ornaments. The snake is the magico-mythical creature that gives both its bodily shape (either straight or undulant) and its metaphysical power to the kris. Distributed from Malacca to Celebes, these swords (the earliest known dated 1342) reached their high point of artistic development in Java. A variety of other motifs originating on the mainland of Asia is found in many of the surviving folk arts of Indonesia. Among them are the "man in the embrace of an animal" (Dayak kris handles) and animals "stacked" one above the other (Timor, Indonesia).

Use of curves and spirals

The ornamental–fantastic style. The styles in which these variations on basic motifs are carried out vary principally according to the preponderance of the sinuous curves and spirals of the ornamental–fantastic style. This style serves as the basis for decoration and as a method of artistic phrasing. It may have made its way into Southeast Asia as late as the 1st millennium BC, being formally related to the spirals used in Chinese Neolithic, Shang, and Chou bronze art. (It should, perhaps, be mentioned that Chinese art, certainly until well into the Christian Era, was itself far more "tribal" than later Chinese tradition recognizes.) Probably connoting spirituality, the spiral imagery appears in Southeast Asian magical art at all levels, from the textiles of Java and the incised bamboo implements or carved doors of Dayak Borneo to the ornament on the costumes of sculptured dancers or deities at every major city site. Given a fiery upward inflection, it appears in the finials on major Indianized stone architecture and on the carved wooden gables of Burmese and Thai Buddhist halls. There is not always complete stylistic consistency within any one cultural group. For example, the fantastic snake-dragon creatures carved in deep relief on the housedoors of the Batak may be extravagantly sinuous, with many spirals, while their figure sculpture adheres to the sterner plastic idiom, virtually without any linear sinuosity. Among the Dayak of Borneo the fantastic style may be confined entirely to surface ornament. On Indonesian islands, ancestral figures may be relatively static and foursquare, while the decorative carving and textiles may display considerable linear fantasy. A special version of the ornamental–fantastic style characterizes the surviving Indianized arts of Bali and Java, intruding even into sculptural inventions derived from strongly three-dimensional medieval Indianizing patterns. Thus, the decoration on the *wayang* cutout leather puppets, with its somewhat stereotyped curlicues, has proliferated at the expense of the three-dimensional sense (see above *Indonesia*). Balinese *wayang* masks may be carved entirely out of curling surfaces and completed in paint with sinuous eyebrows and moustaches. In many parts of Southeast Asia, including Thailand, Vietnam, Burma, Sumatra, and Indonesia, designs originally based upon Indian flowering-scroll patterns can be found in architecture, textiles, theatre costumes, musical instruments, and wooden utensils, all efflorescing with extravagant curling ornament. It is unfortunately true that only in a few of its most serious manifestations does this kind of ornament display substantial artistic invention, with carefully varied, asymmetrical, complementary, and counterchanged curves. Usually, it degenerates into repetitive space filling, without variety or formal meaning.



Brass receptacle in the shape of the mythical serpent, *naga.* From Krui, Sumatra. In the Royal Tropical Institute Museum, Amsterdam. Height 5 cm.

By courtesy of the Royal Tropical Institute, Amsterdam

Batik and *ikat*

Textiles. Perhaps the types of folk art best known in the West are the textiles, especially batik and *ikat.* Both names refer to techniques practiced by different groups of people, who must have learned it from each other. Essentially Javanese but known in other islands, batik may have resulted from the imitation with dyes of South Indian painted cloths, probably before 1700. The essence of the technique is that melted wax is poured from a small metal kettle onto areas of a plain cotton cloth, which is then dyed, only the unwaxed parts taking the colour. The process can be repeated with several different colours. The oldest basic colours are indigo and brown; red and yellow were used later. The possible patterns range from lozenges and circlets through a large repertoire of cursive animal and plant forms. The batik technique can produce sumptuous and complex designs that not even the most elaborate weaving techniques can duplicate. It was encouraged by the Muslim rulers as a major element of social expression in garments and hangings.

Ikat is known among the Batak, in Cambodia, and especially among the dispersed Dayak people. It, too, probably originated in India. The extraordinarily difficult *ikat* textiles (woven cotton and occasionally silk, especially in Cambodia) are made primarily for use in important ceremonials and were regarded by their makers as major works of art. Before being woven, the thread is tightly tied at carefully calculated points in the hank (coiled or looped bundle); this is then dyed, the tied parts not taking up dye. The process may be repeated for different colours. As a consequence of the predyeing, designs appear as the thread is woven. In most *ikat* only the warp (the series of yarns extended lengthwise in the loom and crossed by the weft) is so treated; but in southern Sumatra a tie-dyed floating weft is added to the plain weft. Naturally, *ikat* designs tend to be static and more or less rectilinear. In the finest *ikat,* however, birds and animals, spirits and houses, and, in Cambodia, a vestigial iconography of royal Buddhism may be formalized into extremely beautiful banded compositions. (P.S.R.)

BIBLIOGRAPHY. REGINALD LE MAY, *The Culture of Southeast Asia* (1954), deals with the art and architecture of Southeast Asian peoples. GEORGE COEDES, *Les États hindouisés d'Indochine et d'Indonésie,* new ed. (1964; Eng. trans., *The Indianized States of Southeast Asia,* 1968) and *Les Peuples de la péninsule indochinoise* (1962; Eng. trans., *The Making of Southeast Asia,* 1966), are standard works.

Literature: (*Burma*): MAUNG HTIN AUNG, *A History of Burma* (1967), contains a general survey of Burmese literature. The same author's *Burmese Folk-Tales* (1948), *Burmese Law Tales* (1962), *Burmese Monk's Tales* (1966); and MAUNG MYINT THEIN, *Burmese Folk-Songs* (1970), provide a good survey of Burmese oral literature. Dramatic literature is treated in MAUNG HTIN AUNG, *Burmese Drama* (1937). (*Thailand*): P. SCHWEISGUTH, *Étude sur la littérature siamoise* (1951), is the standard work on Thai literature. The *Journal of the Siam Society* (various issues) contains studies on Thai literature. (*The Philippines*): LEONARD CASPER, "The Philippine Literature: The Unexplored Potential," *Asia,* 9:80–87 (1967); ALBERT RAVENHOLT, *The Philippines: A Young Republic on the Move* (1962). (*Vietnam*): NGUYEN NGOC BICH (ed.), "The Poetry of Viet Nam," *Asia,* 14:69–91 (1969). (*Malaysia*): RICHARD WINSTEDT, *The Malays: A Cultural History,* 6th ed. (1961); and OLIVER RICE and ABDULLAH MAJID (eds.), *Modern Malay Verse, 1946–61* (1963), are important works to consult. (*Indonesia*): T.G.T. PIGEAUD, *Java in the 14th Century,* 5 vol. (1960–63), is a special study of Javanese literature in the Majapahit period. See also A. TEEUW, *Modern Indonesian Literature* (1967). TAKDIR ALISJAHBANA, *Indonesia in the Modern World* (1961), contains good accounts of modern Indonesian literature against the political, cultural, social, and historical background.

Music: WILLIAM P. MALM, *Music Cultures of the Pacific, the Near East, and Asia,* ch. 2 and 5 (1967), discusses the music of the Southeast Asian region; LAWRENCE PICKEN, "The Music of Far Eastern Asia and Other Countries," in *The New Oxford History of Music,* vol. 1, *Ancient and Oriental Music,* pp. 83–194 (1957), discusses intercultural and musical relationships between Far Eastern and Southeast Asian countries. See also D.R. WIDDESS and R.F. WOLPERT (eds.), *Music and Tradition* (1981), essays on Asian music. (*Burma*): U KHIN ZAW, "Burmese Music: A Preliminary Inquiry," *Journal of the Burma Research Society,* pp. 387–466 (December 1940), appends several pages of music notations to a discussion of music history and the structure of Burmese scales. (*Cambodia and Laos*): ALAIN DANIELOU, "La Musique du Cambodge et du Laos" (1957), is a brochure that discusses with the help of illustrations traditional musical instruments of both countries. (*Indonesia*): JAAP KUNST, *Music in Java,* 2 vol. (1949), an important work with a comprehensive bibliography including works of many Dutch scholars, treats matters regarding history, vocal and instrumental music, structure, notation, and tonal systems of the East-Central and the West-Javanese gamelan; MANTLE HOOD, *The Nuclear Theme As a Determinant of Patet in Javanese Music* (1954), modal structure of patet is analyzed according to basic elements in themes of several gamelan pieces of music; COLIN MCPHEE, *Music in Bali* (1966), gives detailed descriptions and specific musical examples of repertoire played in many gamelan ensembles; WALTER KAUDERN, "Musical Instruments in Celebes," in *Ethnological Studies in Celebes,* vol. 3 (1927), a detailed study with a long list of names of musical instruments, 130 figures, and 19 maps showing the geographical distribution of musical instruments; JAAP KUNST, *Music in Nias* (1939), describes both vocal and instrumental music, classifies musical instruments (Hornbostel-Sachs divisions), and illustrates their distribution in 7 maps; *Music in Flores* (1942), lists in a table the native names of 54 instruments of the five divisions of the archipelago; CHARLES S. MYERS, "A Study of Sarawak Music," *Sammelbände der Internationalen Musikgesellschaft,* 15:296–308 (1913–14), 13 gongs from different cultural groups and music of some instruments are analyzed with the help of musical examples; HENRY L. ROTH, *The Natives of Sarawak and British North Borneo,* 2 vol. (1896)—vol. 1, ch. 9 deals with feasts, festivals and dancing, while vol. 2, ch. 26 discusses music in general; EDWIN H. GOMES, *Seventeen Years Among the Sea Dyaks of Borneo* (1911), terms for songs and names of musical instruments that the author cites are still known in Sarawak; IVOR EVANS, *Among Primitive Peoples in Borneo,* ch. 14 (1922), devoted exclusively to a discussion of musical instruments, music, and dancing; JAAP KUNST, *Music in New Guinea* (Eng. trans. 1967), comprises three works first published in 1931 and 1950 that treat of vocal and instrumental music of the Papua in the north and the central range of mountains, of songs in the north and the West, and of music in the West Central range, Southwest, North and West coasts (many musical examples and a distribution map of musical instruments). (*The Philippines*): JOSE MACEDA, "The Music of the Magindanao in the Philippines," (1963, Ann Arbor University Microfilms), discusses instrumental and vocal music of the whole culture with copious musical examples and two long-playing records available at Folkways Records, New York; and "Drone and Melody in Philippine Musical Instruments," paper read at an International Conference on Traditional Drama and Music of Southeast Asia, Kuala Lumpur, August 27–31 (1969), discusses age, spread, and variety of combinations of drone and melody as a root structure. (*Thailand*): DAVID MORTON, "The Traditional Instrumental Music of Thailand," (1964, Ann Arbor University Microfilms), a study that covers both the historical and structural aspects of Thai music. (*Vietnam*): TRAN-VAN-KHE, *La Musique Vietnamienne traditionelle* (1962), detailed and important information about history, musical instruments, and musical theory (references abound with criticism of some works); and *Vietnam* (1967), a clear and concise presentation, containing valuable new historical material.

Dance and theatre: JAMES R. BRANDON, *Theatre in Southeast Asia* (1967), a general survey of major theatre forms, incorporating firsthand observation. (*Burma*): MAUNG HTIN AUNG, *Burmese Drama* (1937), a detailed history and translations of plays, four complete and eight in excerpts; KENNETH SEIN and J.A. WITHEY, *The Great Po Sein* (1965), a lively chronicle of the Burmese stage of the last century, cast in the first person narrative, by a famous contemporary actor; U POK NI, *Konmara Pya Zat* (1952), a complete English translation of a 19th-century play with interpretive notes. (*Cambodia*): JACQUES BRUNET, "Nang Sbek, Danced Shadow Theatre of Cambodia," *World of Music,* 11:18–37 (1969), a brief and excellent description and history of Cambodian shadow theatre; SAMDACH CHAUFEA THIOUNN, *Danses cambodgiennes,* 2nd ed. (1956), the most complete history and analysis of plays, music, and dance of female and male dance-drama. (*Indonesia*): BENEDICT R.O'G. ANDERSON, *Mythology and the Tolerance of the Javanese* (1965), major *wajang* characters are described and their significance as behavioral models is treated; JAMES R. BRANDON (ed.), *On Thrones of Gold: Three Javanese Shadow Plays* (1970), translations, with description of action, music indication, and photographs, of three *wajang kulit* plays; CLAIRE HOLT, *Art in Indonesia* (1967), excellent chapters on *wajang kulit,* dance, and dance-drama in all parts of Indonesia; JAMES L. PEACOCK, *Rites of Modernization* (1968), the plays, structure and content, of *ludruk* in Java as seen through the eyes of a modern anthropologist; W.H. RASSERS, *Pañji, the Culture Hero* (1959), a disputed but brilliant theoretical discussion of the origin and meaning of Indonesian theatre; H. ULBRICHT, *Wayang Purwa: Shadows of the Past* (1970), useful for its extended translations of synopses of Pandawa plays found in J. KATS's famous Dutch work *Het Javaansche tooneel,* vol. 1, *Wajang Poerwa* (1923); BERYL DE ZOETE and WALTER SPIES, *Dance and Drama in Bali* (1938), an authoritative and encyclopaedic pre-World War II description of Bali's performing arts. See also I MADÉ BANDEM and FREDERIK E. DEBOER, *Kaja and Kelod: Balinese Dance in Transition* (1982); and ANA DANIEL, *Bali: Behind the Mask* (1981). (*Malaysia*): JEANNE CUISINIER, *Le Théâtre d'Ombres à Kelantan,* 2nd ed. (1957), Malaysian shadow theatre described and illustrated, with a partial play translation; RICHARD WINSTEDT, *The Malays: A Cultural History,* 6th ed. (1961), contains background information on Malaysian drama, including translations of pre-performance invocations. (*The Philippines*): JEAN EDADES (ed.), *Short Plays of the Philippines* (1950), a collection of recent one-act plays in English; ALBERTO S. FLORENTINO, *Outstanding Filipino Short Plays* (1961), short plays with two appendixes on traditional and modern drama in the Philippines; FRANCISCA TOLENTINO, *Philippine National Dances* (1946), descriptions and brief histories of many folk dances. (*Thailand*): UBOL BHUKKANASUT (trans.), "Manohra," in *Traditional Asian Plays,* ed. by JAMES R. BRANDON (1972), a translation with stage directions of the *lakon jatri* play "Manohra"; H.H. BRIDHYAKORN, "The Nang," 3rd ed. (1956), a short booklet on *nang yai* shadow play; and with DHANIT YUPHO, "The Khon," 3rd ed, (1956), a short booklet on *khon* masked pantomime; DHANIT YUPHO, *Classical Siamese Theatre* (1952), 52 folk and classical dance sequences described and illustrated; and *The Khōn and Lakon* (1963), synopses, commentary, and illustrations for 32 classic dance plays as performed by the Bangkok Department of Fine Arts between 1947 and 1960. (*Vietnam*): SONG-BAN, *The Vietnamese Theatre* (1960), brief descriptions of theatre in Vietnam.

Visual arts: PHILIP S. RAWSON, *The Art of Southeast Asia: Cambodia, Vietnam, Thailand, Laos, Burma, Java, Bali* (1967),

a comprehensive survey with many illustrations and plans; BENJAMIN ROWLAND, *The Art and Architecture of India,* 3rd ed. rev. (1967), sets the art of the region in relation to South Asian, or Indian art; WIM SWAAN, *Lost Cities of Asia* (1966), a pictorial study of Ceylon, Pagan, and Angkor; detailed articles in the *Encyclopedia of World Art* (1960–67): GEORGE COEDES, "Burmese Art," "Khmer Art," and "Cham Art"; A.B. GRISWOLD, "Siamese Art"; MADELEINE HALLADE and ROBERT HEINE-GELDERN, "Indonesian Art"; LOUIS BEZACIER, "Vietnamese Art"; and J.M.R. RIVIERE, "Philippine Art"—all these articles have extensive bibliographies. (*Burma*): A.B. GRISWOLD, CHEWON KIM, and P.H. POTT, *Burma, Korea, Tibet* (U.S. title, *The Art of Burma, Korea, Tibet;* 1964), the only survey of Burmese art; BURMA RESEARCH SOCIETY, *Fiftieth Anniversary Publications,* vol. 2 (1960), source material including important articles in English by G.H. LUCE and W.B. SINCLAIR. (*Champa*): The basic studies are PHILIPPE STERN, *L'Art du Champa (ancien Annam) et son évolution* (1942); and LOUIS BEZACIER, *Relevé des monuments anciens du Nord Viêt-nam* (1959). (*Siam and Laos*): Three basic sources include SILPA BHIRASRI, *Thai-Mon Bronzes* (1957), and *The Origin and Evolution of Thai Murals* (1959); and L. BORIBAL BURIBHAND and A.B. GRISWOLD, "Sculpture of Peninsular Siam in the Ayuthya Period," *Journal of the Siam Society,* 38:1–60 (1951). PIERRE DUPONT, *L'Archéologie mône de Dvāravatī* (1959), is the most authoritative study of this topic so far. The chief illustrated source for dating Buddhist art is A.B. GRISWOLD, *Dated Buddha Images of Northern Siam* (1957). See also CAROL STRATTON and MIRIAM M. SCOTT, *The Art of Sukhothai: Thailand's Golden Age* (1981), covering the period between the mid-1200s and the mid-1600s. (*Indochina*): JEAN BOISSELIER, *La Statuaire khmère et son évolution,* 2 vol. (1955), an exhaustive study of the development of Khmer sculpture; GEORGE COEDES, "Le Culte de la Royauté divinisée . . . ," *Série Orientale,* conference vol. 5 (1952), a basic iconographic study; PIERRE DUPONT, *La Statuaire préangkorienne* (1955), the authoritative book on pre-Angkor sculpture; LOUIS FREDERIC, *Sud-Est Asiatique: ses temples, ses sculptures* (1964; Eng. trans., *The Temples and Sculptures of Southeast Asia* (U.S. title, *The Art of Southeast Asia: Temples and Sculpture;* 1965), a well-documented pictorial survey; BERNARD P. GROSLIER, *Indochine: carrefour des arts* (1961; Eng. trans., *Indochina: Art in the Melting Pot of Races,* 1962), the most comprehensive survey in English; and with JACQUES ARTHAUD, *Angkor, hommes et pierres* (1956; Eng. trans., *Angkor: Art and Civilization,* rev. ed., 1966), a thoroughly documented pictorial survey. (*Indonesia*): F.A. WAGNER, *Indonesia: The Art of an Island Group* (1959), a comprehensive survey of Indonesian art in English; A.J. BERNET KEMPERS, *Ancient Indonesian Art* (1959), a comprehensive illustrated book. Major works on Borobuḍur include T. VAN ERP, *Beschrijving van Barabudur,* vol. 2, *Bouwkundige beschrijving* (1931); N.J. KROM, *Barabudur: Archaeological Description,* 2 vol. (1927); and PAUL MUS, *Barabudur, esquisse d'une histoire du bouddhisme fondée sur la critique archéologique des textes* (1935), which covers the subject of the role of Indian philosophy and theology as a background to Borobuḍur.

Southern Africa

The subcontinent of southern Africa is a vast region that includes the countries of Angola, Botswana, Lesotho, Malaŵi, Mozambique, Namibia, Swaziland, Zambia, and Zimbabwe. The nation of South Africa, as well as the dependent states within its borders, is treated in a separate article. The region of southern Africa has been the background for continuous political struggle of European colonizers and the white elite of emerging nations with the diverse national groups seeking self-determination. (Ed.)

The article is divided into the following sections:

PHYSICAL AND HUMAN GEOGRAPHY

The land

RELIEF

ıree ajor cles of osion

Vast areas of southern Africa are characterized by uniformity and monotony of landform and natural vegetation. Level or slightly undulating surfaces occur widely and sometimes give the impression of being a lowland plain, even though lying at a considerable elevation above sea level. Much geomorphologic investigation remains to be done, and more precise correlation is needed between the surfaces recognized in different parts of the continent. There is a fair measure of agreement on the view that at least three major cycles of erosion have been completed and are represented by well-marked surfaces, certainly in southern Africa where most field study has so far been done. The oldest of these is a high-level erosion surface of Late Jurassic Age at between 7,000 and 8,500 feet (2,100 to 2,600 metres) above sea level. It has been recognized in the Cape ranges, the mountains of Lesotho, and, farther north, in the Nyika Plateau of Malaŵi. The second surface (of Miocene or mid-Tertiary Age) stands at about 4,500 feet and is best preserved where rivers like the Orange and Limpopo and, to the north, the Zambezi and Congo have carved out great valleys in troughs that were filled with easily erodable Karoo sedimentary rocks. The third well-developed surface, occurring at between 2,500 and 4,000 feet, is Pliocene or Late Tertiary and is most marked where the sedimentary rocks are weaker and in the Cretaceous beds of the Luangwa Valley and of the Lake Malaŵi Trough.

The East African Rift Valley extends into southern Africa. The edges are obvious enough to the south in Malaŵi, where a huge crusted block collapsed along the parallel faults that constitute the steeply rising slopes of Lake Malaŵi. The lake is 360 miles (579 kilometres) long but never more than 50 miles wide, and has a maximum depth of 1,226 feet. The rift then follows the line of the Shire Valley to reach the coast of the Indian Ocean near Beira, Mozambique. The western branch, or Western Rift Valley, extends from the northern end of Lake Malaŵi in a great arc.

DRAINAGE

The Zambezi is 2,200 miles long, and its basin covers 513,500 square miles (1,330,000 square kilometres). Its upper course drains a huge shallow alluvial basin more than 4,000 feet above sea level; many of its tributaries there are intermittent because of the highly seasonal nature of the rainfall and the intensity of evaporation. The falls and rapids are in its middle course, the most spectacular being the Victoria Falls, 355 feet high (*cf.* Niagara, 167 feet). Below the falls the river follows a narrow zigzag course developed along successive fault lines. Farther downstream the Kariba Dam has formed a huge lake in the section known by the Matabele (Ndebele) people as Gwembe. Nearer the Indian Ocean the river is more than five miles wide in places.

The Orange River, with only one important tributary, the Vaal, is 1,300 miles long with a basin of 329,000 square miles. It drains some of the higher parts of the South African plateau westward to the Atlantic; but it loses so much of its water through evaporation and, in recent years, through withdrawal for irrigation that in drought years it may even be dry. Much of the rainfall never reaches the main river but collects in shallow lakes (*vleis*) and is evaporated so that the lakes become salt-encrusted.

Also deserving individual mention is the great gray-green Limpopo River, all set about with fever trees, draining an area between the Zambezi and the Orange. (Ro.W.St.)

CLIMATE

A tropical climate is found in Zimbabwe, Zambia, Malaŵi, and Angola, and in part of Botswana. With the exception of the deep Zambezi Valley, the whole of this area lies above 3,000 feet, and in the east some of it is above 6,000 feet. Rainfall is moderate (30 to 50 inches [760 to 1,270 millimetres]), and the temperatures are lower than in the tropical region north of the Equator. The daily maximum temperature in summer at about 5,000 feet is approximately 80° to 85° F (27° to 29° C). The hottest areas are in the deep valleys of the Zambezi (Luangwa and Kafue), where temperatures of more than 100° F (38° C) are common, and the coolest are near the eastern mountains.

Zone of transition

To the south is a dry region forming a zone of transition from the savanna to the deserts. The northern Kalahari and southern Angola belong to this dry region. Rainfall is low (15 to 30 inches), and the rainy season lasts only three to five months. Temperatures are high throughout the year (mean October maximum: 95° F [35° C] at Maun, Botswana). The seasonal range of temperature is small (varying about 18° F [10° C] at Maun), and there is no real winter. The hot weather at the end of the long drought is uncomfortable, and, despite the rise in humidity, the arrival of the rains brings relief from the heat.

Desert climate. The Kalahari Desert in southern Africa extends from the Orange River to the Zambezi and has a summer rainfall varying from 10 inches in the south to more than 25 inches in the north. Practically the whole area supports a natural vegetation of some economic value. Normal summer maximum temperatures are close to 90° F (32° C), and days with temperatures as high as 106° F (41° C) can be expected only about once a month. The winter temperatures are more than 70° F (21° C) in the daytime with light frosts at night and occasional severe frosts at altitudes above 3,500 feet. The Namib and the southern end of the plateau in Namibia also belong to this region of desert climates. Although the rainfall (5 to 15 inches) is meagre, nearly half of it falls in winter, allowing for a vegetation of low scrub that provides grazing for sheep and goats. Summer temperatures are similar to the Kalahari, but winters are somewhat colder.

The coastal zone of the Namib is a narrow sandy desert, cool and almost rainless, stretching from Angola to Cape Province. True desert begins some distance south of Luanda where the mean annual rainfall is 13 inches; Moçâmedes (Mossâmedes) has only three inches of rain, and the coast of Namibia almost no rain at all. The upper air over the coast is warm, but the surface air is cooled by the Benguela Current, and the coast is liable to frequent low cloud and fog. Called *cacimbo* in Angola, it occurs mainly in the cool season, spreading inland from the sea in the early morning and dispersing a few hours later. In Namibia, however, it becomes a thick layer of stratus clouds about 1,000 feet high extending from the seaward limit of the cold water to about 30 miles inland. Almost every morning in summer and on many days in winter, the sky is overcast and a fine drizzle falls and wets the ground. Toward afternoon the sky clears, and a strong, almost violent sea breeze comes up to raise the loose sand of the desert into clouds of dust. The hottest days are in the winter months when the cool weather is occasionally disturbed by the berg winds of South Africa—hot dry winds from the direction of the interior of the plateau that make the temperatures on the cool coast rise to more than 90° F.

Humid subtropical. The plain in the southeast of Mozambique has a warm and humid climate, which is distinctly subtropical in the north but temperate in the south where it becomes cooler and merges into the region of year-round rainfall found on the southern coast. The extent of the region is restricted by altitude, losing its subtropical character above 2,000 feet. Rainfall occurs mainly in summer and varies from about 30 to 50 inches (Maputo averages 32 inches). The warm Mozambique Current exerts a strong influence on the temperature of the whole coastland of Mozambique and Natal, and daily maxima in summer vary between 80° and 90° F (27° and 32° C; the mean January maximum at Maputo is 86° F). These temperatures combined with a relative humidity of about 65 percent are uncomfortably warm, but in the winter the climate of this coast is the most genial in southern Africa (mean July maximum at Maputo: 75° F).

Rainy season

The rainy season in Mozambique (November to April) corresponds to the season of the northern monsoon, which is a southern extension of the Asian circulation into the Mozambique Channel. The southern monsoon, which blows in winter, is part of the trade-wind circulation of the Southern Hemisphere and does not usually bring rain.

Temperate plateau. The grassland plateau is the area from 3,000 to 6,000 feet in altitude and includes the highveld. The main features of the climate are low rainfall, cloudless skies, and large diurnal changes in weather. Local differences of rainfall and temperature are due to relief and exposure, and a number of isolated areas of wet, cool, and misty mountain climates are found on the prominent eastern escarpment. Seasons are sharply contrasted. Summer is warm, and the rains, which are mostly convectional, occur then; winter is cool and dry. Mean annual rainfall varies from about 20 to 40 inches decreasing from east to west.

Highland climate. Highland climates occur where the altitude exceeds 6,000 feet. The Lesotho highlands with their natural vegetation of mountain grass are the only region of southern Africa that can be considered in this category. The higher mountain slopes are cool, misty, and often wet; in the winter they are snow-covered and intensely cold. The population is therefore concentrated in the valleys where conditions are less severe. (Mokhotlong, Lesotho, at an altitude of 7,800 feet, has an average annual rainfall of 22 inches and a mean July maximum temperature of 28° F [−2° C].) (S.P.Jn.)

Traditional cultures

Grouping and distribution. The natives of southern Africa are conventionally classified into four major divisions, which are based primarily on language distinctions: San (Bushmen), Khoikhoin (Hottentots), Bergdama (Damara), and Bantu. The San, originally scattered throughout the region, are now found chiefly in the Kalahari Desert and adjacent areas to the north and west. The Khoikhoin are represented only by the Nama in the southern half of Namibia. Farther south and east there were formerly three other groups: Cape Hottentots, Eastern Hottentots, and Korana. These groups have disappeared almost completely, partly because of the effects of war and disease but mainly as a result of absorption into the Cape Coloured and other mixed-blood communities of which they were a basic element. The Bergdama live chiefly in the centre and north of Namibia, to which in historical times they have always been confined. The Bantu, far more numerous and widespread, consist of the following four subdivisions:

Divisions based on languages

Eastern. Located in the coastal regions east of the Drakensberg Mountains and south of the Sabi River, the eastern division comprises two main clusters, Nguni and Tsonga. The Nguni include the Cape Nguni of Ciskei and Transkei (Xhosa, Tembu or Thembu, Mpondo, and others, as well as the Mfengu, fugitive remnants of tribes broken up in Natal during the great intertribal wars at the beginning of the 19th century); Natal Nguni or Zulu of Natal and Zululand, with their offshoot the Ndebele (Tebele) of Zimbabwe; Swazi of Swaziland and eastern Transvaal; and Ndebele of central and northern Transvaal. The Tsonga, subdivided into Tsonga, Ronga, and Tswa, are found chiefly in Mozambique, with offshoots in eastern and northern Transvaal.

Central. This subdivision occupies most of the interior plateau north of the Orange River and west of the Drakensberg. It also comprises two main clusters: (1) Sotho, including the Southern Sotho of Lesotho (formerly Basutoland) and adjoining districts, Tswana (Thlaping [or Tlhaping], Rolong, Hurutshe, Kwena, Ngwaketse, Ngwato, Tawana, Kgatla, and others) of Botswana (formerly Bechuanaland) and western Transvaal and Northern Sotho (Pedi and many others) of central and northern Transvaal, and (2) Venda, a homogeneous cluster inhabiting the Soutpansberg district of northeastern Transvaal.

Shona peoples. These groups of Zimbabwe and Mozambique are subdivided linguistically into Zezuru, Manyika, Karanga, Kalanga, Korekore, and Ndau.

Western. The fourth group comprises the Ambo in northernmost Namibia and the Herero in central Namibia and northwestern Botswana. The Ambo consist of several distinct tribes, notably Ndonga, Kwanyama, and Kwambi; the Herero are customarily divided into Herero, Mbanderu, and Tjimba.

Major population movements. Because of the remote affinities of San (Bushman) languages with the Hadza and the Sandawe languages of Tanzania, it used to be held that the San had entered southern Africa from the northeast. Later research, however, indicates that they developed by local differentiation from some preexisting southern people and then spread north of the Zambezi. Although relating existing to prehistoric populations is largely guesswork, it is nevertheless possible that some of the Late Stone Age (about 8,000 years ago) inhabitants of southern Africa may have been the ancestors of the modern San. Relics of their much more recent ancestry (*e.g.*, stone artifacts, pictorial art, and place-names) are found throughout the region in areas where the San no longer live—notably Cape Province, Orange Free State, Lesotho, and Natal.

Relics of the San

An East African (and partly Hamitic) origin has also been claimed for the Khoikhoin, but this belief is now likewise largely rejected. Since some speakers of Khoi languages still practice hunting and gathering, one may suppose that this was perhaps the original means of livelihood of an indigenous Khoi population (also arising from local differentiation) and that sections of this population later received livestock from a source that remains uncertain.

The origins of the Bergdama are also problematic. If less importance is attached to their physical appearance, however, and more is attached to the fact that they speak a Khoi language, it may be possible to regard them as descendants of a section of the early Khoi population that did not acquire livestock.

Migration of the Bantu

The Bantu are universally held to have moved into South Africa from the north, although there is no certain evidence of their place of origin. They apparently came in many separate bodies and along diverse routes, each of their main groups representing a different series of migrations. Archaeological evidence, notably from the Great Zimbabwe and similar ruins, suggests that the ancestors of the Shona were already in what is now Zimbabwe by the 8th century AD. Other evidence indicates that the Nguni and Sotho had reached their present habitats by the 14th century, that the Venda migrated from Zimbabwe to the Soutpansberg about 1600, and that the Ambo and Herero entered what is now Namibia from the upper reaches of the Zambezi at about the same time. More cannot be said with much plausibility, and even the dates mentioned are hypothetical.

The impacts of different native peoples, upon one another, as well as the later advance of European settlers inland from the Cape, resulted in many local upheavals, during which various groups of San and Khoikhoin ceased to exist and the Bantu also were greatly affected. In particular, the wars of conquest initiated by the Zulu chief Shaka (1818–28) led to widespread devastation and enforced migration. One sequel was the creation of several other strong native states. Mzilikazi, after dominating central and western Transvaal (about 1821–37), went on to conquer the Shona in Rhodesia, where he established the Matabele (Ndebele) kingdom. Soshangane subjugated most of the Tsonga in Mozambique and founded the Gaza kingdom, while Zwangendaba and others founded Ngoni kingdoms north of the Zambezi. Mshweshwe (or Moshesh or Moshoeshoe), from the remnants of many scattered tribes, similarly created the modern Basuto nation in Basutoland (now Lesotho). Thulare and his son Sekwati brought many northern (Transvaal) Sotho tribes under the rule of the Pedi; Sebetwane, leading a horde of Sotho refugees from the south, established in far-off Barotseland (Zambia) the great Kololo kingdom. The wholesale destruction of life and dispersal of people resulting from these upheavals facilitated the subsequent extension of European settlement over the country. By the end of the 19th century, European armies had destroyed most of the new kingdoms, but such large tribal states as the southern Sotho (of Lesotho), Swazi, and Ngwato survived as examples of Bantu political development. (I.S./W.J.Ar.)

THE KHOISAN PEOPLES

The term Khoisan is a composite for Khoikhoi herders and San hunter-gatherers; it designates the indigenous southern African peoples once known, respectively, as Hottentots and Bushmen. The nature of the relationship between the two Khoisan groups has given rise to much discussion. They closely resemble each other in physical appearance and are markedly different from any of the other southern African populations. The prevailing view is that they are of the same genetic stock and differ only in their cultures, the Khoikhoin (Hottentots) being pastoral while the San (Bushmen) were formerly all hunter-gatherers. The Central San languages, Naro (Nharo), G/wi, and G//ana (/ and // represent click sounds), have such close lexical and structural resemblance to Khoi languages that all have been classified as belonging to the same language family.

hysical ppearance f the hoisan eoples

San and Khoikhoin are lightly built with thin limbs and wiry muscles. They are shorter in stature than most southern African peoples; San males average about five feet two inches (1.57 metres), women being an inch or so shorter, and Khoikhoin are perhaps two inches taller than San. Their skin colour at birth is coppery yellow, darkening to yellowish brown with exposure to the sun. Their dark-brown eyes are narrowed by the fullness of the upper lids and the pronounced epicanthic folds, a feature that has led some erroneously to link them with Mongoloid peoples. Facial and body hair are sparse, and few men develop beards before middle age.

The now-extinct San south of the Molopo River were smaller and lighter coloured than are the present-day San of Botswana, Namibia, and Angola. The latter are thought to have acquired Negroid characteristics from intermixing with Bantu-speaking peoples.

THE KHOIKHOIN, OR HOTTENTOTS

Origin of the name Hottentot

The origin of the name Hottentot is uncertain. The most plausible theory is that it was invented in 1647 by the survivors of a Dutch shipwreck who camped in the vicinity of present-day Cape Town for a year before being rescued. They established friendly relations with the local inhabitants but were quite unable to learn their language. One theory is that the Dutchmen took the refrain *autentou* from a ¾-time dance of the people and used it in friendly, joking reference to them. Another theory is that the Dutch formed the word in an attempt to imitate Khoi click sounds in Dutch. Whatever its origins, the form Hottentot was later adopted by the Europeans and subsequently by the Hottentots themselves. The original Hottentot self-designation had been Khoikhoi ("Men of Men" or "Proper People").

Knowledge of traditional Khoi culture is fragmentary and imprecise. The only records come from travellers and missionaries who were untrained observers and therefore strongly influenced by their own values and preconceptions. By the time capable and interested observers came to the southern African scene, war and disease had taken their toll, and only vestiges of the traditional cultures remained.

Subsistence and economics. The Khoikhoin were pastoral, grazing their fat-tailed sheep and long-horned cattle over much of the drier, western half of the subcontinent. The native sward and the meagre supplies of water were unsuited to sustained stock raising, and the Khoikhoin had periodically to trek their flocks and herds in search of new pasture and water. No crops were cultivated. The diet of milk and meat was supplemented by game and wild plants.

Their settlement pattern appears to have varied with the conditions of the locale, but a typical community may have consisted of 200 or 300 people, together owning upward of 1,000 head of stock. The temporary settlement was enclosed by a circular fence of thornbush, within which were the kraals (corrals) for lambs and calves. No special enclosures were made for adult beasts, which were driven into the main enclosure at night among the huts of their owners. These huts were hemispherical structures made by planting a ring of wands in the ground and bending the tops over, tying them together to form a dome that was covered with woven grass matting. Weatherproof and cool, the huts could be assembled or taken down in a few hours, loaded onto oxen, and carried to the next settlement site. The oxen were used as mounts as well as pack animals.

The hut was the property and domain of the wife; she had complete authority over it, to the extent of being able to exclude her husband if she wished. It appears that the Khoikhoi woman's status was equal or nearly so to that of the man; women could own stock and were economically important in their role as gatherers of plant food. The men were herders and hunters and, among some tribes, metalworkers. The principal weapons were arrows and spears tipped with bone or iron. An extensive trading network seems to have existed through which scattered settlements exchanged aromatic, medicinal, and narcotic herbs; pottery; and metalware such as iron and copper beads, arrowheads, and knife and spear blades. Some Khoikhoin, particularly in northeastern Botswana, had developed rather advanced mining techniques; the old shafts indicate a high order of geologic insight.

Religion of the Khoikhoin. The Khoi deity Tsuni-g//oab is celebrated in myths as the first of that people from which all others take their origin. A great chief and a powerful magician, he worked many wonders, including the creation of men and women. He died several times and was resurrected, eventually becoming a purely spiritual being. As such he was omniscient and omnipresent and the giver of life and of rain. Supplication was made to him in an elaborate ritual to persuade him to send the rain that was essential for survival of man and stock.

Another deity, Heitsi-eibib, was the hero of a cycle of myth. He could change his form and had died and been reborn many times. He was worshipped at his many graves scattered about Khoi country; the passerby would add a stone to the heap marking the grave and pray to Heitsi-eibib for prosperity and success.

G//aunab was a god with varying attributes who, with appropriate phonemic changes of his name to fit a particular language, appeared in the theologies of a number of Khoisan peoples. In the myths of the Khoikhoin he is the adversary of Tsuni-g//oab and the bringer of death, disease, and other misfortunes. In some accounts he appears as the personification of the malevolent spirits of the dead. Among those peoples exposed to Christianity, he has been equated with the devil.

The Khoikhoin are also said to have worshipped the Moon, although not much is known about this. Certainly the Moon figures in their very rich mythology.

The practice of magic does not seem to have been extensive. Curing and divination were the main activities of magicians, although they occasionally misused their powers for evil purposes. When they did, their influence could be negated by the application of cold water, either to the sorcerer or to his victim—an illustration of the high value placed on water by the Khoikhoin.

Khoi confederation of clans

Political organization. Khoi communities were linked by language and kinship into a loose confederation of clans that formed a nation. Several nations existed in the 17th century. Each had its own territory, its own name, and apparently its own dialect. (The dialectal differences must have been minor, as there was great similarity in the Khoi languages across their total range of nearly 2,000 miles from the eastern Cape Province in South Africa to the highlands of Namibia.) The clans comprising a nation were ranked in order of seniority, the chief of the nation being drawn from the senior clan. Marriage within the clan was strictly forbidden; the need to find spouses in other clans helped to keep the nation together. The nation as a whole had exclusive rights to grazing and hunting land, water holes, and mineral deposits. In the larger nations there seems to have been some apportionment of land among the clans.

Leadership, despite its clan basis, was not strong; it was persuasive rather than autocratic. Many instances are recorded of individual clans hiving off from the main body, attracting a following from other clans, and setting up as an independent nation with a new rank order of constituent clans. It has been suggested that this anarchistic tendency was reinforced by social and environmental pressures. Among a people lacking a strong and coercive political system and totally dependent on a scarce resource such as water, the use of that resource can be effectively controlled only if harmony prevails in the community. Harmony is threatened if the community reaches an unwieldy size and imposes too heavy a demand on the limited supply of water. The only solution then is for the community to divide. The "daughter" component must put sufficient distance between itself and the "parent" community to avoid predation by the latter. The effect is to limit the size of communities and to encourage their wide dispersal over habitable country.

Khoi nations did not develop great strength or solidarity. As nomadic pastoralists, they were forced to give ground before the incursions of more cohesively organized peoples: the Bantu who moved down the southeastern coastal areas (probably about the 10th century AD) and European settlers who began moving inland from Table Bay in the 17th century. By the 19th century some of the Khoi nations felt themselves desperately pressed. Strong leaders were able to unite fragments of clans and nations under them, and large groups moved across the Orange River out of Cape Colony into South West Africa, thus retracing the route that may have been taken by forebears, who are thought to have migrated into southern Africa from the central African lakes. In South West Africa some of the Khoikhoin had troubles with the Herero group of the Bantu, who were roving farther and farther south. Caught between the Europeans and the Bantu, the Khoikhoin fought for nearly 100 years, giving up the struggle only after the suppression of the uprising of the Bondelswarts in 1922.

Remnants of the Khoi culture

Little now remains of a distinctively Khoi culture, despite the efforts of missionaries and governments to soften the impact of alien cultures. The Nama and Garikwe communities retain their identities as Khoikhoin, and the former have kept their language and some of their customs. Many other communities live on reserves, where they make a living from their livestock, supplemented by jobs as unskilled or semiskilled labourers, but theirs is more the culture of poverty than that of the traditional Khoi people.

THE SAN (BUSHMEN)

The term Bushman is an Anglicization of the 17th-century Dutch *bosjesmans* ("bush dwellers"; *i.e.,* people without fixed abode). The historical evidence concerning San is incomplete and often unreliable. As hunters and gatherers, they were unobtrusive in their occupation of the land. Few early explorers took much interest in them, and fewer had the time and means to study the small, widely dispersed communities that shifted their encampments every few weeks. The San occupied most of the country south of the Zambezi River that was not in the hands of the Khoikhoin or Bantu. Skeletal and cultural remains, particularly rock paintings and engravings, indicate that they once ranged over the whole subcontinent.

Present-day San. The San population lives in Botswana (chiefly in the western Kalahari) and Namibia, and a small number in Angola. Some of the San move among Zambia, Zimbabwe, and Botswana. Two groups lived in the Republic of South Africa, one in the Kalahari Gemsbok National Park in Cape Province and the other, a very small community, near Lake Chrissie in the eastern Transvaal.

It is often stated that the surviving San migrated to their present inhospitable desert habitats in recent times after being driven out of other territories. There is, however, ample evidence that they occupied their present locations in the mid-19th century, when other San were still living in parts of southern Africa where they are now extinct. In addition, there is no evidence that the latter groups migrated into the desert areas; rather, it is clear that they remained where they were until they were either overwhelmed by war and disease or absorbed by intermixing with Europeans and Bantu. The distribution of dialects among the present-day San population is consistent with their having lived in their present locations over a very long period.

The majority of today's San have been dispossessed of their territories by European and Bantu cattle raisers. The intrusion of cattlemen has reduced the supply of game animals and esculent plants that constituted the greater part of their diet. Most San have therefore turned to livestock raising for their living, either by hiring themselves as labourers to ranchers or by entering into master–serf relationships with Bantu owners of cattle posts. A few of them still follow the traditional hunter-gatherer style of life.

Ranch labourers. On the marginal cattle ranches in the arid regions of southern Africa, San compete with the Bantu for available jobs. Because of their relative lack of strength and marketable skills and because most do not know any language other than their own, they get only the lowest-paid jobs. In addition, a person who finds employment often attracts a number of dependent kin and friends who come and live with him on the ranch. It is common, therefore, for ranchers to increase the labourer's issue of rations (the usual supplement to wages) to accommodate the train of dependents. Some ranchers give an annual bonus of small livestock, which are allowed to be kept on the ranch and multiply.

Ranch San, although retaining their languages and some aspects of social organization such as the kinship structure, have lost much of the rationale of their old culture.

In the traditional life-style, women had great economic importance as the gatherers of food plants, which constituted the main part of the diet. On the ranches, however, a labourer's wife may be something of a liability when he is competing with other men for a job. Deserted wives and children are numerous; they turn to already overburdened kinsmen for help or to prostitution and petty crime.

Hunting-gathering San. Independent San are confined mainly to the western part of Botswana. In the northwest are the !Kung (! represents a click sound), who inhabit an area with some permanent water holes and numerous esculent plants, including extensive groves of mugongo trees (*Ricinodendron rautanenii*), which bear pleasant-tasting and nutritious nuts. The !xõ (!xu) live in the west central part of Botswana, a region without natural supplies of permanent water but rich in game and esculent plants. In the central Kalahari region of Botswana are the G/wi, /aba, and G//ana groups. This region does not have the seasonal abundance of game that is seen farther west and is also waterless except during the wet season. Other hunter-gatherer San are found elsewhere in Botswana, Namibia, and Angola, but they are not independent and autonomous, and their way of life is therefore less representative of the traditional San cultures.

ıltural
riety
San

San are not culturally homogeneous, although the cultures of some of the groups resemble one another. The !Kung, !xõ, and G/wi, though all hunter-gatherers, live in rather different habitats containing different resources, and the three groups show marked differences in their technologies and their patterns of exploiting resources. Organizationally, these groups also exhibit marked contrasts. The lexical and structural differences among their languages are so great that they have been classed in three different language families. Generalizations about the San can therefore be made only at the level of a "southern African hunter-gatherer culture" and limited to such superficial common features as the use of bows and poisoned arrows in hunting; a low population density and small communities; generally diffuse and ephemeral leadership; egalitarian communities, the government of which is by consensus; an absence of the lineage principle in most, if not all, kinship systems; and a lack of formal legal and judicial systems. Both the breaking up and coalescence of communities are common. Violence and warfare, although frequently reported in the literature of now-extinct San, are not characteristic of the present-day hunting-gathering San peoples; on the contrary, their ethics stress cooperation. A detailed description of one of the San peoples follows.

The G/wi people of the Kalahari. The G/wi inhabit the western half of the Central Kalahari Game Reserve and the eastern fringe of the Ghanzi ranches. Fewer than 1,000 G/wi live on the ranches as labourers and their dependents, and another few hundred move into the ranching area during periods of drought. Sporadic trading visits are made at other times in order to exchange the desert produce of game-hide leather for tobacco and iron rods as well as for fencing wire, which is used in the manufacture of tools and weapons. The remaining G/wi people, however, are permanent desert dwellers who have only sporadic contact with Bantu and Europeans and have retained their traditional hunter-gatherer culture.

viron-
nt of
G/wi

The habitat. The G/wi country lies in the centre of the Kalahari Basin, a depression in the vast inland plateau of southern Africa. The basin is covered by a mantle of fine-grained sand, up to 400 feet deep, poor in plant nutrients and of a texture that renders it susceptible to wind erosion, leaching, and scorching. Rainfall is confined to 10 or 12 weeks after midsummer. The country responds with startling rapidity to the first good rains, and within four days the landscape is transformed by a luxuriant growth of grasses and herbs and by the flush of foliage on shrubs and trees.

Rain is the key to all life in the Kalahari. The extreme seasonal variations in rainfall, humidity, and temperature lead to corresponding seasonal contrasts in the amounts of plant and animal food available to hunter-gatherers. There are a few localities in which resources are sufficient to meet the year-round needs of the hunters and gatherers, and these are the areas that the G/wi communities occupy as their territories.

The G/wi world view. In G/wi theology the universe was created by N!adima. He is the owner of all that is and may dispose of it as he wishes. His actions are, however, bound by the natural systems he is believed to have ordained. Therefore, he cannot suspend or reverse their operation.

N!adima is remote from his creatures. They cannot communicate with him or influence his will to their own advantage or favour. G/wi therefore have no religious rites such as prayer, worship, or sacrifice and no priests. They stoically accept N!adima's occasional caprice—*e.g.,* when "he grows tired of a man's face" and sends a marauding lion or some other misfortune to kill him.

G//amama is a less powerful being than N!adima. He sporadically attempts to do harm to man and is often successful, but he can be frustrated by a number of means that N!adima created and man has been able to discover. Several herbal medicines have the power to counter G//amama's influence; certain communal dances also serve to exorcize the evil he is believed to send in the form of invisible, magical slivers of wood.

The logic of the G/wi worldview is that N!adima, having created man, intends him to survive and thus permits him to make use of what is available in the environment to that end, subject to the restriction that N!adima will be angered by wastefulness or greed on the part of those using his property. Man, at least G/wi man, must devise for himself the best means of survival that he can, including the regulation of interpersonal behaviour. Social usages and customs are therefore seen as man-made and not sacrosanct. The stability of behaviour and the predictability of others' actions require agreement by all concerned. Such agreement can be won only by persuasion and ensured by the fulfillment of obligation. It is therefore imperative that individuals and the community remain on good terms with everybody in the circle of social contact.

Social and political organization. The communities, or bands, usually have between 40 and 60 members, the extremes being 25 and 85. They are open communities, in the sense that there are no particular qualifications for membership, but members tend to stay with a group over long periods. Eventually, some catastrophe such as severe drought or epidemic disease (*e.g.,* the smallpox outbreak of 1950) may lead some bands to merge, in order to ensure their members' survival.

G/wi
kinship
and social
structure

The G/wi kinship system is partially classificatory (*i.e.,* it subsumes several different biological relationships under a smaller number of categories). There is also a high degree of equivalence among those classified by the same kinship term. These features allow the kinship system to be applied to persons other than actual kin, including the whole membership of the band. The kinship system can thus be used as the organizing principle of society and, by further extension, to govern relationships between people in different bands.

The kinship system permits not only extension but also truncation: groups can be enlarged by inclusion of new members and also easily split when necessary by substituting other members for those who have left. This is of great advantage to the G/wi. Not only are their bands subject to periodic disasters, but they also grow; in some cases they outgrow their territorial resources and need to be able to split up. The bands normally respond to the decreased density of food resources in winter and summer by reducing their population density. This is achieved by fragmenting into constituent households, each of which goes off into a separate part of the territory and remains there until the approach of the wet season, when food supplies increase again. The organization of the band must therefore compromise between making the community cohesive enough to hold the members together and keeping it loose enough to allow them to separate when necessity dictates.

The kinship system also provides sets of rules for appropriate behaviour. All kin of any individual are either his avoidance or his joking relatives. Behaviour between avoidance relatives is reserved and polite. The avoidance category includes those who are in a relationship involving submission to or exercise of authority—firstly, the parent–child relationship and, by extension, relationships between all those who are classificatorily equated with parents and children. Siblings of opposite sex are in the avoidance category and, by extension, all those whose classificatory relationship prohibits intermarriage. All other kin are joking relatives, among whom behaviour is much less restrained: possessions are freely shared; ribald, bawdy joking is in order; and pointed but friendly criticism of behaviour is permitted in public without offense.

The regular retreat into isolation requires that each household be capable of independent existence for three or four months of the year. This precludes the development of any centralized authoritarian political system either within the band or among bands. Leadership is diffuse and ephemeral, arising in response to specific situations. The function of a leader is to make proposals for action that are acceptable to the members, and he has little power to force agreement. Because of its lack of firm leadership and its equalitarianism, the band would have only tenuous solidarity were it not for the emphasis placed upon harmonious relationships and the fact that social relationships are structured so as virtually to exclude competition.

The G/wi economy. About 30 plant species provide the basis of G/wi subsistence. None is available the year round, the number available at one time varying from 29 in autumn to only four in early summer—of which only two are sufficiently plentiful to make a significant contribution to the diet. The work required in gathering food for a household therefore varies greatly with the season; the more species there are to be taken, the easier and shorter is the search and the greater the number of people able to subsist on the resources of a given area. It is for this reason that the band separates in winter, when the frosts blight the plants and sharply reduce the amount of plant food available. Dispersed in individual households, each exploiting a particular area, the relative load on the food resources is reduced, as is also the amount of work required to gather enough to eat. Just before the onset of the wet season the variety and amount of food plants increase; the band can again reform and resume its normal life as one community, shifting its encampment every three to five weeks to a new supply of plant food.

Means of subsistence

The work of gathering is done mainly by the women and girls. They range over a radius of up to five miles, picking fruits and berries and digging the roots and tubers that comprise about 80 percent of the G/wi diet. A three-foot stick sharpened at one end serves as digging tool and carver. An antelope-hide cloak, worn over the shoulders and fastened around the neck, does duty as a blanket at night. The lower points of the cloak are tied about the waist, forming a bag.

Hunting and trapping is the work of men. They use a light bow and flimsy, unfeathered arrows tipped with poison prepared from the pupae of the beetle *Diamphidia simplex.* Arrowheads were formerly of bone, but most are now made of fence wire hammered by stone into precise shape; the poison is applied just behind the cusps of the arrowhead. Careful stalking is necessary if the hunter is to get close enough to the antelope and small desert mammals to place his arrow effectively and inject a dose of poison. The poison requires between six and 96 hours to take effect, depending on the animal, during which time the animal must be tracked and followed; when killed, its meat has to be carried many miles back to the encampment.

The beetle grubs for making poison can be gathered only in summer, and, since by the middle of winter supplies are usually exhausted, recourse must be had to traps. These are snares, loaded by attaching the free end of the line to a bent-over sapling. The noose is held in place by a trigger, which, when stepped upon, releases the sapling and causes the noose to tighten around the leg of the antelope and hold it until the trapper returns, which he does about four times a day. The monthly intake of meat per person varies from about four pounds (two kilograms) in October, the end of winter, to 15 pounds (seven kilograms) in January, when fresh arrow poison is available and the antelope population is increased by the arrival of migrating herds.

The G/wi have drinking water only for about eight weeks of the year. They meet their fluid requirements at other times from the juices of plants and by drinking the blood and rumen liquor (squeezed-out cud) of antelope; they also drink the amniotic fluid of pregnant does.

Technology

Tools are light and multipurpose. The desert does not provide many raw materials, and migrations between campsites impose a need for easy portability. Iron rods obtained from the ranches are beaten out on stone anvils to make the heads and blades of spears, knives, adzes, and axes. Pestles and mortars are adzed out of wood. Bowls and other receptacles are also carved from wood. Tin cans and enamelware are imported and used as cooking pots, replacing the fragile pottery that was formerly obtained from the Boteti (Botletle) River area and has now become rare. Antelope hides, pegged out to dry in the sun, are scraped clean of fat and other waste and then tanned, using a mixture of plant juices, bone marrow, urine, and rotted brains. The leather of the steenbok and duiker is worked into a fine kid having the softness of cloth and is one of the main commodities of exchange with the bands nearer civilization. Other antelope hides are made into cloaks and skirts or sewn to make bags or sandals.

Shelters vary with the season. The dry-season structure is a roofless windbreak of shrubs, branches, and grass. The wet-season hut is of a beehive shape, made from a frame of branches thatched with clumps of grass. When a band migrates, the huts or shelters are abandoned and new ones set up within several hours after arrival at a new campsite.

(G.Si.)

HISTORY

Southern Africa to 1800

EARLY MAN AND THE STONE AGE

The story of mankind in southern Africa may go back 2,000,000–3,000,000 years. Hundreds of fossils and artifacts found in the dolomitic limestone caves of the Transvaal and northern Cape suggest that for hundreds of thousands of years southern Africa was in the forefront of human development and technological innovation. With East Africa, the subcontinent has seen the evolution of hominids into the beings who, deliberately fashioning tools to a regular and set purpose, are commonly defined as human.

The hominid way of life

For a long time small groups of manlike creatures roamed the open savanna lands of southern and Central Africa, collecting fruits and berries—and perhaps roots—and either scavenging or catching small animals to supplement their diet. During the Early Stone Age the hominids, using first pebble tools and, later, all-purpose stone axes and natural traps, gradually became more skilled as hunters. About 200,000 years ago, with the discovery of the controlled use of fire, the hunting bands were able to make their camps not only in the watercourses in the dry savanna and coastal areas to which they had been previously restricted but also in rock shelters and in the more heavily wooded country of what are now Zambia and Angola.

After hundreds of thousands of years of relatively undifferentiated tools came the experimentation, regional development, and diversification of the Middle Stone Age, dated in what is now South Africa to between 120,000 and 40,000 years ago, with regional variations. Larger, more mobile populations with more varied settlement and subsistence patterns were able to exploit their environment more effectively than before. Over most of the area in this period, the large, crude Acheulian hand axes of the Early Stone Age were replaced by more diverse tools, made by striking flakes off a core into flake-blades and flaked spear points, crescents, and scrapers. Specialized adzes, gouges, and machetes were developed in the forest zone of the Zambezi and Zaire river valleys.

The ever-increasing number of radiocarbon dates available for the Late Stone Age in southern Africa suggest that its new small-blade technology developed from ancestral Middle Stone Age industries between 40,000 and 15,000 BC. In Zambia bored stones for weighted digging sticks and grinding stones suggest a greater emphasis on food gathering, while faunal remains everywhere show that there was a move to the hunting of smaller woodland animals rather than the exploitation of the large gregarious creatures of Middle Stone Age times.

In what is now South Africa, a number of distinctive Late Stone Age industries have been identified. Initially, these industries were based on macrolithic scrapers and the hunting of larger herds of antelope and separated the microlithic cultures of the Middle Stone Age from the widespread microlithic Late Stone Age Wilton industry. From the 7th millennium BC the Wilton industrial complex emerged over a large part of southern Africa and was characterized by the predominance of small convex scrapers and backed microliths. Despite regional variation, it seems to represent a single culture, reflecting changing fashions and local adaptations rather than any cultural or ethnic distinctions among the toolmakers. It was followed during the 1st millennium BC or perhaps a little earlier by another phase of a large scraper-based industry in many parts of what is now Cape Province, Namibia, and perhaps Zimbabwe as well.

Though there is considerable dispute and terminological confusion over the nature of the Middle and Late Stone Age skeletal remains, the Late Stone Age populations of southern Africa encountered in historic times—the Pygmy, San, and Khoikhoi peoples—were basically genetic developments of an ancient African stock that had evolved in the subcontinent. Khoisan remains are associated with the Wilton culture over much of southern and eastern Africa. There is some slight evidence also of a small non-Khoisan population, which may be associated with the stone bowl cultures of East Africa and possibly with the origins of sheep pastoralism in the Cape.

During the Late Stone Age hunter-gatherers appear to have been organized in loosely knit bands, of which the family was the basic unit. By using bows and arrows, often tipped with poison, and a variety of snares and traps, they achieved greater dominance over their environment. With hooks, barbed spears, and wicker baskets, they were able to catch fish and exploit rivers, lakeshores, and seacoasts more effectively. Each group had its own territory, in which special importance was attached to natural resources, and in many instances bands moved from summer to winter camping sites, following the game and vegetation, as they were to do in historic times. It seems likely, too, that the sexual division of labour was much the same as in historic times, with men responsible for hunting, women for collecting plant foods and for domestic chores.

one
ge art

Associated with the Late Stone Age are magnificent cave paintings and engravings on rocky outcrops. Although most of the surviving rock paintings are probably comparatively recent, in Namibia some have been dated to a pre-Wilton industry, while on the southern coast similar paintings are associated with Wilton remains dated to the 6th and 4th millennia BC. Whereas the art in the northern woodlands is stylized and schematic, that of the savanna and coastlands is naturalistic, showing scenes of hunting and fishing and of ritual and celebration. An analysis of the symbolic structure of the paintings suggests their deep religious significance; they also vividly portray the way of life of the Late Stone Age. In the rock art also is the first hint of the new groups of herders and farmers who were to revolutionize the economies and societies of southern Africa.

At a number of sites along the Cape coast and the Orange River, dating from about 2,000 years ago, there is clear evidence of the presence of sheep associated with distinctive pointed base, lugged pottery. Although as yet no cattle bones have been discovered so early, one of the outstanding problems for Late Stone Age research is to elucidate the spread of cattle and sheep pastoralism among the Khoisan.

THE COMING OF THE IRON AGE

The coming of the Iron Age to southern Africa almost 2,000 years ago brought with it the food-producing revolution. Agriculture combined with pastoralism supported far larger settled communities than had been possible and enabled more complex social and political organization to develop; it is with the evolution of these Iron Age farming communities into present-day African societies that the precolonial history of the region is concerned.

The Early Iron Age. Though documentary sources and oral tradition go back at most only 500 years, archaeological, linguistic, anthropological, and ethnobotanical data suggest that the complex of Early Iron Age traits—agriculture, settled village life, ironworking, and pottery—entered the subcontinent in fully developed form in the first half of the 1st millennium AD. The 3rd- to 5th-century-AD dates for Early Iron Age sites deep into South Africa are remarkably close to the first Early Iron Age dates to the north and suggest the rapid spread of the new culture over a very wide area. Although the language of these early farmers is unknown, the chronology and scatter of the Iron Age sites is consistent with what is known of the spread of the closely related Bantu languages spoken in the region today. Thus these Early Iron Age sites probably represent the first settlements of Bantu speakers, the precursors of the majority of the region's contemporary population. (As noted above, despite its widespread misuse, the term Bantu refers solely to this language family.)

Expansion
of Iron Age
culture

This movement into southern Africa was part of a wider expansion of Iron Age culture. Penetrating to the woodlands south of the moist forest zone of Central Africa, Early Iron Age farmers settled in substantial villages in the savanna grasslands, along rivers and watercourses, searching out richer soils. Their economy was based on the cultivation of millet and sorghum, and they had herds of sheep and goats and small numbers of cattle. Thus they avoided both the dangerous tsetse fly belts on the east, and the more arid western half of the subcontinent, which was suited only to a specialized pastoral way of life. There the first Iron Age dates are later than those of the better watered easterly regions.

Most of the Early Iron Age communities in the subcontinent shared a common, if regionally diversified, culture, a phenomenon suggesting the rapid dispersion and proliferation of small trickles of people. As they spread they absorbed the sparser Stone Age peoples, leaving small remnant groups in the southern and western fringes of the subcontinent. In some areas hunter-gatherers may well have adopted the food-producing techniques and even the language and culture of the newcomers. Thus, today the Bantu-speaking peoples of southern Africa are neither racially nor genetically uniform. The old picture of invading hordes of Bantu speakers moving into southern Africa, brandishing iron weapons, and wiping out the earlier inhabitants is not supported by the evidence. In South Africa the close relationship between hunter-gatherers and farmers is manifest at a number of archaeological sites, where Iron- and Stone-using peoples seem to have lived together, and by the presence of "click" sounds from the Khoisan languages in the South Eastern Bantu languages.

From early times Iron Age farmers made use of metals, settled near mineral as well as salt deposits, and engaged in trade, often with Late Stone Age people. Iron implements, such as hoes, and possibly pottery as well, passed from hand to hand, as some communities became known for their specialist craftsmen or settled near exceptionally good iron or clay deposits.

The Later Iron Age. From about the turn of the 1st millennium AD, in some areas of what are now central Zambia, southeastern Zimbabwe, and parts of Malaŵi, the Transvaal, and Natal, changes in ceramic style were paralleled by a change in the location and nature of settlements and a far greater emphasis on cattle keeping and milking. More sophisticated techniques of metalworking, more extensive gold and copper working (perhaps in response to the development of trade with the east coast), and a great extension in stone building suggest the evolution of more complex state structures, increased social differentiation, and new religious and ritual ideas. These changes were, however, neither simultaneous nor evenly spread.

The transition from the Early to the Late Iron Age is still poorly understood. Although many scholars are convinced that the transition is marked by a sharp break, best explained by new arrivals, research on sites in central Botswana and elsewhere suggests that the natural buildup of herds in the Early Iron Age may have provided the catalyst for change.

Importance of cattle

Growing numbers of cattle allowed for the accumulation of wealth in societies where there were few other ways of storing a surplus, and this, in turn, may have led to increased stratification and social control. Historically, cattle have been the main form of bridewealth in much of southern Africa, so that control over cattle has meant control over women and their productive and reproductive capacities. More wives, and thus more children, increased a homestead's ability to expand its agricultural production and its following: this was particularly important in societies where wealth and power were measured in terms not of land but of people. Patrilineal and polygynous cattle keepers had immense advantages over Late Stone Age hunter-herders and over Early Iron Age farmers without the new forms of wealth and social organization.

Greater stratification and more complex social organization were also probably accelerated by the growth of long-distance trading. In the early centuries AD the northeast African coast was well known to the traders of the Greco-Roman world. These contacts diminished with the rise of Islām, and the east coast became part of the great Indian Ocean trading network. By the 8th century Arab traders began to visit more southerly harbours, and between the 11th and 15th centuries they founded some 37 new towns. Though these towns never united politically, they developed a common Afro-Arab, or Swahili, culture and a splendour that amazed the Portuguese when they encountered them at the end of the 15th century.

The Limpopo and Save rivers appear to have been early arteries of the trade to the southernmost Arab trading posts and settlements, with African intermediaries initially bringing ivory, and later copper and gold, to the coast.

Prehistoric mining

From the late 1st millennium AD, the mining of copper and gold had become part of Iron Age culture. In Zimbabwe alone there are more than 1,100 prehistoric gold mines and 150 copper mines, while in the Transvaal there are thousands of iron, copper, and gold mines. While most of these date from the 2nd millennium AD, at Phalaborwa in the eastern Transvaal iron was being mined by the 8th century, and a little later the area became a rich source of copper.

The move from ivory to gold and copper trading is perhaps reflected in the archaeological record with the shift in power from sites like Schroda and Bambandyanelo in the Limpopo Valley, where ivory seems to have been of major importance, to the most prosperous of the Late Iron Age sites known in southern Africa—Mapungubwe, Great Zimbabwe, and Ingombe Ilede, whose prosperity was based on the mining of gold and copper. Although they do not typify the Iron Age as a whole, these sites share certain broad cultural similarities. They appear to be the settlements of a wealthy and privileged trading elite who built with stone and could afford to be buried with gold and copper ornaments, exotic beads, fine Chinese pottery, and imported cloth. Their homes, diet, and ostentatious burials are in stark contrast to those of the common folk, whose dwellings cluster at the foot of the sites where they probably laboured.

Mapungubwe. Located atop a steep hill in the mid-Limpopo Valley in the northern Transvaal, Mapungubwe flourished in the 11th and 12th centuries AD—rather earlier than the more famous Great Zimbabwe, from which site it was long considered a subordinate offshoot. The walling could only have been built by bringing large quantities of stone to the top of the hill, which suggests the organization of considerable labour, while gold and copper came in trade or tribute from settlements about 100 miles away. Mapungubwe therefore represents a far more powerful class-based state than anything witnessed in the area before. Fierce controversy has raged around the racial identity of its occupants. Early excavators refused to accept that it could have been built by blacks. Subsequent work on the Iron Age in southern Africa, however, puts both the skeletal and cultural evidence in the same category as that from other Late Iron Age settlements on the subcontinent, and there is little reason to doubt its association with Bantu speakers.

Mapungubwe's heyday lasted until 1150, and it had all but disappeared 50 years later. By the early 1980s there had been no equivalent site excavated in South Africa, although it would seem that Manekweni in southwestern Mozambique, dated to the 12th and 13th centuries, and Great Zimbabwe, which flourished in the 14th and 15th centuries, represent similar accretions of power based on wealth in cattle, which in turn enabled a preferential access to the coastal trade.

Great Zimbabwe. With 60 acres (24 hectares) of imposing ruins, Great Zimbabwe has dominated the discussion of the Late Iron Age of southern Africa. Long the subject of mystery and romance, the Zimbabwe complex—Great Zimbabwe is the largest and most impressive of more than 150 similar ruins extending from southwestern Zimbabwe into Mozambique—has been clearly established as being of African origin and of medieval date. From about 1300 to 1450 prosperity at Great Zimbabwe led to the import of fine cloth and celadon china, Indian beads, and elegant ornaments. Although the names of the rulers of Great Zimbabwe, the precise extent of their power, and the identity of their people are unknown, they probably were, like the contemporary majority of Zimbabwe's population, speakers of the Shona language.

In the second half of the 15th century the kingdom came to an abrupt end. Its successor in the southwest was the Torwa state, with its centre at the stone-built site of Khami. The new culture that emerged at Khami developed both the stone-building techniques and the pottery styles found at Great Zimbabwe and seeded a number of smaller Khami-type sites over a wide region. The Torwa kingdom seems to have lasted until the end of the 17th century, when it was replaced by the Changamire dynasty of the Rozwi from the central plateau.

Ingombe Ilede. The decline of Great Zimbabwe may also have been associated with the rise of a new trading centre at Ingombe Ilede, in Zambia. It was located in a strategic position on the Urungwe plateau just to the north of the middle Zambezi, and its chief export appears to have been copper, which the local people wrought into finely made crosses and traded over a wide distance. Neither the local people nor their rulers were likely to have been Shona, but they probably drew their supplies of gold from south of the river, and may thus have limited the expansion of Great Zimbabwe.

Other kingdoms. During the heyday of Great Zimbabwe a considerable number of similar settlements on a smaller scale were built by expanding dynasties, probably related to the Karanga ruling group at Great Zimbabwe. As wealthy, polygynous cattle owners, they would have had both the means to expand at the expense of surrounding people and the impetus to send their scions in search of fresh pastures. Between the 13th and 16th centuries dynasties bearing Zimbabwe-type culture expanded in every direction, absorbing earlier non-Shona-speaking peoples. Of these states, most is known about the kingdom of Mutapa, and its offshoots in the eastern highlands of Zimbabwe, because of their encounter with the Portuguese.

Farming communities. If the stone buildings of Great Zimbabwe and its satellites are readily visible remains of state formation in southern Africa, the majority of farmers in southern Africa in the first half of the Late Iron Age were relatively unaffected by long-distance trade and the formation of larger states. Most Iron Age farmers lived in small-scale societies, based on kinship, in which political authority was exercised by a chief who claimed seniority by virtue of his royal genealogy, but who may have risen to power through his access to mineral resources, hunting, or ritual skills. By 1500, Iron Age farming communities had stabilized in roughly their present-day habitats, reaching their ecological frontier on the southern Highveld of South Africa, where rainfall averages 16 to 20 inches annually, and gradually clearing the forests of the southeast coast.

By the time the peoples of the coastlands were encountered by the first literate observers in the late 15th and early 16th centuries, they had acquired part of their contemporary cultural and ethnic identity. This does not mean, however, that these societies were static and unchanging. Within these frontiers there was considerable movement as population expanded and found available resources inadequate. Thus, between the 17th and 19th centuries there was migration of northern and eastern Shona speakers into the centre and south of the plateau, while in South Africa, too, there was colonization of new land by people and cattle. In some areas this inevitably led to conflict as the newcomers came up against settled communities; in others the slow process of absorbing Late Stone Age people was completed, while in yet others the colder, more arid mountain lands were colonized.

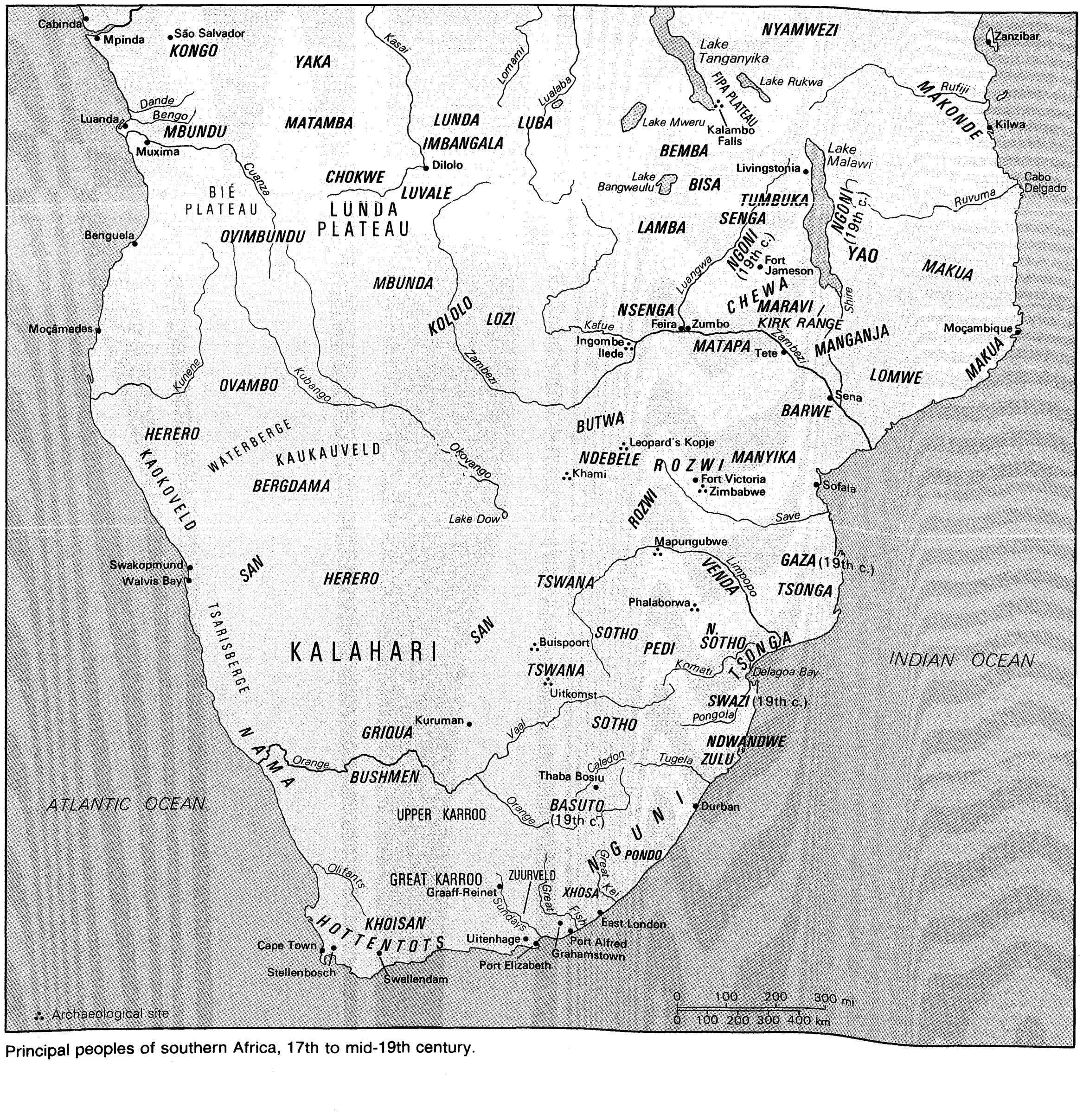
THE PORTUGUESE

In the hope of outflanking Islām, finding a sea route to the riches of India, and discovering additional sources of food, the Portuguese edged their way around the African coast from the 15th century onward. They reached the kingdom of the Kongo, to the north of present-day Angola, in 1482–83; early in 1488 Bartolomeu Dias rounded the tip of the continent; and just over a decade later Vasco da Gama sailed along the east coast of Africa before striking out to India. Within eight years of their arrival, the Portuguese had seized the divided if wealthy centres along the east coast, and by 1507 they had built a fortress and refreshment station on Mozambique Island, which soon became the key port of call for ships on the way to India.

The Portuguese in southeastern Africa. The Portuguese conquests led to the economic and cultural decline of the east coast cities. The Portuguese soon discovered that they were unable to control the vast area they had conquered. They faced resistance from the Swahili communities on the coast throughout the 16th century, and the profits they had expected from the gold trade failed to materialize. In an attempt to control the trade and to discover the precious minerals for themselves, the Portuguese, following in the tracks of Muslim traders from the coast, expanded into the Zambezi Valley. There they penetrated the Mutapa kingdom, with its heartland in the northeast between the Zambezi and Mazoe rivers.

Contrary to earlier historical opinion, there is little evidence to link the origins of Mutapa directly to Great Zimbabwe, nor did it reach the size suggested in some accounts. It was, nevertheless, a considerable kingdom by the beginning of the 16th century, the capital alone containing several thousand people. As at Great Zimbabwe, the gold trade was of great importance, although subsistence depended on agriculture and cattle. The Mutapa kingdom

In the 1530s the Portuguese dominated the trade exits from the coast and had established fortresses at Sena and Tete and trade fairs along the Zambezi River and on the plateau, where Africans came to exchange ivory and gold for beads and cloth. After 1541 Portuguese residents at



Principal peoples of southern Africa, 17th to mid-19th century.

The *prazo* warlords

these outposts elected representatives who were delegated certain powers by the Mwene (or ruler of) Mutapa. Individual Portuguese and Goans were also able to get land grants and judicial rights from local rulers, which enabled them to extract tribute from the local population. Out of these early grants grew the *prazo* system of landholding. Between the 17th and 19th centuries the *prazo* holders became immensely powerful and were able to interfere actively in local African politics and to resist attempts by the Portuguese crown at reform. By then an Afro-Portuguese society had been created in the lower Zambezi Valley independent of either African or Portuguese jurisdiction. Assisted by African and Portuguese soldiers known as the Chikunda, Afro-Portuguese warlords engaged in the slave and ivory trade, unsettling a wide area of east central Africa.

The effect on the Mutapa state of Portuguese traders along the Zambezi Valley was minimal until the late 16th century. The hold of the Portuguese over the Mutapa kingdom was strengthened in part, however, by the sudden irruption of Maravi peoples from north of the Zambezi. They were known as the Zimba, a term applied to any marauding group, and their migrations were probably set off by rivalries among chiefs in southern Malaŵi. They appeared along the Zambezi in the 1580s, attacking Sena and Tete between 1592 and 1593 and then the Portuguese settlements along the coast. By 1601 the Mwene Mutapa was forced to call on the Portuguese, while his attempts to expel the invaders also triggered a widespread and debilitating internal revolt.

In return for their assistance the Portuguese tried to control the Mwene Mutapa, and this proved increasingly disruptive. After an unsuccessful attempt by the reigning Mwene Mutapa to drive them from his domains in 1629, he was replaced by a rival from whom the Portuguese extracted extensive trading and mining privileges. Yet this appearance of Portuguese power and prestige was deceptive; the Portuguese never had the resources to control the interior. It was the Afro-Portuguese *prazo* holders who truly exploited the weakness of the Mwene Mutapa.

The Rozwi kingdom

The rise of the Rozwi kingdom under the Changamire dynasty also undermined the Portuguese presence in Mashonaland. By the late 17th century the Changamire, originally subject to the Mutapa, had expanded over much of central and southwestern Zimbabwe, and may even have had an eye on the Mutapa throne. In response to an appeal from the Mwene Mutapa, who was facing revived Portuguese pressure, the Rozwi swept up the Zambezi between 1693 and 1695 and forced a Portuguese retreat. Despite this, however, Rozwi intervention could not save the Mwene Mutapa from the civil wars that ensued and reduced the Mutapa state to a small chiefdom in the Zambezi Valley lowlands, where it maintained itself into the second half of the 19th century. Elsewhere, the hegemony of the Rozwi was probably recognized over a wide area. Heading a confederacy, which left a great deal of autonomy in the hands of local dynasties, the Changamire had both religious prestige and military power. Wealth was based on the agricultural and pastoral resources of the kingdom, internal tribute, and trade with African neighbours and coastal Arabs, Indians, and Portuguese.

The Maravi confederacy

In addition to gold, the Portuguese were interested in ivory and other mineral resources, particularly after 1700 when the goldfields on the plateau were exhausted. A search for silver mines first led them into Malaŵi in the 17th century, and from that point there is direct, albeit fragmentary, evidence of developments in the region. From Portuguese records it would appear that after a period of upheaval and migrations from the mid-16th century, during which newly arrived Maravi chiefs attacked Portuguese allies and attempted to cross the Zambezi, a single dominant Maravi chiefdom emerged. By about 1620, a single chief controlled the north bank of the Zambezi. During the 17th century Maravi Karonga chiefs of the Phiri dynasty formed a confederacy and exacted tribute from their neighbours. Like the Zimba, the Maravi attacked Sena and Tete and tried to conquer the Mutapa state during the civil wars of the 1620s. Frustrated in these efforts, they invaded Mozambique, where they traded with the Portuguese. By the end of the 17th century the rise of Yao traders to the east of Lake Nyasa (Lake Malaŵi) and of Bisa to the west contributed to the disintegration of the Maravi confederacy into several more or less autonomous fragments. To their north lived Tumbuka-speaking peoples, whose differentiation from the southern Maravi peoples probably dates to the mid-18th century, when new patrilineal migrants entered the region from the north and west.

By the early 18th century the Portuguese had also penetrated into present-day Zambia, establishing trading fairs at Zumbo in 1714 and at Feira in 1732. Though there were no highly organized broker kingdoms in the area, a number of different African peoples were trading with the Portuguese by that time.

From the beginning of the 17th century the Portuguese faced increasingly severe competition from Dutch and British ships in their colonial waters, while north of Cabo Delgado the Arabs also took advantage of Portuguese weakness. In 1631 a series of revolts began, which, by the beginning of the 18th century, had driven the Portuguese from the coast north of the Ruvuma River. The Portuguese then turned their trading interests southward. From the mid-16th century onward, they had been trading at Delagoa Bay with the local Tsonga inhabitants, and difficulties in the north and the decline in the gold trade reinforced this trend. In the late 18th century European competition over the ivory trade and the new market for foodstuffs significantly affected political relationships at Delagoa Bay and in its hinterland and transformed the nature of states as far south as northern Natal.

The Portuguese in west central Africa. Portuguese influence in west central Africa radiated over a far wider area and was far more permanent, dramatic, and destructive than on the east coast. Initially, however, it was on essentially peaceful links forged by the Portuguese crown and Jesuit missionaries with the kingdom of the Kongo, north of present-day Angola. Almost immediately, however, slave traders followed in the wake of priests and teachers, and west central Africa was tied to the demands of the São Tomé and transatlantic slave trade.

The Ndongo

Establishment of Portuguese power. Until 1560 the Kongo kings had an effective monopoly over trade with metropolitan Portugal; the vanguard of Portuguese expansion further south were the Afro-Portuguese traders and landowners who established themselves at São Tomé and in small numbers on the coast. When, in the 1520s, the Portuguese crown turned a cold shoulder to the overtures of the Ndongo kingdom further south, local Afro-Portuguese stepped into the breach. They relieved Ndongo of its war captives and supported the king in his military campaigns. With their assistance Ndongo, originally a small chiefdom among the agricultural Mbundu peoples under rulers called the *ngola,* became a major kingdom by mid-century. By 1556, when it threw off Kongo control, the kingdom extended over a wide area between the Dande, Lukala, and Kwanza (Cuanza) rivers.

By the last third of the 16th century, with the waning of its fortunes in the east and the diminishing return from its mines on the Guinea coast, the Portuguese crown began to take a more aggressive interest in its African possessions. The dispatch of Francisco Barreto in 1569 to discover the sources of gold in the Mutapa kingdom was paralleled in 1575 by the appointment of Paulo Dias de Novais as captain general, or governor, over an undefined area between the Dande and Kwanza rivers, which later became known as Angola. Basing himself at Luanda Bay, Novais soon found, however, that his plans for metropolitan development of the south were defeated by tropical disease, the mythical nature of the reputed silver mines, the unsuitability of the region for agriculture, and strenuous African resistance to settler encroachment.

Though the *ngola* held their own against Novais for a generation, within a few years of his arrival the first war of a century of almost constant warfare had begun. The wars soon resolved themselves into efforts to acquire slaves on Portuguese terms. After his death Novais was succeeded by a series of military governors, who found that they could not break into the existing Afro-Portuguese trading networks that depended on alliances with local rulers. Impelled by the increased demand for slaves for the plantations of Brazil, and relying on African mercenaries and allies, they initiated a policy of armed expansion. For the Ndongo kingdom the results were disastrous, as the political balance was overturned and rival factions turned on one another in an effort to secure the royal title.

Mbangala warbands and the kingdom of Kasanje

Civil violence was compounded by years of bitter drought, a recurrent feature of the local ecology. New warlords emerged at the head of bands of starving refugees, who from the late 16th until well into the 18th century swarmed down from the hills, fought one another, and devastated the settled kingdoms. By the end of the 16th century marauding cannibals in well-organized military bands, known as the Mbangala (Imbangala), began to appear along the coasts south of Luanda. In their anxiety to swell slave numbers in the second decade of the 17th century, Portuguese governors were ready to ally with these war bands, and together they dealt the final blow to the Ndongo kingdom by about 1622. By that time the Mbangala had retreated to the middle Kwango, where they established the kingdom of Kasanje. After the 1650s they reestablished their links with foreign slavers and became the chief slave-trading entrepôt between the coast and the Lunda kingdoms in the east over the next two centuries.

Queen Nzinga

As successful as the Mbangala warlords in surviving these turbulent times was Nzinga, half sister of the *ngola* of Ndongo, who created a new refugee-based state in the middle Kwango Valley based on trading slaves to the east. Nzinga dominated the politics of the region for nearly half a century. After the Mbangala–Portuguese victory of 1622, she attempted to use the Portuguese to secure her leadership in Mbundu politics. When this failed she moved eastward with her followers. Conquering the small Matamba polity, she developed it into a powerful slave-trading centre, with networks reaching north to the Kongo and Luanda and eastward to the Lunda.

Both Nzinga and the Mbangala kingdom welcomed the Dutch conquest of Luanda in 1641, but their relief was short-lived. By 1648 the Portuguese had reconquered Luanda and were thus free to reconsolidate their hold on their African possessions. Nzinga ended her days as a Roman Catholic, collaborating with the Portuguese in the slave trade; in 1681 her successor as *ngola* of Matamba was finally subjugated by the Portuguese, 10 years after the remnant Ndongo kingdom further west had been completely dismantled.

At the beginning of the 17th century, as they were penetrating inland from Luanda, the Portuguese also moved southward. In 1617 they established a colony at Benguela, which, like the Kongo kingdom, was annexed as part of Angola in the 19th century. Expansion inland from Benguela, however, like the initial expansion further north, was spearheaded by Afro-Portuguese slave traders, who used the southerly ports to outflank Portuguese control. As the slave frontier moved south, so the entire process of the construction and then dissolution of the slave-trading warrior kingdoms was repeated. By the 18th century the process had reached the Ovimbundu kingdoms on the lower Cuanza; these came to rival Kasanje as suppliers of slaves. By that time the slave trade from Benguela probably exceeded that from Luanda, with slave routes stretching as far as the Kunene River in the south.

Angola in the 18th century. Through the 18th and early 19th centuries the slave trade remained at the heart of Angola's economic existence, with Benguela replacing Luanda as the chief port. The expansion of plantations escalated the demand for slaves, which peaked in the last third of the 18th century. Routes stretched into the heart of Africa, with most of the trade in the hands of African and Afro-European traders. Rich chiefs and warlords still exchanged their captives for the luxury goods of Europe and for supplies of firearms; for the poor, mass-produced textiles, cheap guns, and the fiery Brazilian alcohol were of the essence.

Cheaper Dutch, British, French, and Brazilian manufactures increasingly undercut the Portuguese, and after 1763 the French became the chief traders on the southwest coast. Portuguese attempts to control the coastal trade and Ovimbundu resistance led in the 1770s to drastic Portuguese intervention in the Benguela hinterland in an attempt to install compliant rulers. By that time the Ovimbundu kingdoms, which probably originated in the late 16th and 17th centuries with the arrival of raiders and refugees from the Mbangala and Mbundu kingdoms, were formidable foes. Despite their military victory, the Portuguese were unable to control them effectively until more than a century later.

By the late 18th century Portuguese fortunes in both Mozambique and Angola were at a low ebb. Fewer than 1,000 settlers in each colony huddled on a number of estates around inland forts and along the Bengo and Dande rivers in Angola and along the lower Zambezi in Mozambique. Most of them had intermarried with local peoples and were independent of Portugal. The metropolitan Portuguese were unable to control either the coastal trade or the activities of merchants and warlords, who often acted in their name. In the absence of regular taxation or an effective system of customs and tariffs, the economies of the territories were poor, and their administrations weak and corrupt. Despite a mythology that held that the Portuguese did not differentiate according to race, from quite early times it is clear that whites had superior status and prestige, if not always greater power, in Angola as in Mozambique. Although both territories gained somewhat from the Napoleonic Wars, it was not until the end of the 19th century that the Portuguese regained anything of their initial colonizing energy.

Effect of the slave trade

It is not possible to compile an exact balance sheet of the devastation caused to west central Africa by the slave trade. In the 17th century some 10,000–12,000 slaves were exported annually from Luanda. Although this includes captives from the north as well as the south of the bay, it does not include those smuggled out in order to escape official taxation. The figure may indeed have represented a relatively small proportion of the total population of a vast area, but it was a significant proportion of the economically active young adults. Nor do these figures take account of the depopulation and social dislocation resulting from incessant warfare and banditry and from the famines whose effects were exacerbated by the withdrawal of productive labour.

The better watered regions were probably able to recoup their population losses within a couple of generations, and the introduction of new food crops, such as manioc and corn (maize), which the Portuguese imported from South America, supported larger populations. Nevertheless, the effect of the slave trade was, in social terms, incalculable. Accounts of Ndongo in the 16th century as rich and populous gave way to lamentations about its desolation in the 17th. And the processes of border raids, wars of conquest, and civil strife, which affected the Ndongo and then the Kwango Valley kingdoms in the 17th century, were repeated both to their south and east in the course of the 18th century as the slaving frontier expanded. The ending of the more overt violence as the slave frontier moved on left the weak—women, children, and the poor—vulnerable to innumerable personal acts of kidnapping and betrayal, a process exacerbated by the indebtedness of local traders to coastal merchants and their dependence on the transatlantic economy.

THE DUTCH AT THE CAPE

Apart from the enclaves of European settlement in Angola and Mozambique, the only other area of European settlement in southern Africa in the 17th and 18th centuries was the Dutch settlement at the Cape of Good Hope. In the late 16th century the Cape had become a regular port of call for the crews of European ships, who found the Khoikhoi (Hottentot) herders there ready to barter cattle in exchange for iron, copper, beads, tobacco, and brandy. By the mid-17th century Khoikhoi intermediaries traded far into the interior. These trade relationships profoundly affected the nature of contact between the Khoikhoi and the Dutch when, in 1652, the Dutch East India Company dispatched Comdr. Jan van Riebeeck and 125 men to set up a permanent station at the Cape.

Although initially intended as a refreshment station, the outpost soon grew into a colony of settlement. In 1657 a number of company servants were released as free burghers in order to cultivate land and herd cattle on its behalf; in the next year the first slaves were imported into the colony, mainly from the East Indies. In contrast to the Portuguese, the Dutch feared to imperil the cattle trade that was the raison d'être of the colony, and they prohibited the enslavement of the local population. At the Cape there was no indigenous slavery, and local chiefs had neither the people nor the power to engage in the trade. Nor did Late Stone Age peoples have the agricultural skills of Bantu-speaking farmers.

Dutch expansion and the trekboers. Although the Dutch settlers, called Boers (Dutch *boer*, "farmer") did not raid the local inhabitants for slaves, they soon undermined the loosely organized social structure of the Khoikhoi by their increasing demands for cattle and by encroachment on their grazing and hunting lands. As one Khoikhoi group became impoverished and reluctant to trade, another would take its place. Though the Khoikhoi offered considerable resistance to Dutch expansion, their small numbers enabled the colonists to move beyond their base at the coast far more easily than the Portuguese had been able to in Angola. Moreover, the Mediterranean climate of the Cape was well suited to European settlement. While the Portuguese in Angola and Mozambique were ravaged by disease, at the Cape the Khoisan were decimated by the epidemics brought by Europeans.

The greatest geographic barrier to Dutch expansion was the range of mountains inland from Cape Town. Once these were crossed, settlers relatively rapidly moved both east and north. There, in the arid scrubland, economic necessity and geography dictated a nomadic, pastoral way of life for the settlers, as it had for the indigenous peoples before them. Poor soil and inadequate rainfall could be compensated for by size, and 6,000-acre farms became the norm.

Apart from sporadic attempts, the Dutch East India Company made little effort to follow the cattle farmers, or trekboers, as they came to be known. The new districts of Stellenbosch (1685), Swellendam (1745), and Graaff-Reinet (1785) were large and unwieldy, and their centres far from the expanding colonists. Central governmental authority was even more remote, and the trekboers were left to their own defense, through the "commando" system. They became accustomed to handling emergencies on their own and to ruling over their slaves, servants, and clients as they saw fit.

European settlers at the Cape

The number of white settlers at the Cape initially grew slowly, then rose rapidly from about 1,000 at the end of the 17th century to about 15,000 by the end of the 18th. Despite their varied origins, company insistence that they all speak Dutch and conform to Calvinism, and their shared vicissitudes, led to a certain cultural uniformity. By the mid-18th century class divisions were marked between the affluent merchants and status-conscious company servants of Cape Town; the "gentry" of the southwest, who were engaged in the intensive production of wine and wheat for the world market, based on slave labour; and the trekboers of the interior, who—though always dependent on the market at the Cape to dispose of their cattle and animal products for arms and ammunition and luxuries—had evolved a semi-nomadic way of life that superficially resembled that of their African neighbours more than that of their kin in the western Cape.

As the settler population increased, so did the number of slaves. Experiments in the use of indentured European labour were unsuccessful, and in 1717 the governing Council of Policy decided overwhelmingly against its continuation. By the mid-18th century about half the free burghers at the Cape owned at least one slave, though relatively few owned more than 10. Slavery at the Cape is commonly regarded as a benign institution, but slave mortality rates were high and their fertility low. Manumission, baptism, and intermarriage rates were also extremely low, although a considerable number of newcomers and poorer burghers married slave women, and even, though more rarely, Khoi women. In the absence of white women, miscegenation with Khoi women was much more common, especially in frontier districts. Their children, however, took on the unprivileged status of their mothers, so that the practice did not affect the racially defined class society that was forming at the Cape. By the late 18th century the Cape society was closed and rigid; most blacks (though not all) were "servants" and most Europeans (though again not all) were "masters."

The existence of an enslaved labouring population affected the status and opportunities of the Khoisan once they lost their land and cattle and had to enter the labour market in increasing numbers. Although theoretically the indigenous inhabitants were "freemen," compulsion governed the relationship between master and servant or slave, and the legal status of the Khoisan increasingly approximated that of slaves. As they lost their cattle and grazing lands, they became virtual serfs on the settler farms, though some groups were able to escape beyond colonial borders. In the east they continued to be absorbed by the Xhosa, while on the Orange River, where they became known as the Koranna and Griqua, they established polities that strongly resembled the republics later established by Dutch settlers.

Khoisan resistance to the Dutch

Despite their ultimate subordination, Khoisan resistance to Dutch colonialism erupted into guerrilla war on three occasions in the 17th century; on the first, it almost destroyed the settlement. Cattle raids punctuated almost every decade of the 18th century, a continuation of this resistance. The raids and counter-raids mounted in violence as the Dutch expanded into the northeast; by the last quarter of the century the entire northern frontier was under arms and numbers of settlers were driven from their lands. The vigour of the Khoisan response to trekboer expansion, especially in the district of Graaff-Reinet, was probably related to the increased involvement of the Cape in the world market, as the demand for cattle by European ships led to more intensive exploitation of labour. The entire colonial order was challenged, and there were fears that the Khoisan would penetrate to the wine and wheat farms of the southwest. Those trekboers not ruined by the robberies were exhausted by constant military service; though they wore the Khoisan down through sheer bloodletting, further colonial expansion northward was prevented for some 30 years.

The Cape's eastern frontier. Settler expansion was simultaneously blocked in the east, where trekboers had come up against the populous and settled agricultural Xhosa people in the area of the Great Fish River. Although trading contact had existed between the colonists and the Xhosa from early in the 18th century, by the 1770s they competed over grazing, water, and the terms of the cattle trade. Demands for fresh meat at Cape Town increased during the Anglo-French wars and the American Revolution; Khoikhoi reserves of cattle had been exhausted, and the trekboers expanded to the edge of Xhosa settlement. Various attempts to draw a frontier between the colonists and the Xhosa were unavailing: when, in 1778, the Dutch decreed the Great Fish River the boundary between the colony and the Xhosa, there were Xhosa in the contested area known as the Zuurveld to the west and trekboers embedded in Xhosa society to the east.

The establishment of the district of Graaff-Reinet, with a full-time magistrate, hardly improved matters. The area of his jurisdiction was vast, and its inhabitants were ill-disposed to listen to a government that had left them alone for more than half a century. Before the 18th century was over, minor cattle raids had escalated into two frontier wars. They were the prelude to a struggle that lasted almost 100 years; the trekboers only expanded again after moving north and outflanking the Xhosa.

Despite the firearms of the Dutch, the two combatants were by no means unevenly matched, for the Xhosa were superior in numbers. Both sides suffered from internal divisions that complicated the struggle. Thus, the first two frontier wars ended in a stalemate.

Southern Africa, 1800–1910

In the 19th century new internal and external forces radically changed the demographic, economic, social, and political configurations established over the previous two centuries in much of southern Africa. Though certain areas remained unaffected by change until the late 19th century, for the area as a whole the first three-quarters of the century brought many new developments. Some, like the rise of the Zulu kingdom, appear to have had a largely internal dynamic; others, like the expansion of European settlement, trade, and missionary endeavour in the interior, were more directly a response to wider transformations in the international political and economic order.

THE MFECANE AND ITS EFFECTS

The first of these changes began toward the end of the 18th century in the area of present-day Zululand, where Zulu were one of numerous Bantu-speaking Nguni chiefdoms, closely related to the Xhosa on the Cape's eastern frontier. Their rise to power in the second decade of the 19th century profoundly altered demographic and political relationships in most of the subcontinent. Their expansion and its repercussions are known in the south as the Mfecane, or Difaqane—the "Crushing." Yet out of the Mfecane new polities were formed, and for those not destroyed by their advent, the invaders brought a challenge and an example of military and social organization they could use in resisting later intruders.

Shaka and the Zulu

Though it has been customary to attribute the warfare solely to the evil genius of Shaka, who became king of the Zulu in 1816, its pattern suggests that the pressure of people and cattle on Zululand's highly differentiated ecology at a time of reduced rainfall gave the rise of the Zulu empire its explosive force. Shaka's internal policies, which attempted to dramatically change existing social and economic relations through population control, lend force to these arguments.

The rise of the Zulu kingdom must also be put in the wider context of the rise of larger states in the region. From the middle of the 18th century the increased demand and price paid for ivory from Delagoa Bay had led to important political and social changes in those kingdoms supplying the market—both the result of attempts by chiefs to monopolize the trade and the increasingly centralized control that elephant hunting gave rulers over the military powers of young men. Out of this grew the *amabutho,* or age-graded regimental system, which became a key feature of warfare in the early 19th century. Access to traded luxuries and control over the *amabutho* enabled those chiefs closest to Delagoa Bay to greatly extend their authority. By the late 18th century a cattle market was established at the bay, and the *amabutho* began raiding their neighbours for cattle. This trend intensified during the severe drought of 1802–03. Whereas in the past population pressure and conflict had led to the colonization of new areas, this option was no longer open by the end of the 18th century. As the skirmishing increased, chiefdoms and fragments of chiefdoms were sent ricochetting off from the centre, only to dislodge their neighbours in turn and thereby increase the circle of disorder at the periphery.

While the opening stages of the Mfecane predated Shaka's accession to power, his military genius lay behind the rise of the Zulu. He used readily available military tactics with originality and within a few years had brought all the peoples from Delagoa Bay to the north and Pondoland to the south under his sway. Out of the numerous chiefdoms between the Tugela and Pongola rivers he welded a highly centralized military kingdom; to the south he created a buffer zone of havoc and destruction.

)rigins of he Swazi

The Ndwandwe under the leadership of Zwide were initially Shaka's most formidable rivals. In 1816 the Ndwandwe had already pushed the Ngwane people and their Dhlamini ruler, Sobhuza, to the north of the Pongola River, where the Dhlamini conquered and absorbed the local Nguni and Sotho chiefdoms and created the Swazi state. Under frequent attack from the Zulu, the Swazi withdrew northward and escaped the fate of those who remained within the Zulu orbit.

By 1820 Shaka had defeated Zwide's armies, which then broke into three groups. Zwide himself retired, but two of his commanders, Soshangane and Zwangendaba, fled northward, to be followed shortly thereafter by Nxaba, Msane, and Ngwane Maseko. Battles among Zwangendaba, Nxaba, and Soshangane on the middle Save River left the last victorious, and he built up the Zulu-type Gaza empire, which at its height stretched between the Zambezi and the Komati rivers; reduced both the Portuguese and the Shona to tributary status; and incorporated many of the peoples of southern Mozambique.

The Ngoni across the Zambezi

After their defeats both Zwangendaba and Nxaba moved to the highlands of Zimbabwe, fighting one another, disrupting the Mutapa and Rozwi kingdoms, and incorporating Karanga peoples. In 1835–36 Zwangendaba crossed the Zambezi, while Nxaba's followers who had joined up with the Maseko Ngoni split in two, one section remaining in what is now Zimbabwe, the other eventually settling in the Kirk Range in the 1860s. From there, they raided a wide area west of the Shire River and, after 1875, to the east of Lake Nyasa.

Across the Zambezi, Zwangendaba's followers raided into what is now Tanzania, Zambia, Malaŵi, and Mozambique, establishing tributary relations with many surrounding chiefdoms and incorporating many of the defenseless Cheŵa, Nsenga, and Tumbuka agriculturalists into their state. After Zwangendaba's death on the borderlands of Zambia and Tanzania in 1843, the Ngoni split into several groups. Two sections settled in Tanzania, and three of Zwangendaba's sons, including Mpezeni and Mbelwa (or Mombera), moved southwestward and southeastward. Having conquered the enfeebled Bisa, Mpezeni clashed fiercely with the Bemba, by then becoming one of the most powerful polities of northeastern Zambia; he finally reoccupied Nsenga country by the mid-1860s.

Mbelwa, with the largest fragment of Zwangendaba's following, moved back into Tumbuka country, where in 1855 he overthrew the local Kamanga chiefdoms, weakened by Swahili slave raiding. Ngoni movements were dictated by the need to avoid the more powerful African polities and to find new food resources after local cattle and crops had been exhausted through their raids. Though the Ngoni aristocracy retained political and social dominance through their monopolization of cattle, there were many tensions between the incorporated peoples and Ngoni commoners. They nevertheless retained their cohesion into colonial times.

Origin of the Mfengu

While the fragments of the Ndwandwe confederacy were making their way into east central Africa, chaos was being spread southward and westward by Shaka's campaigns against the chiefdoms of Zululand and Natal. The ever-widening circle of disorder reached the western Sotho and Tswana groups across the Vaal River, in Botswana, and in the northwestern Cape, as raiders and refugees moved off in all directions. Many went south to join fugitives making their way from Shaka's campaigns in Natal to the territory east of the Cape Colony. All these refugees, whose social structures had been disrupted by the years of warfare, became known as Mfengu (or Fingo). They rapidly emerged as a group of enterprising peasants, many of them converts to mission Christianity and ready to collaborate with the British (see below *British occupation of the Cape*).

Mshweshwe and Mzilikazi

Yet other shattered people scrambled to safety in the mountain fortresses of what is now Lesotho. There Mshweshwe, the Koena leader, built a new kingdom at Thaba Bosiu, based on his ability to provide his followers with raided cattle and protection and a series of shrewd diplomatic marriages. He invited missionaries in as advisers and was able, during his lifetime, to retain a precarious independence for his people.

Other dislodged Highveld peoples joined the Griqua polities along the Orange River or continued raiding along the Vaal and in the western Transvaal, where the disorders prepared the way for the coming of Mzilikazi. He moved out of Zululand in 1823 with 300 of his followers, known as the Ndebele (or Matabele), and over the next 15 years moved westward across the Transvaal in flight from the Zulu armies. Mzilikazi built up a 20,000-strong raiding kingdom by absorbing local Sotho peoples into his regiments. In 1837, after conflicts with the Zulu, with Griqua raiders from the Orange River, and with the expanding white farmers (or Voortrekkers) from the Cape, he led his followers across the Limpopo into what is now southwestern Zimbabwe.

Across the Limpopo, Mzilikazi found it easy to establish himself, for the Shona polities were ill-prepared for the new form of warfare and were weakened by drought. As in the Transvaal, Mzilikazi absorbed the local populace into his age-set regiments; a castelike society evolved, with the original Nguni nucleus on the top, the Sotho in the middle, and the Shona at the bottom. Beyond their immediate settlement, the Ndebele established relationships with surrounding peoples that ranged from exacting regular tribute to random raiding. Beyond the range of Mzilikazi's armies many Shona chiefdoms remained independent; by the 1870s they were trading firearms to resist Ndebele incursions.

The Kololo

Yet another group dislodged by the Mfecane, the composite Sotho group known as the Kololo, made their mark in west central Africa. Defeated in the warfare among the western Tswana, Sebetwane led his followers across the Zambezi, where in about 1840 they conquered the Lozi kingdom, which had been built up in the 18th century and then dominated what is now western Zambia. Weakened by a succession dispute and deserted by its tributaries, the ruling Lozi dynasty was swiftly crushed. Kololo triumph was short-lived, however; by 1864 the ravages of malaria, the accession of a weak and diseased king, and the revival of Lozi royal fortunes put an end to their hegemony. Nevertheless, a variant of Sotho is still the language of the region.

THE SLAVE AND IVORY TRADE NORTH OF THE ZAMBEZI

Long before the beginning of the 19th century, Central Africa was crisscrossed by trade routes to both the east and west coasts. By the second half of the 18th century there was probably a route all the way from Luanda in the west to Mozambique in the east. Until the mid-19th century the Luanda trade was mainly concerned with slaves. With the growing demand in Europe for ivory and the attempted legal abolition of the slave trade in Angola in 1836, however, ivory became the most important export, and Benguela replaced Luanda as the main outlet. At the same time, the Mbangala of Kasanje and the Portuguese traders on the coast were displaced as intermediaries by the Ovimbundu of the Bié Plateau, whose Chokwe neighbours were already renowned as elephant hunters. They penetrated deep into Central Africa, and with their firearms decimated the elephants in one region after another. By 1850 they were in Luvale and Lozi country and were reaching the southern Congo forests.

The ending of slave exports to the Americas did not end slavery in Africa, as local merchants, chiefs, and elders used slave labour to produce the tropical goods demanded by British merchants and to carry to the coast growing quantities of wax and ivory in the 1840s and '50s and rubber in the 1870s. By 1910 wild rubber accounted for 77 percent of Angola's exports by volume. Successful in the short term, the over-collection of wild rubber destroyed an irreplaceable natural resource, while new concentrations of population destroyed the ecology of a drought-prone environment.

Slaves and ivory in east central Africa

In east central Africa the ivory trade was more significant in the 18th century and remained so in the first half of the 19th. It greatly expanded the demand for slave carriers, however, as did the labour demands of the tropical plantations of Réunion, Mauritius, Pemba, and Zanzibar, and later of Cuba and Brazil, at the very time that the abolition movement was at its height. With the abolition of the trade north of the Equator in 1815 and 1817, slaving in east central Africa acquired enhanced significance, while the incursions of the Nguni greatly increased the numbers of slaves available. The Gaza and other Nguni offshoots engaged in the slave trade on a considerable scale, especially in the early years of raiding when they off-loaded their captives at the coast. By the 1830s Mozambique ports were supplying as many slaves as Angola, while the arrival of the Voortrekkers in the Transvaal opened up another outlet. The trade was further augmented by slaves captured by the Chikunda soldiers of the Zambezi warlords, whose rival armies were by the 1840s competing over the Zambezi Valley and over control of trade routes to Central Africa.

Failure of abolition

In both east central and west central Africa, British efforts to abolish the slave trade in 1836 and actual slavery some 20 years later were ineffective until the end of the century. In Angola the attempts led to widespread resistance, and to revolts in Luanda and Benguela. In Angola and Mozambique the external trade was simply replaced by increased internal exploitation of slaves and new ways of extracting forced labour from the African populace. On the east coast British pressure to ban the trade through the sultan of Zanzibar was easily circumvented and, though the abolition treaty forced on the Zanzibaris in 1873 was more effective, the reduced coastal demand for slaves led to even more ruthless methods in the interior of east central Africa, as slaves, no longer needed for export, were exploited locally.

East coast Arabs began to play a much more dynamic role in the interior. Though initially they worked through local chiefs, they came to exercise a wide military and political jurisdiction over the northern routes from their strategically placed commercial centres, many of which became plantations with their own slave labour forces.

The most important area of slave raiding appears to have been in present-day Malaŵi and northeastern Zambia, where predatory overlords, such as Msiri of the Nyamwezi and the Arab Tippu Tib, devastated a wide area from bases in the Congo. To the east of Lake Nyasa, the Yao—keen traders from the 17th century—moved from ivory trading to slave raiding, obtaining firearms from the Arabs, subjugating the Cheŵa agriculturalists, and building up powerful polities under new commercial and military leaders. Similarly, the Bemba were also able to increase their power through the slave and ivory trade, raiding the loosely organized Maravi peoples to the west of the lake from their stockaded villages on the infertile Zambian plateau. Although they never became large-scale slave traders, preferring to incorporate their captives, the Ngoni invaders also added to the dislocation and turmoil. Although the first European observers probably exaggerated the degree of depopulation, the political geography of the region was transformed and vast numbers of people were torn from their social settings.

Arrival of missionaries

It was into this fluid and turbulent world that European missionaries and traders, many of them inspired by David Livingstone, arrived in the 1870s. In 1875 the Free Church of Scotland established the Livingstonia Mission, while the Church of Scotland began work among the Yao at Blantyre. From Lake Nyasa the Scottish missions spread inland to northeastern Zambia, to be followed by a large number of representatives of other Christian denominations in the last decades of the century. In general, the missionaries found that while the weaker, raided peoples quickly adopted Christianity and the education that went with it, people whose livelihood depended on the slave trade were unenthusiastic.

In 1883 the Scottish missionaries and the African Lakes Company (a Scottish-financed firm set up to support the missionary endeavor and encourage "legitimate commerce") persuaded the British government to appoint a consul for Malaŵi and Mozambique, with a watching brief on the slave trade. The British government became increasingly embroiled in struggles with the Arab traders—which resulted in open warfare in 1887–88—and with the Portuguese. With the appointment of Harry (later Sir Harry) Johnstone as consul in 1889, they came to play a more aggressive imperial role in the region.

BRITISH OCCUPATION OF THE CAPE

By the beginning of the 19th century the small Dutch East India Company outpost at the Cape had grown into a sprawling settlement in which 22,000 whites dominated a labouring class of some 25,000 slaves and an indeterminate number of Khoisan peoples, many virtual serfs, as well as a growing number of Xhosa in the eastern districts. During the Napoleonic Wars the settlement passed first to the British (1795–1803), then to the Batavian Republic (1803–06), and then to the British again in 1806.

The substitution of British for Dutch rule at the Cape meant the supplanting of a mercantile company by an imperial state already undergoing the complex economic, social, and ideological changes of its industrial revolution. These developments in Britain had momentous effect at the Cape. Much of the colony's 19th-century history can be seen in terms of the transformation of a company outpost into a more fully capitalist state. The initial impulse behind Britain's annexation of the Cape was to protect its sea route to India. The demand, however, that the colony pay for its administration, produce raw materials for the metropole, and provide a market for its manufactures and a home for its unemployed locked the Cape into the world economy.

Constitutional developments

In its constitutional development the Cape Colony followed the pattern set by Britain's other settler colonies in the 19th century. It was initially a crown colony governed by an autocratic governor, whose more extreme powers were modified by the arrival of some 5,000 British settlers in 1820 and the presence at the Cape of an articulate middle class. After a decade of intense constitutional struggle, the colony was granted representative institutions in 1853. Elections were based on a colour-blind franchise that included Coloured voters (*i.e.,* voters of mixed racial descent) who, it was hoped, would support the British and counteract the influence of Dutch-speaking colonists.

In 1872 the Cape was granted full, responsible government. The colour-blind franchise was retained but came under increasing attack. As a strategy for incorporating the more prosperous black peasants and artisans, it was supported by white merchants, professionals, and officials. With the annexation of African territories and the creation of a mass black working class, however, it proved vulnerable (see below), and in 1887 and 1892 the franchise qualifications were changed in an effort to restrict the numbers of black voters.

Property and labour relations

From the outset, the Cape colonial state intervened to secure and extend existing property relations. Soon after their formal acquisition of the colony, the British attempted to reform the nature of Dutch landholding to immobilize the trekkers, provide the conditions for capital-intensive farming, and raise government revenue through the auction of crown lands. Later, land policy was geared toward fostering independent peasant production among Africans, though by the end of the century forms of individual tenure were encouraged mainly to force out labour.

Indeed, the most dramatic area of state intervention in the 19th century, as in the 20th, was in the creation and control of labour. By the 1820s and '30s the imposition of British notions of "free labour" had an explosive effect on the class relations of a colony dependent on both slaves and serfs. In 1807 the slave trade was abolished in the British Empire, and in 1834 slaves were emancipated. At the same time ordinances were passed attempting to regulate relations between "masters and servants," to ensure standards of treatment and equality before the law, and to transform the basis of the labour supply from serfdom to contracted wage labour. The most important of these measures was Ordinance 50 of 1828, which conferred a degree of civil liberty on the Khoisan while facilitating their movement into the labour market. The "modernizing" policies of the British government threatened the world of the Dutch farmers, and several thousand of them left the Cape Colony between 1834 and 1845 in the movement known as the Great Trek.

Labour and property relations were not simply a matter of internal policy, however. In the 19th century, as in the 18th, the expansion of commodity production led to the extension of colonial frontiers and the conquest and incorporation of the Bantu-speaking farmers to the north and east. By the mid-century the Cape's northern frontier was the Orange River, while in the east the frontier had moved from the Great Fish to the Kei River.

The eastern frontier

The British were also confronted by a series of interlocking crises in the eastern districts. In the space of eight years they were faced with three rebellions by the trekboers of Graaff-Reinet (1795, 1799, 1801) and the third war between settlers and Xhosa on the eastern frontier, which coincided with a mass uprising of the recently subordinated Khoisan of Graaff-Reinet, who deserted with their masters' horses and guns. The British managed to divide the African forces, but were unable to dislodge the Xhosa from Zuurveld.

Although the appearance of regular British troops on the frontier in 1811 pushed the Xhosa west of the Great Fish River and tilted the balance in favour of the settler, in 1819, 1834–35, and 1846 the almost constant unrest there exploded into war. Increasingly the object of the wars was not only African land and cattle but also labour. For most of the century the Cape was dependent on British troops for its defense and for the conquest of African territory. Each conflict pushed the Xhosa closer in on one another, and the tensions between Xhosa chiefdoms grew as the colony extended into the eastern Cape. By mid-century the Xhosa were formidable foes, using firearms and adopting guerrilla tactics. Thus the eighth war in 1851 was the most drawn-out and expensive of all. As in 1799, a simultaneous uprising of Khoisan–Coloured people at the Kat River settlement (established as a buffer for the colony in 1828) exacerbated the situation for the colonists.

In the end, it was not so much British arms or settler prowess that defeated the Xhosa as the internal tensions that had built up as a result of the expansion of traders, missionaries, and settlers. These pressures were increased by the expansion of commercial sheep farming and the arrival of the Mfengu refugees. During the 1834–35 war the Mfengu fought on the British side and were rewarded with Xhosa lands and cattle—a pattern that was to be repeated in each subsequent war. In 1857 the Xhosa—exhausted by years of attrition, in the midst of severe drought, and faced with the aggressive policies of Gov. Sir George Grey—turned to millenarian prophecies. In the belief that to do so would raise their ancestors from their graves to drive the whites into the sea, they slaughtered their cattle and destroyed their crops. When the awaited salvation failed to materialize, some 20,000–30,000 Xhosa streamed across the frontier to seek work in the colony. An equal number died of starvation in the Transkei. Although Xhosa fought the colonists again in 1877 and 1879, by then their lands were impoverished and eroded and many were forced to work in the colony. Thereafter, the annexation of the remaining African territories between the Cape and Natal proceeded peacefully, if piecemeal. The last of the independent kingdoms to pass into Cape hands was Pondoland, in 1895.

If the Cape's eastern frontier saw some of the fiercest and most prolonged resistance to colonial expansion anywhere in 19th-century Africa, it was also the scene of the closest black and white interaction. Trading relations in the frontier zone existed from the early 18th century, and by the 19th, with the coming of the British, trade was expanding rapidly. The terms of this trade were unequal: Africans exchanged their cattle and crops for beads and brandy; yet by the last third of the century, they were earning sufficient cash to purchase the plows, seed, and fertilizer that were transforming African agriculture. The advent of missionaries from the beginning of the century also encouraged the growth of peasant communities, cultivating for the colonial market.

Role of the missionaries

The first converts in the Cape were the Khoisan and mixed groups along the Orange River—the Griqua—and their example was followed early in the century by a number of southern Tswana people. In the eastern Cape, Mfengu refugees were soon the focus of mission activity. By the 1850s the Mfengu were regarded as the most progressive element in the eastern Cape. After the cattle killing of 1857, and even more after 1879, an increasing number of Xhosa also turned to Christianity and to new political techniques with which to confront colonialism. By the last quarter of the 19th century missionaries of almost every denomination were working among the people of southern Africa.

THE ORANGE FREE STATE AND LESOTHO

When the British withdrew from Transorangia in 1854, a far greater proportion of the small settler community was already tied into the world economy through wool production. Many of its burghers (citizens) had trekked across the Orange River before the Great Trek and wished to retain links with the Cape and British economies. After the British withdrawal they established a Volksraad, or parliament, and drew up a U.S.-style constitution. Of a population in 1875 of some 125,000, only the 26,000 whites had burgher status, but many European observers in the 1870s considered the Orange Free State a "model" republic. Despite the Dutch ancestry of the majority of the settlers, English was the language of commerce and education into the 20th century.

Even before the British withdrawal of 1854 there had been constant friction between the Sotho and the settlers over the land. To the east of the trekkers, Mshweshwe had built up a formidable mountain state, and by 1848 he had attracted some 80,000 followers. With the restoration of peace, many hoped to return to their homes, only to find that the Boers had settled their most fertile lands. Overcrowded, the Sotho spilled over onto Boer farms, as well as onto lands occupied by other African groups.

Despite Mshweshwe's attempts to keep the peace, cattle raiding from those of his subjects who saw no hope of recovering their land by negotiation, together with increasing demands for land and labour from the settler-sheep farmers, led to war in 1858 and again in 1865. On the first occasion, the Orange Free State was forced to sue for peace. On the second, Basutoland, starved of arms and torn by internal jealousies, was beaten. Some chiefs, especially in the north, offered their allegiance to the Boers and, with their followers, became labour tenants on their farms; others fled into the Transkei. In 1868, in response to repeated appeals from the Sotho, the governor of the Cape intervened to declare the Sotho British subjects and their land British territory—much to the anger of the Free Staters. Nevertheless, the Orange Free State gained in security and held the Caledon Valley. In 1869 the frontiers of Basutoland were delimited and shortly thereafter it was handed over to the Cape. In 1881, however, when the Cape government tried to disarm the Sotho, a war that the colony could not control broke out, and in 1884 the Cape returned the territory to Britain.

GOLD AND THE SCRAMBLE FOR SOUTHERN AFRICA

Even more important than the discovery of diamonds in transforming the subcontinent was the discovery of vast quantities of gold deep in the southern African interior. While the diamond discoveries had begun a series of social and economic changes, the discovery of gold—and especially the deep seams of gold of the Witwatersrand, or the Rand, in 1886—intensified and carried this revolution far beyond the Limpopo. From the late 1880s, gold outstripped diamonds as the region's most important export, and by 1898 the value of gold production had risen to £16,000,000, about one-fifth of world output.

When local capital resources from the Cape and the diamond mines proved inadequate, new funds flowed in from Britain, Germany, and France. The coastal colonies competed to control the lucrative trade. Immigration no longer needed official encouragement: in 1870 the total white population of southern Africa was probably under 250,000; by 1891 it had increased to more than 600,000; and by 1904 it was more than 1,000,000. Skilled miners were drawn from the mining frontiers of the world to the southern African interior by the lure of high wages.

Mineral discoveries led to economic development in almost every sphere. Roads, railways, and harbours had to be built. The need for coal led to the renewed exploitation of mines in the Cape and the opening of others in Natal and the Transvaal. Farming was commercialized. Manufacturing, though in its infancy, responded to the new markets, and land prices soared. The demand for labour was universal and insatiable. South of the Limpopo, the remaining independent African territories were annexed, and both within and outside the frontiers of the colonial states land, labour, and mineral concessionaires were spurred by prospects of further discoveries and the availability of speculative capital.

The Limpopo constituted no barrier to the flood of concessionaires, and between 1889 and 1895 all of the African territories up to the Congo were annexed. In Central Africa the British competed with the South African Republic, Portugal, Germany, and Belgium. In east central Africa, to the west and south of Lake Nyasa, the thrust from the south encountered the less powerful, but nonetheless still significant, antislavery missionary frontier from the east.

From the 1860s it was known that there were "ancient gold workings" between the Zambezi and the Limpopo. By the mid-1880s Lobengula, king of the Ndebele, was surrounded by concessions hunters. In 1887–88 alarm at possible Transvaal expansion across the Limpopo led the high commissioner at the Cape to declare the region a British sphere of interest. It was at this point that Cecil John Rhodes decided to enter the arena.

Cecil John Rhodes

The story of how Rhodes came to South Africa to repair his frail health and stayed to become a millionaire on the diamond fields before he was 30 is legendary. At Kimberley chaotic underground conditions—the result of the uncontrolled enterprise of individual miners and the unrestricted flow of diamonds to the world market—had led to monopolization. In 1880 Rhodes entered the Cape parliament to protect his interests, and he played a key role in securing the British annexation of Bechuanaland. Backed by international financiers, Rhodes bought out his rivals by 1888 and created the consortium of de Beers. By 1890 he had become the prime minister of the Cape Colony, by far the most powerful man in South Africa.

By 1888 Rhodes was looking to Central Africa to find a "second Rand," open a new area for exploitation, and outflank the increasingly powerful South African Republic. His agents secured exclusive mining rights in Lobengula's kingdom in 1888, and although the king soon regretted his concession, Rhodes immeasurably strengthened his hand by forming the British South Africa Company (BSAC). The company was granted a royal charter by the British government to exploit and extend administrative control over a vast area of southern and central Africa. An amalgamation of London financial interests, bedecked by an array of aristocratic shareholders, the BSAC represented the most naked exploitative enterprise.

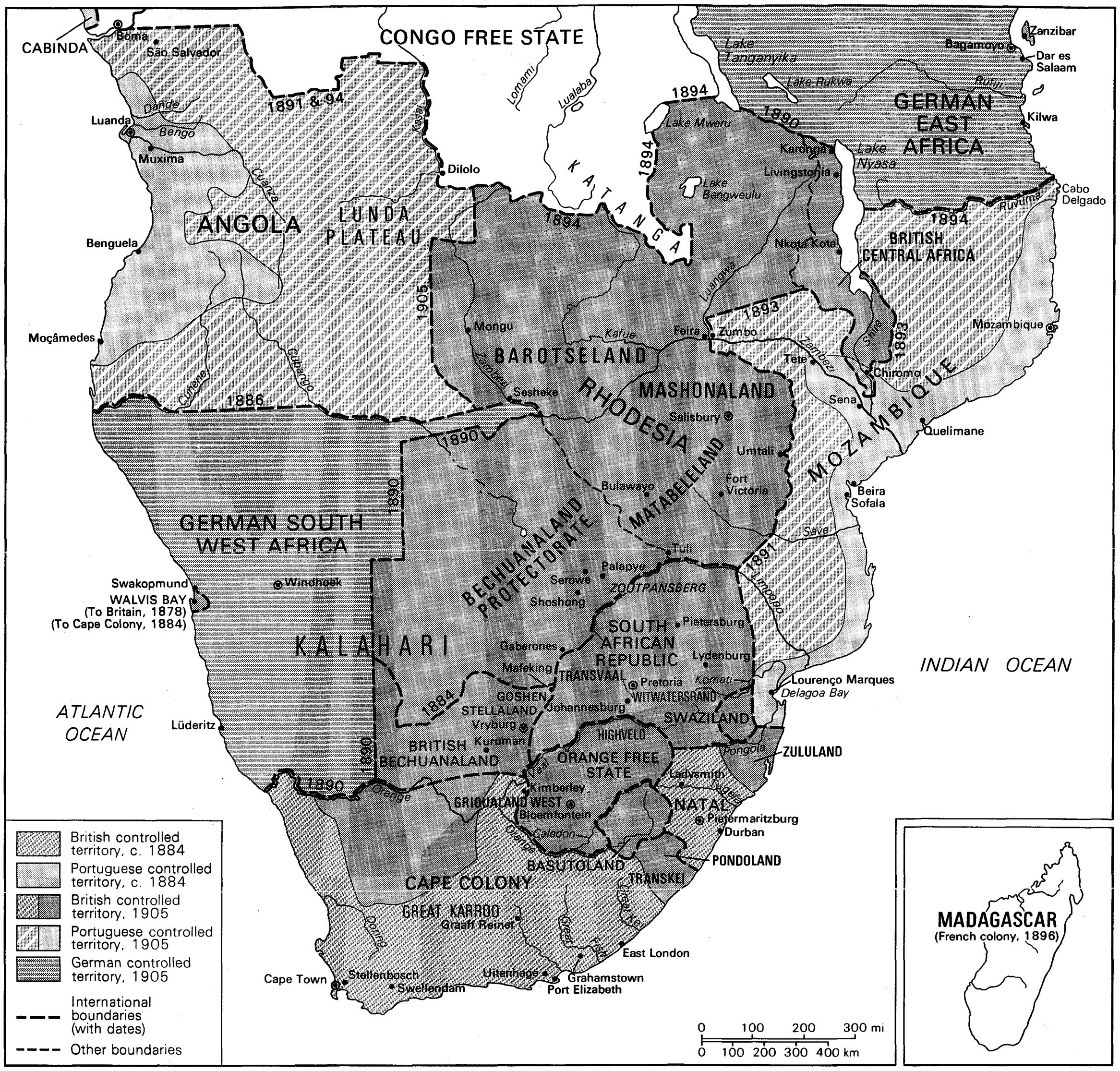
British South Africa Company

Even before his charter was ratified, Rhodes was encouraged to stake his claims across the Zambezi, where the British were anxious to preempt their European rivals. In 1888–89 there was a flurry of treaty making by Rhodes's agents. These included Harry Johnston, who had been appointed British consul in Mozambique. A major objective was Msiri's kingdom in Shaba (Katanga), where the existence of rich copper deposits was already suspected, but Rhodes and Johnston were anxious to extend imperial hegemony over as wide an area of south central Africa as possible. None of Rhodes's agents persuaded Msiri to grant the BSAC concessions, and in the following year Shaba was forcibly annexed to the Congo Free State (now Zaire). Nevertheless, the whole of present-day Malaŵi and Zambia was staked out for crown and company.

The precise nature of the treaties and their dubious legality was of little importance. In general they secured mineral and sometimes land concessions for the BSAC, while the chiefs agreed to accept British jurisdiction over non-Africans in their domains and over external relations. But the frontiers of Africa were by then being decided in the chancelleries of Europe, and there the treaties provided useful bargaining counters. In 1890–91, British, Portuguese, and German conventions essentially established the frontiers of the modern states of Zimbabwe, Mozambique, Malaŵi, and Tanzania in the east and of Zambia, Angola, and South West Africa in the west.

For the British government the BSAC's great advantage was its promise to people Central Africa with white immigrants who would make British occupation effective against contending European powers and bring the necessary capitalist development to the interior at minimum cost. In 1890 Rhodes sent a "Pioneer Column," consisting of 200 white settlers and 150 blacks, backed by 500 police, into Mashonaland. From the outset, the real goal was the Ndebele kingdom. In 1893, to solve the company's financial problems, its resident commissioner, Leander Starr (later Sir Leander Starr) Jameson, deliberately provoked war with the Ndebele. The war was followed by a boom in BSAC shares and the flotation of several new speculative mining companies, with the BSAC receiving a half-share

Invasion of the Ndebele kingdom



Colonial southern Africa, 1884–1905.

From J. Fage, *An Atlas of African History;* Edward Arnold (Publishers) Ltd.

in any profits. Yet, by the end of 1894, it was clear to Rhodes that the "second Rand" was not in Matabeleland, but on the Rand itself, where new deep-level mines were coming into operation. As their hopes of instant wealth through gold waned, so settlers and the BSAC turned to African cattle and land.

Settlers who participated in the war were granted farms on a lavish scale and in addition were rewarded with mineral claims, both of which soon passed into the hands of speculative syndicates. A land commission perfunctorily set aside two reserves for the Ndebele on poor soils. Vast quantities of African cattle were looted, and forced labour and taxation were prized out of both Shona and Ndebele communities by the newly formed Native Affairs Department and unpopular Ndebele police.

Ndebele
nd Shona
prisings

The Ndebele rose in March 1896 when an outbreak of rinderpest ravaged what was left of African cattle and the BSAC removed its police in anticipation of an attack on the South African Republic (see below *The South African War*). The Ndebele were joined in the middle of the year by a number of eastern Shona polities. Only the arrival of imperial troops and the collaboration of other Shona groups saved the company state. The Chimurenga, as the uprising was called by Africans, led to direct British government intervention in BSAC affairs for the first time with the appointment of a British resident commissioner in Bulawayo responsible to the imperial high commissioner in Cape Town.

Despite the hopes of the British government that the BSAC would effectively occupy the territory north of the Zambezi, these events left Rhodes little time, energy, or resources to spare until the late 1890s. Opposition from the missionaries and the African Lakes Company ensured that the region around Lake Nyasa and the Shire Valley was separated from the BSAC sphere; it was declared the British Central African Protectorate in 1891, with Johnston as commissioner. During the early years of the protectorate Johnston engaged in a spate of wars against the Swahili and Yao slave and ivory traders, who feared the loss of their livelihood. Given the ethnic fragmentation and social divisions of the region, he found little difficulty in implementing a policy of divide and rule. His antislavery wars had the advantage of releasing labour for European employment. His vision was not dissimilar to that of Rhodes: the protectorate's future development should be based on the marriage of white enterprise and black labour, assisted by Asian middlemen.

Lewanika
and the
Lozi

West of the protectorate Africans were more gradually drawn under colonial rule. Neither the BSAC nor Johnston had time to pay attention to the pleas from the Lozi king, Lewanika, that the British honour the treaty he had signed with Rhodes's agent in 1890, in which he had been promised technical and financial assistance in exchange for mineral concessions. Lewanika's "scramble for protection" in the 1890s was dictated from the same circumstances that initially led him to invite whites into his kingdom in the mid-1880s. The 20 years after the restoration of the Lozi monarchy, after the Kololo interregnum, had been filled with civil war and succession disputes. By inviting in the missionaries, and subsequently the BSAC, Lewanika hoped to bolster his internal position and gain the skills to enable him to deal with the intruders.

Finally, a BSAC administrator was sent to the Lozi kingdom (Bulozi) in 1897. Contrary to Lewanika's expectations, his arrival spelled the end of Lozi independence. Although his choice of manoeuvre and accommodation, rather than resistance, and the lack of minerals in his domains meant that he secured a protected status from the company, over the next decade the powers of king and aristocracy were whittled away. British insistence on the abolition of serfdom and slavery in 1906 undermined the cultivation of the floodplain on which Lozi agriculture depended, and Lewanika's hopes to retain control over the modernization of his state were in vain. Bulozi became a protectorate within a protectorate, tied to the southern African political economy.

In northeastern Zambia, too, the process of imposing colonial rule was later, but in the end swifter and less violent, than it had been to the south or west. A series of natural disasters in the 1890s weakened the ability of even the more powerful groups to resist, while for the others BSAC rule seemed a release from the raids and exactions of their Bemba, Ngoni, and Swahili overlords. Above all in this region, however, it was the lack of resources that spared people major confrontations with colonialism: in the one area where gold was believed to exist, among Mpeseni's Ngoni, the onslaught was as dramatic as in Zimbabwe and the expropriation as brutal.

THE SOUTH AFRICAN WAR

It was in the south that the most bitterly fought and costly colonial war of the period took place. By 1895, despite the progress of his plans all over southern Africa, Rhodes's hopes of gold in Matabeleland were disappointed and the development of his deep-level mines on the Witwatersrand were a heavy drain on capital. The South African Republic's inability to adequately create and coerce a labour force and its liquor, railway, and dynamite policies were irksome to the deep-level mine owners with their huge demand for labour, tight working costs, and great need for explosives. Making use of the largely fomented clamour of British immigrants in the Transvaal, known as Uitlanders ("aliens"; *i.e.,* non-Boers), over their lack of voting rights, and with the full knowledge of Joseph Chamberlain, the British colonial secretary, Rhodes plotted the armed overthrow of the republic by his lieutenant, Leander Starr Jameson.

The Jameson raid

The affair was a complete fiasco. The hoped-for uprising of Uitlanders failed to materialize, and Jameson was soon arrested by the Boers. Rhodes was forced to resign from the premiership of the Cape Colony, and the alliance he had constructed between the English and Afrikaners in the Cape was destroyed. Cape Afrikaners united behind the republican Pan-Afrikaner opposition to British imperialism. The dispatch of Sir Alfred Milner as British high commissioner to South Africa in 1897 exacerbated matters and Paul Kruger, president of the South African Republic, began to rearm. The war broke out in 1899 and lasted until 1902. In all, nearly 500,000 British troops were deployed against a Boer force of 60,000–65,000, at a cost to the British taxpayers of £222,000,000. Some 6,000 British soldiers died in action, another 16,000 of infectious diseases. The Boers lost some 14,000 in action and 26,000 in so-called concentration camps. The camps left a deep scar on Afrikaner consciousness and led the British Liberal leader, Sir Henry Campbell-Bannerman, to refer to his country's "methods of barbarism." The total number of African dead is unrecorded, although according to low official estimates more than 13,000 died in the camps. In the end, Britain's greater resources wore the Afrikaners down, and their leaders were forced to sue for peace, which was signed on May 31, 1902.

The clashes between Kruger and the representatives of British imperialism—Rhodes and Milner—are usually portrayed in personal terms, but the issues were essentially economic and political. The struggle was for supremacy over the richest gold mines in the world, and a leading role was played by the most important mining houses on the Rand. While strategic calculations played a role, essentially the Jameson raid and South African War (Boer War) were designed to eliminate the dangers of foreign intervention in the subcontinent and to substitute a state in the Transvaal that could fulfill the demands of the deep-level mining concerns.

Even before the war was over, Milner had appointed a group of young men from Oxford University to "reconstruct" the Boer states: the most serious grievances of the mine magnates were removed and an efficient bureaucracy was established. Central to Milner's plans was the smooth functioning of the mining industry for political and economic reasons. An acute shortage of African unskilled labour was resolved by the importation of 60,000 Chinese, despite the bitter opposition of white workers, and ambitious schemes were set on foot to reduce the cost of both black and white labour. Tax and pass laws were made more efficient and settler agriculture was improved. By 1906–07 the British were sufficiently confident of the new order they had established to grant self-governing institutions to the conquered territories. Although much British propaganda before and during the war had been concerned with the political rights of British subjects, regardless of colour, in neither of the colonies were these extended to blacks.

ANGOLA AND MOZAMBIQUE

The discovery of minerals in the southern African interior in the late 19th century was accompanied by a reassertion of Portuguese colonial ambitions, although a more active colonial policy predated the scramble for Africa by a few decades.

Attempts at plantation agriculture

From the mid-19th century attempts were made to foster plantation agriculture, and a certain amount of Portuguese capital began to enter the colony. Land grants were made in the Luanda hinterland, and planters experimented with coffee, cotton, cacao, and sugar, using the slaves who could no longer be exported. In the absence of an adequate administration or communications network, the plantations in Angola were never highly successful, although coffee cultivation spread among African peasant farmers in the region. The appropriation of land for plantations was resisted, and Portuguese attempts to expand their colonial nucleus led to a series of wars with African peoples, followed by famine and epidemics. The instability of the last decades of the 19th century paved the way of the colonial period to follow.

Portuguese attempts to develop Mozambique along capitalist lines met with even less success, given the lack of investment and the disorder that prevailed in the 19th century. This was exacerbated in the last decades of the century by the escaped slaves, soldiers, and porters who formed bandit bands in broken country in both Angola and Mozambique and attacked Portuguese settlements and African villages alike. In many areas domestic slavery underpinned the migration of young men to the labour markets of the south.

Liberal governments in Portugal from mid-century were anxious to outlaw the feudal aspects of the *prazo* system, but they were unsuccessful despite four military campaigns. The declaration in 1880 that the *prazos* were crown property brought little change. Until 1895 the Gaza kingdom controlled a wide area between the Komati and Limpopo rivers, while the Portuguese had little authority over Africans in the rest of the territory. The only bright spot in Portuguese fortunes in southeastern Africa was the growing prosperity of Delagoa Bay, as trade with the Transvaal increased. In 1875 Portuguese rights to the bay were formally recognized and, with the discovery of gold on the Witwatersrand, the bay acquired a new importance as its closest outlet. In 1888 Lourenço Marques became the capital of the territory, and the focus of Portuguese colonization shifted south.

Portugal's colonial aspiration

The late 19th-century scramble for African territories threatened even those territories the Portuguese considered theirs by historic right. The mineral discoveries, and knowledge that there was copper in Shaba, revived Portuguese dreams of linking their African territories. But

Portugal received short shrift from the other imperial powers. At the Berlin West Africa Conference of 1884–85, Portugal received the Cabinda exclave and a portion of the left bank of the Congo—considerably less than it had claimed. In 1886 the Portuguese ceded the Cunene–Okavango frontier between South West Africa and Angola, hoping to gain German diplomatic support for their colonial ambitions. Portugal fared even worse in the hinterland of Mozambique: with the declaration of the British Central African Protectorate and the expansion of the British South Africa Company in Southern Rhodesia, its economic hinterland was in the hands of the British. The instability of Portugal led the British and Germans to draw up a treaty to partition the Portuguese colonies between them should it be forced to give them up, and until World War I this remained a possibility.

Portugal nevertheless acquired about 800,000 square miles (2,000,000 square kilometres) of African territory, of which it controlled about one-tenth. In both territories "pacification" became a sine qua non of economic development, and there were military campaigns or police actions in almost every year between 1875 and 1924, a measure of Portugal's weakness as a colonial power. The greatest resistance came from those people with the longest experience of Portuguese rule and with the necessary firearms. In Angola the major campaigns were against the Kongo, Mbundu, and Ovambo peoples; in Mozambique, in the Zambezi Valley, and among the Islāmized Makua and Yao people. It was only the war against the already weakened Gaza kingdom in southern Mozambique that achieved rapid and dramatic success.

The majority of Portuguese troops in both territories were black, a situation that turned every campaign into a potential civil war. This, together with the fragmentation of political authority, the resistance of traditional elites who were threatened by colonial rule, and the precipitate introduction of tax and forced labour policies, made resistance in the Portuguese colonies the most prolonged in early 20th-century Africa.

The colonies were of particular importance as a market for Portugal's textiles and cheap wine, and tariff barriers were erected to protect Portuguese manufactures. Starved of capital and wracked by financial crises, Portugal planned to develop the colonies by attracting immigration and foreign capital. In both colonies, considerable tracts of land were granted to chartered companies with wide mining, agriculture, and commercial concessions. In Mozambique these companies, based on the old *prazos,* controlled more than half of the colony. None brought the anticipated level of development, and in both colonies it was the recruitment of migrant labour for South African, Rhodesian, and German enterprise that provided revenue for tax and trade.

THE GERMANS IN SOUTH WEST AFRICA

The Germans were the last imperial power to arrive in Africa, annexing South West Africa in 1884. An Anglo-German commission settled the frontiers of the new territory in 1890. The annexation signalled a change in Germany's colonial policies, accelerated the scramble for southern Africa, disturbed British hegemony in the region, and aroused fears of an alliance between Germany and the South African Republic.

nker frikaner

A vast, arid region, South West Africa was sparsely populated by pastoral Nama and Herero people in the south and centre and by the more densely concentrated agricultural Ovambo, who straddled the frontier with Angola in the north, as well as by scattered San and Bergdama (Damara) hunter-gatherers. Throughout the 19th century communities of Khoikhoi and Oorlam (mixed Khoikhoi-slave groups displaced from the Cape by colonial expansion) made their way into South West Africa, increasing the competition for water and grazing land. Initially, small groups of Oorlams settled peacefully on land granted them by the local populace, some of them establishing mission communities. Jonker Afrikaner, however, who brought his well-armed followers from the Orange River in the 1830s, significantly altered the political balance of power in the region. Responding to an appeal for assistance from the Nama, who were being pushed out of their grazing lands by Herero expansion, he settled at Windhoek. Gaining control over the all-important trade routes from Walvis Bay and the Cape Colony, he ensured Nama dominance over the Herero until his death in 1861.

Conflict between the Nama and Herero was exacerbated from the mid-19th century by the increasing tempo of the cattle and ivory trade from the south. As both sides acquired increasing numbers of guns, the conflict between them became even more destructive, and, apart from a breathing space between 1870 and 1880, it continued from 1863 to 1892, opening the way for German annexation and pacification.

Nor were the Ovambo to the north immune to the effects of the accelerated cattle and ivory trade. Initially trading in salt, copper, and iron from the Etosha Pan region to the north, and supplying hides and ivory to Portuguese traders, they had been largely able to stand outside the slave trade that ravaged their more populous neighbours. In the mid-19th century the volume of the ivory and cattle trade increased, only to collapse as the elephants were exterminated in the 1880s. With the firearms they had acquired through the trade, the Ovambo raided for cattle and people, and new warlords further destabilized the region.

Hendrik Witbooi

The Germans initially hoped to exploit the territory through a concession company, but it proved incapable of raising the necessary capital. The German government was increasingly forced to intervene, especially when settlers appropriated Herero cattle and grazing lands. The most formidable opponent of the Germans was Hendrik Witbooi, the Nama chief who tried unsuccessfully to unite the Herero and Nama against the Germans. He was defeated only in 1894, after a lengthy guerrilla war.

An outbreak of rinderpest that decimated African cattle, the alienation of the better watered highlands, unfair trading practices, and increasing indebtedness, together with the absence of legal redress, led in 1904–07 to an uprising of the Nama and Herero peoples. They were ruthlessly crushed: the Herero population was reduced from 70,000 to 16,000, many of them dying in the desert in an attempt to escape. The Nama were reduced by two-fifths. Although South West Africa was then open to European exploitation, the handful of settlers had to look for labour to the Ovambo in the north (only formally brought under colonial rule in 1915) and to the British territories. German rule lasted until World War I, when South West Africa was conquered by South African troops. After the war, South West Africa was transferred to South Africa as a mandate of the League of Nations.

The political economy of southern Africa, 1910–45

By the beginning of the 20th century new frontiers had been created that became the boundaries of modern states. The penetration of capital throughout the subcontinent wrought vast changes, though north of the Zambezi and in the Portuguese territories they were initially less dramatic. The exploitation of minerals, the capitalization of settler agriculture, and the establishment of manufacturing industries drew Africans into the world economy as workers and peasants, transforming old and creating new classes and political alignments.

Although everywhere the majority of Africans on the subcontinent lived by farming, south of the Limpopo the century-old battle over land had been won by settlers, though increasingly crowded pockets of land were reserved for Africans in the Transkei, Zululand, and the three protectorates that became High Commission Territories (Lesotho, Botswana, and Swaziland). In the reserves and protectorates government made use of African political organization, creating "tribes" where these did not exist and governing through compliant indigenous chiefs and headmen. The priority of every colonial administration was to raise taxation, which brought in revenue and pushed out labour, already a well-established expedient in South Africa. By 1910 Africans north of the Limpopo were all subject to taxation and inroads were being made on their lands.

Union of South Africa

In the south, where this process had gone furthest, settlers increased their hold through the unification of the South African colonies, creating, by 1945, the most highly industrialized state in Africa. The Union (later Republic) of South Africa, formed in 1910, has dominated its neighbours for much of the 20th century. International capitalist expansion northward was paralleled by vast labour migration to the mines and farms of South Africa.

In the early 20th century by far the most dramatic demand for labour came from the mines of South Africa, and it was there that much of the subsequent pattern of southern Africa's industrialization was established. In the 19th century a migrant labour system had evolved in the diamond and gold mines of South Africa, the result of struggles between traditional ruling groups and the new mine magnates over the labour power of young men. This was expensive so long as Africans had access to land and a state sufficiently powerful to create a proletariat and control desertion was lacking. By the end of the 19th century a nexus had grown up to control the work force through the compound system (first established to bring down labour costs at Kimberley between 1885 and 1887), pass laws, and masters' and servants' legislation. Similar institutions came to characterize the mining industries to the north, and the advantages of a work force of unorganized migrant labourers whose welfare and reproduction costs were borne in the reserves became more clear.

Demands for African labour on the Witwatersrand went far beyond the frontiers of South Africa. In 1889 the Chamber of Mines was formed in the Transvaal with the object of controlling African wages. It formed two major recruiting organizations to monopolize the labour supply, and by the early years of the 20th century its recruiters had spread north of the Zambezi.

In all the territories the economic exploitation of blacks was facilitated by the development of a racist ideology. Discriminatory policies were prompted by settlers' fears of the growth of African class consciousness and competition; they were lent respectability by anthropologists and paternalist administrators anxious to preserve African societies from rapid social change. Though "native policy" in the Portuguese territories was theoretically designed to assimilate Africans to a Portuguese way of life, obstacles to progress for the Afro-Portuguese and acculturated elite were enforced more in the 20th century than they had been in the 19th.

SOUTH OF THE LIMPOPO

With the granting by 1907 of self-governing institutions to the conquered Transvaal and Orange Free State, the way was open for the unification of the South African colonies. In 1910 the Union of South Africa was formed, dictated largely by the mineral revolution. Weak, with their economy dominated by foreign capital, 1,250,000 enfranchised whites nevertheless achieved political control over 4,250,000 Africans, 500,000 Coloureds, and 165,000 Indians.

High Commission Territories

There was talk of incorporating the landlocked High Commission Territories of Basutoland (now Lesotho), Bechuanaland (now Botswana), and Swaziland into the union. The opposition of their African inhabitants and of the humanitarian lobby, as well as Britain's desire to maintain a toehold in the region, however, prevented this. In a schedule to the Act of Union, conditions for their ultimate inclusion in the union were set out. This included the continued nonalienability of African lands, although in Swaziland in 1907 the British government confirmed the grant of one-third of the land to settlers and speculators on the basis of temporary concessions granted by the Swazi king in the late 19th century. Only Swaziland had a substantial number of white settlers.

Until the mid-20th century both Britain and South Africa assumed that the protectorates would ultimately become part of the union. Although this did not happen, the territories were fully locked into South Africa's regional economy through the migrant labour system and highly discriminatory tariff policies. All three territories, which had been grain and cattle exporters at the turn of the century, became increasingly dependent on the South African labour market, especially after measures were implemented by the South African state to protect white farmers. Administrators were often South African, and the form of indirect rule practiced left little room for political progress. This dual administration, as well as the dependent economies, was severely castigated by the Pim Commission of 1934–35, but despite modest reforms the territories remained poor and neglected.

Acquisitio of South West Africa

After World War I the union acquired South West Africa. Despite a League of Nations mandate that the territory be administered as a "sacred trust" for its indigenous inhabitants, white settlement was encouraged and subsidized, and the Africans never regained the land lost to the Germans. In 1922 the Nama and Herero were confined to reserves and subjected to police checks and pass laws. Although their reserves were large, they were located in the dry hinterland of the Kalahari or on sandy veld unsuited to man or beast. By the 1930s some 30,000 settlers (mainly from South Africa) controlled all of the potentially fertile, watered land, more than one-third of the total.

The main concern of the mandatory power was to foster international mining capital, which dominated the economy, and to subsidize white settlement. The Permanent Mandates Commission of the League of Nations constantly criticized the lack of economic and educational development for Africans and the violent confrontations between the South African government and the Africans. This had little effect and culminated in the withdrawal of the mandate by the United Nations in 1946, with even less effect.

White politics in the union

In the early 20th century international mining capital dominated the political economy of the union as well, though it had increasingly to come to terms with commercial farming and the growing manufacturing sector. The white population was divided by both class and ethnicity, and the years after union were turbulent, with major strikes by white workers on the Rand and a civil war between Afrikaners when South Africa joined Britain in World War I. Afrikaner nationalist feeling had been aroused in the aftermath of the South African War and was deliberately fostered in the years that followed by the Dutch Reformed Church and Afrikaner intellectuals. Poorer Afrikaner farmers and the northern petty bourgeoisie, as well as English-speaking workers, felt their interests were ignored by the South African Party (SAP), which held parliamentary power under the leadership of Louis Botha (1910–19) and Gen. Jan Smuts (1919–24). The SAP's essential aims were to promote capitalist farming and mining interests, maintain the imperial connection, and reconcile English and Afrikaans-speaking South Africans. In 1913–14, Gen. J.B.M. Hertzog broke away from the SAP to form the Afrikaner National Party.

Native Lands Act of 1913

For white farmers, mining capital, and industrialists, the crucial issue continued to be labour. In 1910 about 40 percent of the African population lived on reserves and about the same number lived on white-owned farms, either as labour- or rent-paying tenants or as sharecroppers. The mining industry wanted the reserves as the home base of migrant labour, while white farmers demanded the reserve lands for their own use, the elimination of competition from African peasant producers, and the reduction of various forms of tenancy to labour service. These objectives were reconciled in the Native Lands Act of 1913. Under this and the supplementary Native Land and Trust Act of 1936, about 12–13 percent of the land was scheduled for more than 4,000,000 Africans, while 87 percent was reserved for 1,250,000 whites. In the interests of the mining industry, the existing reserves were recognized, while Africans could only remain on white-owned land as labour tenants, subject to the masters' and servants' legislation. The land acts were part of a battery of legislation aimed at assisting and subsidizing settler agriculture at the expense of peasant production.

White working class

Maintaining the reserves and improving agricultural production was one way of reducing mining costs; another was by reducing the high wages of skilled white workers, who came increasingly to play a purely supervisory role over Africans, while the Africans replaced the whites at lower pay. In 1907, 1913, and most seriously in 1922, white miners struck work. In defense against their structural insecurity, they did not attempt to join hands with black workers, but demanded a colour bar that protected certain jobs for white workers only. Initially formulated by the Milner administration to reconcile white workers to the importation of Chinese labour, this was formally established under the Mines and Works Act of 1911 and its amendment in 1926. The years immediately after World War I saw a rising tide of worker militancy among both blacks and whites, the result of wartime inflation and postwar depression, with a series of strikes and the formation of the South African Communist Party in 1921. In 1922, at a time of falling gold profits—and, in part, in response to a massive strike of black mine workers in 1920—the Chamber of Mines challenged the job colour bar. White miners struck, martial law was declared, and a five-day battle between white workers and troops on the Witwatersrand ended with 230 dead and the ringleaders hanged.

The 1922 strike brought to a head growing disaffection with the South African Party, which had merged with the Unionists, the party of big capital, in 1920. In 1924 an alliance of Hertzog's Nationalists and the predominantly English-speaking Labour Party removed Smuts from office. Between 1924 and 1929 the "Pact" government made a number of concessions to white workers and incorporated them into the state through industrial conciliation machinery. This largely removed the political danger from white miners. Through raising tariffs and adding a nationalized Electricity Supply Commission to the state-run Iron and Steel Corporation established by Smuts, the Pact government promoted the interests of the manufacturing industry while attempting to resolve the problem of white unemployment.

For its nationalist constituency, Afrikaans was recognized as the second official language of South Africa (replacing Dutch), and a new flag and national anthem were established after great acrimony. At the same time, partly as a result of Hertzog's efforts, the nature of the connection with the British Commonwealth was redefined in 1926 and 1930.

Although the Nationalists won the 1929 elections, they were gravely weakened by the gold standard crisis in 1933. Hertzog formed a coalition government with Smuts and in 1934 their parties fused to form the United Party. South Africa's departure from the gold standard and the enormous rise in the price of gold enabled the new government to use the taxation from gold profits to foster secondary industrialization on an unprecedented scale, to subsidize white farming, and to largely resolve the "poor white" problem.

Poor
hites"

From the beginning of the 20th century the stream of black migrants making their way into the towns was paralleled by a stream of Afrikaners displaced from the land by capitalist agriculture. By 1930 one in five was classified as "poor white." Unskilled and semiliterate, they were in direct competition with Africans, some of whom had more industrial skills. Although their plight was frequently attributed to their refusal to do "Kaffir" work (a derogatory term for manual labour usually done by Africans), poor Afrikaners were at a double disadvantage in the towns. Unlike Africans who had some access to the land, they were totally dependent on their urban wage, while they lacked the skills of English-speaking workers.

The Pact government's "civilized labour" policy—which made government tenders to private industry dependent on the employment of white labour—had more effect in spurring capital-intensive manufacturing and the employment of cheaper, female Afrikaner operatives than in eliminating white poverty, though blacks were replaced by Afrikaner males in government services. A huge rise in gold tax revenues enabled social welfare services to be implemented for whites, while the expansion of manufacturing during World War II largely eliminated white poverty.

rban
acks

Urban blacks were not nearly so fortunate. By 1921 their numbers had risen to 500,000; they increased sharply to 1,250,000 by the late 1930s and to 2,000,000 by the end of World War II. Until the war years their welfare needs were largely ignored by the state, though poor health and the hazard this posed to the reproduction of the work force and to whites aroused concern. Through segregation and influx control the state hoped to restrain African urbanization and tie down farm labour. Increased manufacturing and the rapid development of unions among Africans improved African wages in the early 1940s, and there was talk of some form of stabilization and social reform. In general, however, wages remained low, and housing and health conditions appalling.

Purified
Nationalist
Party

The outbreak of World War II ruptured the fragile consensus of white political parties. As in World War I, Afrikaners were divided, and Hertzog resigned when the parliament voted by a narrow margin to join the Allies. Even before this, in 1934, a group of nationalists supported by the Broederbond—a secret society formed to advance Afrikaner economic interests and ultimately capture the South African state—had left the United Party to form the Purified Nationalist Party. Though their electoral support remained small in the 1930s, they built up their power through the establishment of extensive cultural, welfare, and economic subsidiary organizations directed, on the one hand, at promoting the interests of the Afrikaner petty bourgeoisie and, on the other, to winning over the Afrikaner poor from the predominantly Labour- and Communist-dominated trade unions.

During the war General Smuts, world statesman and for the second time a member of the British war cabinet, headed the government, supported by English-speaking whites. Divisions among Afrikaner nationalists and the adherence of certain factions of the party to fascist-type extra-parliamentary, paramilitary organizations masked the fact that the Purified Nationalists could potentially gain a parliamentary majority by mobilizing ethnic power.

NORTH OF THE LIMPOPO

Though the Union of South Africa had by far the most powerful and complex settler polity, in the first half of the 20th century settlers also acquired political control in Southern Rhodesia (now Zimbabwe). Even in Nyasaland (Malawi) and Northern Rhodesia (Zambia), it was accepted that settlers would provide the necessary economic development under the aegis of international capital, with Africans providing cheap labour.

Southern Rhodesia. The early years of the century saw intensified recruiting of African labour for the hundreds of small mines working scattered deposits. The Rhodesian Native Labour Recruiting Bureau drew in low-paid coerced labour from much of Northern Rhodesia, Mozambique, and Nyasaland to undercut local labour. The dubious profitability of mining in Southern Rhodesia meant that wages, food, housing, and health conditions were cut back ruthlessly and mortality and morbidity rates were exceptionally high.

After the South African War the British South Africa Company (BSAC) turned to the development of commercial corn (maize), tobacco, and cattle farming. Settlers replaced the speculative companies of the early days, and in 1907 they were granted an elective majority in the Legislative Council, first established in 1898. The settlers attacked competition from African producers and their continued access to the land that had been reserved for the Africans at imperial insistence after the uprisings. The colonial office, however, fearful of further disturbances, and the mining industry, with its interests in migrant labour, ensured the continuation of the reserves. By the mid-1920s, 31,000,000 acres of land in Southern Rhodesia were in the hands of about 11,000 whites, 22,500,000 acres had been reserved for the 676,000 Africans, and 45,500,000 acres remained open to purchase and occupation by people of any race.

Land
Apportionment Act
of 1930

The Land Apportionment Act of 1930, modelled in part on the Native Lands Act in South Africa, barred African land ownership outside the reserves, and a special freehold purchase area of 8,000,000 acres was set aside for "progressive farmers." These allocations soon proved wholly inadequate, as African "squatters" were displaced from the white farms and made their way to the increasingly congested reserves.

In 1922, when the BSAC handed over administrative control to the colonial office, the settlers in Southern Rhodesia opted for self-government rather than join the

Union of South Africa, as the company had hoped. The new constitution effectively excluded Africans (95 percent of the population) from the government, and provided an imperial veto, never openly exercised, over discriminatory legislation. Settlers controlled the police and armed forces. Between the 1920s and 1950s, the party of government generally remained closely allied to international capitalist interests, while the opposition represented settler farmers and workers in the face of African competition.

Thus, international capital's fears that self-government would jeopardize their interests proved unfounded. The BSAC continued to own the railways and coal mines, while other base and precious minerals were dominated by a handful of foreign-owned companies. The majority of small, undercapitalized, and struggling white farmers were also initially dominated by the interests of big capital. The Depression and the example of the Union of South Africa led to state assistance both for small mining companies and for white farmers through the establishment of subsidies, systems of control, and marketing boards. These policies had mixed results and were at the expense of the African peasantry and workers, although African corn production increased in the 1930s as rural stratification intensified and falling prices increased self-exploitation. It was only during World War II and immediately thereafter, with the tremendous boom in Rhodesian tobacco, that white farmers began to make money; in 1946 tobacco replaced gold as the most important export. Growing numbers of landless Africans meant that wages in the mines and agriculture fell sharply in the 1930s. White workers were protected by the Industrial Conciliation Act (1934), modelled on South African legislation, while black workers were tightly controlled.

Northern Rhodesia. Across the Zambezi the BSAC's territory of Northern Rhodesia was little more than an appendage of the south. In the northeast, administrative attempts to impose closer settlement and interference with local agricultural techniques interacted with the effects of the natural disasters of the 1890s to produce extremely high morbidity and mortality in the first decades of BSAC rule. These were exacerbated by the forced labour policies of the company. When the BSAC failed to find mineral wealth, Africans had to migrate in search of money for food and taxation to the mines in Shaba (Katanga), Southern Rhodesia, and South Africa. The absence of 270,000 young men recruited as porters during World War I further accelerated agricultural decline. Disease, malnutrition, and inflation followed in the wake of the war, so that by the time the BSAC handed over control to the British colonial office in 1924 the territory was no more than a massive labour reservoir.

Settlers were not initially encouraged in Northern Rhodesia, although a handful of white farmers and traders made their way to the territory. They were granted an Advisory Council in 1918, and in 1924 the BSAC handed over administrative control to the British government, while retaining its mineral rights. In an attempt to increase settler numbers, the colonial office threw open the colony's best lands for white farming, while reserves were drawn up for African occupation. Although extensive, the reserves' soils were poor and they rapidly became overcrowded with people and livestock.

Discovery of copper

It was, however, the discovery and exploitation of rich deposits of copper by U.S., South African, and British capital from the mid-1920s that transformed Northern Rhodesia's economy. Although copper mining was interrupted by the world depression, by the late 1930s Northern Rhodesia had joined the world's major producers, with nearly 90 percent of its export earnings coming from copper. Until 1947 the British government received 50 percent of the taxation on the mines, while the BSAC also exacted a royalty. Thus, despite the importance of the copper industry, the revenue accruing to the Northern Rhodesian government remained small until the boom in copper prices after 1949 and an agreement with the BSAC that gave the local government a share in the royalties. The BSAC's title to mineral royalties was eventually bought out by the Zambian government at the time of independence in 1964.

The inception of copper mining brought the number of white settlers up to 11,000 in 1930, largely skilled miners and farmers from South Africa. Their pressure—together with the reluctance of the British government to confront problems of African urbanization and proletarianization—led until the late 1940s to policies of indirect rule and segregation, with scant attention paid to the needs of urban Africans. The job colour bar, an enormous differential between black and white wages, and a variety of discriminatory practices characterized the Copperbelt in the 1930s and '40s, although by mid-century the development of a skilled and unionized African work force, the determination of the mine companies to cut costs, and the political weakness of the white working class enabled the companies to replace white miners with black.

Nyasaland. In Nyasaland, too, though the principle of African paramountcy was frequently proclaimed, British imperial and settler interests prevailed and led to the territory's depressed economy. In the early colonial period a three-sector economy developed: the ecologically marginal north soon came to act as a labour reserve; in the Shire Highlands an enclave plantation economy was established by settlers dependent on labour tenants; and in the south and centre a cash-cropping peasantry emerged.

In the north and centre, by the turn of the century the activities of missionaries and the growth of consumer demand, the violent imposition of taxation, and the need to replenish cattle herds decimated by the rinderpest epidemic meant that several thousand Africans left their homes each year in search of work on settler plantations or in neighbouring territories. Elsewhere the process was slower, and until the 1920s taxation, often brutally extorted, was necessary to break the autonomy of African societies and force out labour.

Nyasaland had no mineral wealth, transportation costs were high, and white plantation agriculture in the Shire Highlands remained poor and inefficient until the 1920s and '30s, when tobacco and then tea replaced the earlier unsuccessful efforts to grow coffee and cotton. In 1907 the planters were given membership in the Legislative Council, where they demanded state control over labour. With their appallingly low pay and working conditions the planters, even though they extracted labour from their tenants, were unable to compete with the mines in surrounding territories. It was only the influx of successive waves of Lomwe refugees from Mozambique that solved their labour shortages. Harshly exploitative conditions on the Shire Highlands estates led in 1915 to an uprising directed by John Chilembwe.

John Chilembw

From about 1910 the administration began to encourage African cotton production, notably in the Shire Valley and along the shore of Lake Nyasa, where rice was also grown. By the 1920s Africans in the south were also successfully producing tobacco. Yet a racial bias against African producers, and the poverty of the administration in the interwar period—the result of a massive railway debt, heavy freight rates, and the effect of the Great Depression—stultified peasant agriculture. By the 1930s, because of soil erosion and overpopulation, even areas where cash crops had previously provided an alternative were sending out increased numbers of migrants. And, whereas from 1913 the administration had banned official recruitment for the mines, the expansion of the South African and Southern Rhodesian economies in the later 1930s led to a renewal of systematic migrant labour, despite its destructive impact on village life.

ANGOLA AND MOZAMBIQUE

In Angola and Mozambique, too, the economy was sustained by labour migration, while forced cotton, coffee, and rice production increased the burdens on the African populace. To an even greater extent than in British Central Africa, and until considerably later, overt force was a central feature of colonial rule.

Labour migration

In the 1890s the Portugese government hoped to foster plantation agriculture in Mozambique, but the territory was already linked to the labour markets of South Africa. By 1897 more than half the workers on the Rand came from Mozambique. The Portuguese government controlled the flow, through conventions signed with the

South Africans in 1901, 1903, 1909, and 1928. Capitation brought revenue to the state, while deferred pay ensured the migrant's return, his tax, and his purchase of Portuguese manufactures. In exchange for black labour Mozambique received a set proportion of the Transvaal's railway traffic. In Angola this system had its counterpart in the contract labour sent to São Tomé: when this was terminated as a result of slavery scandals in 1908, the São Tomé planters turned to Mozambique for labour.

Most of the Mozambique labourers going south migrated from the region south of the Save River; in the north and centre of the territory were the concession companies that the Portuguese had allowed in the 1890s. In terms of productive capacity, the most important was the Sena sugar estate in the Zambezi Valley, established by British capital. The company totally controlled economic life in its zone of operations, and local Africans had little choice but to work on its plantations. The remaining companies were largely parasitic, extorting taxation and the crops Africans were already producing by force. Conditions were little short of slavery. The most lucrative export was manpower, with local food and cash-crop production falling on the women and children.

ettler
emands

Despite Portugal's schemes to settle immigrants in its colonies, their numbers remained minute. Many were illiterate and unskilled peasants who had little political power, which remained in the hands of the governor general, the highest colonial representative of the Portuguese government. The loosening of political authority in Portugal itself during the republican period (1910–26) was accompanied by a flurry of activity among settler political groups, some of them in alliance with Afro-Portuguese and with members of the creole elite who were angry at Portuguese maladministration, bureaucratic inefficiency, and corruption. A handful of planters dominated settler politics. They attacked migrant labour, demanded protection against African competition, and opposed the ban on distilling and the abolition of slavery without compensation. In Angola the collapse of rubber prices in 1913 added to settler problems, and many went bankrupt. In northern Mozambique campaigns against the Germans during World War I led to famine, forced labour, and high mortality both in combat and from disease. After the war, however, the colonies attracted new settlers as their economies recovered in response to the rise in world prices for tropical products. In Angola, diamond production in the northwest was an additional stimulus.

ntónio
alazar's
conomic
olicy

With the premiership (1932–68) of António de Oliveira Salazar of Portugal, however, immigration schemes were dropped and strict vigilance was exercised over all political and economic activity in the colonies. Minimal consultative institutions disappeared, and in 1940 there were only 27,400 whites in Mozambique and about 44,000 in Angola. Grand imperial rhetoric accompanied a return to protectionism in order to foster Portugal's needs for cheap raw materials and a closed market. Southern Mozambique was entrenched as a labour reserve for the Rand; elsewhere in the colony, from the mid-1930s, Africans had to produce fixed quotas of cotton. During the war years the demand for other tropical products led to a severe shortage of labour, and in both territories women were drafted to cultivate cotton and rice. Confiscations and assaults were legion, despite a plethora of protective legislation. By 1945 more than 80 percent of Portugal's raw cotton came from Mozambique and Angola.

THE AFRICAN RESPONSE

African peoples who had been drawn painfully into the capitalist-dominated economy of southern Africa and subjected to ever-increasing administrative, economic, and political control did not all experience their subjection in the same way. The 20th century witnessed the rise of new classes, with the emergence of an African petty bourgeoisie and working class in the towns and a considerable degree of stratification in the countryside. Migrant labour both undermined and strengthened the authority of the chiefs, especially in areas where the colonial state was anxious to retain traditional structures for purposes of social control. Side by side with the growth of nationalist movements among the new elite and trade union organization among workers, there was a continuation of royal family politics, a restructuring of ethnic identification, and a resort to millenarian solutions.

Royal family politics. In regions where large centralized states had existed at the time of the colonial takeover, royal politics continued to be of significance. In Barotseland, Swaziland, and Basutoland, where paramount chiefs were recognized by the British, the traditional aristocracy combined with the educated elite to protect their position and demand the redress of grievances. In both Matabeleland and Zululand, where the royal families had been militarily defeated, royalists combined to demand state recognition of the monarchy, while in Nyasaland in the 1930s there was an attempt to create a Tonga "paramountcy" and to restore the Ngoni king. In general, attempts to manipulate tradition were most successful in the climate of "indirect rule" of the 1930s and in areas where settlers were weakest.

South
African
Indian
Congress

Political organizations and trade unions. The experience of white rule for Africans has been longest and most intense south of the Limpopo, where the existence of substantial Coloured and Indian minorities gave an extra dimension to anti-colonialism. In South Africa, between 1906 and 1913, Mahatma Gandhi formed the South African Indian Congress and led the first large-scale nonviolent resistance campaign against anti-Indian legislation. He gained limited success, though restrictions on Indian movement and immigration to South Africa remained in force. After his departure in 1914, however, the militancy of the Indian Congress was lost until after World War II, when younger, more radical groups won power from the middle class that had dominated the organization.

The Coloureds of the Cape and Transvaal were never as successfully mobilized. Though the African Political (later African People's) Organization, founded in 1902, was the first black nationwide political organization, it was soon eclipsed by more vigorous African nationalist movements, and for most of the century the Coloured population was divided and relatively voiceless. The Non-European Unity Movement, founded predominantly by Coloureds in the 1940s, influenced African political thought, however, and numbers of Coloureds came to identify with the Black Consciousness Movement.

In 1912 educated Africans united the local associations that had developed in the late 19th and early 20th centuries into the South African Native National Congress, in 1917 renamed the African National Congress (ANC). It aimed to represent African grievances, overcome tribal divisions, and gain acceptance from whites through self-help, education, and the accumulation of property. Though demands for industrial education, individual land tenure, and representation in Parliament were accompanied by attacks on the pass laws, the colour bar, and the Native Lands Act of 1913, the ANC's methods remained strictly constitutional and appealed mainly to the black petty bourgeoisie.

Such associations had their counterpart in Central Africa, partly because many early nationalists either studied in the leading mission schools of the Union or had worked there. Thus, African National Congresses were established in Southern Rhodesia in 1934, in Nyasaland in 1944, and in Northern Rhodesia in 1949–51. Despite regional differences, the class composition and methods of struggle of these organizations were broadly similar until the 1950s, with South Africa leading the way.

Industrial
and Commercial
Worker's
Union

Despite the greatly increased numbers of Africans in South African industry by the end of World War I, African trade unions were hampered by the pass laws, lack of recognition, and police harassment; strikes were illegal and often were put down with violence. Nevertheless, the period 1918–22 saw a great deal of working-class militancy, and in 1920 Clements Kadalie, an immigrant from Nyasaland, founded the Industrial and Commercial Workers' Union (ICU). Based initially among dockworkers in Cape Town, it spread rapidly as a mass movement in towns and in the countryside, where the evicted responded with millenarian zeal to its message, arousing fear and hatred among white farmers. At its height it claimed 100,000 members and had branches in Southern Rhodesia South West Africa. Internal divisions, the mismanagement of finances, and the expulsion of its most active Communist members led to its disintegration by 1929. By that time the Communist Party was beginning to organize Africans in industrial unions; there were some 20,000 unionized Africans by the beginning of World War II and 158,000 in 1945. In the early years of the war the unions were able to win considerable gains for their members, but these were lost after the crushing of the African mine workers' strike of 1946 and the victory of the Afrikaner Nationalists in 1948.

From the early 1920s, to preempt political militancy, the South African government attempted to provide channels for the expression of African grievances through a variety of local consultative councils. This example was followed in Central Africa in the 1930s, where the authority of the chiefs was even stronger. Despite their powerlessness, educated Africans participated in them in the hope of gaining concessions from the state.

In Angola and Mozambique there were few political rights, apart from the republican period (1910–26), when political organizations, trade unions, and the press flourished. It seemed as if Africans and settlers in Angola would strive for similar reformist goals, but the Africans broke away to form organizations demanding limited welfare and educational benefits and to publicize black grievances, though they remained dominated by middle-class moderates. Crushed even before the advent of Salazar in 1932, they revived as ostensibly social and educational organizations in the 1930s and '40s. It was only in the 1950s that they returned to overtly political expression.

Role of Christianity. Even at their height, political organizations and trade unions never reached more than a fraction of the African population. By the beginning of the 20th century, however, parts of South Africa had experienced almost a century of Christian missionary presence. The Roman Catholic Church revived its presence in Angola and Mozambique, where it had a far longer history (it had largely disappeared by the late 19th century when Baptist and other Protestant missions had begun work there). In the early days of colonialism Christianity tended to advance rapidly among the disaffected and dispossessed, and it was usually only after a major disaster undermined the authority of traditional religion that considerable numbers turned to the new religion.

In general, the missionary heritage was ambiguous; by deliberately inculcating the individualism of laissez-faire capitalism and encouraging the stratification that was to lead so many of their converts onto the colonial labour markets, they simultaneously undermined African societies while championing them against more aggressive settler policies and enabling individuals to acquire the skills with which to survive in the white man's world. Frequently critical of settler colonialism, they also attacked much that was valuable in African tradition and developed an ideology to accompany colonial subordination.

Thus many Africans turned instead to the independent churches that first grew up in South Africa after the mineral discoveries and spread rapidly throughout the subcontinent. The churches ranged in size and structure from small family groups to vast organizations with links to black churches in the United States.

In the early years of the century Europeans in southern Africa interpreted the independent church movement as politically sinister and in some areas they sought to suppress it. By and large, however, it probably diffused African political energies. Though the initial break with a mission church betokened a desire for independence from whites, there were many motives for separatism. Independent churches and prophet movements drew on traditional religious and cultural beliefs, offering hope to a sorely pressed and poverty-stricken populace.

Southern Africa since 1945

When, after World War II, the imperial powers turned their attention to their African colonies, it was in the context of a changed world. Great Britain and France were weakened economically and under international pressure to decolonize. The founding of the United Nations, the Bandung Conference of African and Asian nations (1955), the liberation of South Asia, and events in West Africa led Britain to prepare for black independence.

The transfer of power in southern Africa to an African majority was greatly complicated by the presence of entrenched white settlers. Although Nyasaland (as Malaŵi), Northern Rhodesia (as Zambia), and the High Commission Territories (as Botswana, Lesotho, and Swaziland) moved to independence relatively peacefully, the 1960s and '70s saw escalating wars of liberation elsewhere in the subcontinent. By 1980 only South Africa and Namibia remained under minority rule. South Africa increasingly experienced violence and internal opposition; the dissident African National Congress, banned in South Africa from 1960 to 1990, carried on operations underground and outside of South African territory. In response to the transformed geopolitical situation, South Africa massively increased its defense expenditures and waged a war of destabilization against its neighbours. Namibia, however, gained its independence from South Africa in 1990.

South Africa and Namibia

The states of southern Africa remained, despite their political independence, embedded in a regional economy consisting of areas of high capital investment and rural areas of increasing impoverishment. The South African economy accounted for about 77 percent of gross regional production in the early 1980s. Attempts by surrounding states to break this dominance through, for example, the Southern African Development Coordination Conference, were strengthened by the end of the war in Zimbabwe (formerly Rhodesia) in 1979–80, though immense social, economic, and political problems remained.

For coverage of related topics in the *Macropædia* and *Micropædia,* see the *Propædia,* sections 945, 96/11, and 978, and the *Index.* (Sh.M./Ed.)

ANGOLA

Angola is situated on the southwest coast of Africa, with an area of 481,350 square miles (1,246,700 square kilometres) and 1,025 miles (1,650 kilometres) of Atlantic coastline, stretching from the Congo River estuary in the north to the Kunene (former Cunene) River estuary in the south, with a short stretch north of the Congo in the Cabinda enclave.

Independence

For 500 years a Portuguese colony, Angola became independent on November 11, 1975, as the People's Republic of Angola (República Popular de Angola), by proclamation of the Movimento Popular de Libertação de Angola (MPLA; People's Movement for the Liberation of Angola), one of three rival nationalist groups that had engaged in armed resistance since the early 1960s. The nation is bounded to the north and east by Zaire, to the southeast by Zambia, and to the south by Namibia.

Angola was recognized as a colony of Portugal at the Berlin West Africa Conference of 1884–85. The Cabinda enclave, which is 20 miles north of the Congo River and bounded by the People's Republic of the Congo to the north and Zaire to the east and south, was added to the colony by the Berlin conference and is an integral part of the People's Republic of Angola. The capital is the port city of Luanda.

The People's Republic of Angola became a member of the Organization of African Unity in February 1976 and of the United Nations in December.

Physical and human geography

THE LAND

Relief. The land rises from low coastal plains to an interior plateau. The Lunda Divide forms a watershed separating north- and south-flowing rivers.

The land may be divided into four zones, the flat coastal zone and three plateau zones. The principal plateau, the *planalto central,* lying from 3,500 to 4,500 feet (1,100 to 1,400 metres) above sea level, forms part of the great southwest African plateau and covers three-fifths of the total area of Angola. The higher tablelands (with altitudes up to 7,500 feet) include the Malanje (former Rand) Plateau in the north; the Benguela, Bié, Huíla, and Lunda Divide plateaus in the central regions; and the Humpata Highlands and Serra da Chela in the south. The uplands slope down to the coast in a series of escarpments joined by level areas. In the east they dip gradually toward the drainage systems of the Congo (Zaire) and Zambezi rivers. In the south, the Moçâmedes Desert stretches inland from the sea to the Serra da Chela, crossed by seasonal rivers that are dry for most of the year.

Soils. The coastal belt of Angola belongs to the partly Tertiary formation that follows the African coast from the Gulf of Guinea to the Cape of Good Hope. The southern zone generally consists of Recent superficial deposits of dunes and alluvia. The west of the *planalto central* has a crystalline bedrock, and mineral deposits lie close to the surface. The east, except for the Cassange Depression in the north, is mainly buried under relatively recent sand formations. In the coastal area are large concentrations of granite, crystalline limestones, and gneiss, as well as stretches of quartzite. The Zambezi and Congo areas display sandstone, schist, and quartzite and a wide variety of volcanic rocks. There are diamonds in the gravel and laterite in the Lunda area. In the coastal zone the soil is sandy, merging to desert in the south. Petroleum-bearing structures are found in Cabinda and in the Congo and Kwanza basins south to Benguela. There are limestone and sandstone cliffs as well as sandbars and spits. Alluvial soil along the riverbeds is fertile.

wanza
iver

Drainage. The principal rivers flow either from the plateau westward to the Atlantic or, in the eastern part of Angola, to the north and south. The Kwanza (former Cuanza), which flows into the sea some 40 miles south of Luanda, drains much of the central region. Other important rivers are the Kwando (Cuando) and Kubango (Cubango; the name in Angola of the Okavango), flowing through the southeast and into the Okavango Swamps of Botswana; the Zambezi in the east; and the Kunene (Cunene), which rises in the Bié Plateau and flows south and west for 587 miles inside Angola before reaching the Atlantic Ocean.

Climate. Angola has a tropical climate with two distinct seasons. In general, the average annual temperature decreases with distance from the Equator as well as with altitude; temperatures rise with proximity to the sea. To the south, the cool Benguela Current of the South Atlantic creates atmospheric conditions not conducive to rainfall. Soyo (former Santo Antônio do Zaire), in the north, has an average temperature of 79° F (26° C). In Huambo (former Nova Lisboa), in the *planalto central,* the average is 67° F (19° C). The rainy season lasts about seven months, from October to mid-May. The greatest amount of rainfall—about 70 inches (1,800 millimetres)—occurs in the Mayombe Forest in Cabinda, while the smallest amount—as little as two inches (50 millimetres)—falls on the coast. The northern coast near Luanda receives about 13 inches, Lobito 11 inches.

Plant and animal life. As far south as Benguela the coast is rich in oil palms and mangroves. In the north are dense forests, some of which, especially in Cabinda, contain valuable timber; in the south, toward the Kunene, are regions of dense thorn scrub. Rubber vines and trees are abundant but have been reduced by ruthless tapping; the commonest are various root rubbers, and species of *Landolphia* are also found. Coffee, cotton, and Guinea pepper are indigenous, and tobacco flourishes in several districts. Among timber trees are the *tacula* (*Pterocarpus tinctorius*), which grows to an immense size, its wood being blood red in colour, and Angola mahogany. The bark of the *musuemba* (*Albizia coriaria*) is used in tanning leather. A unique Angolan growth is the tumboa plant, (*Welwitschia mirabilis*), a primitive, low-growing plant of great botanical interest; described as a living fossil, it is peculiar to the Moçâmedes Desert.

The larger plant-eating animals include giant sable antelope, elephant, hippopotamus, white and black rhinoceros, giraffe, zebra, gazelle, and gorilla. Among the carnivores are lion, cheetah, and leopard. The land is rich in bird species and in a wide variety of reptiles, including crocodiles. Butterflies, poisonous spiders, white ants, and locusts are indigenous. Marine life includes the conger eel, the black (or southern) right whale, crustaceans such as lobster and giant prawns, tunny, and many temperate-water coastal fish such as mackerel, sardine, catfish, and mullet. Thirteen national parks and nature reserves have been created, the largest of which are the Parque Nacional do Iôna (about 6,000 square miles) and the Parque Nacional da Kameia (5,579 square miles).

Parks and
reserves

Settlement patterns. *Rural settlement.* The sparsely populated countryside of the north, east, and south contrasts with the more densely inhabited central part of the country. Eighty-five percent of the people are peasant subsistence farmers living in traditional and relatively small villages. In the southern desert, nomadic and seminomadic pastoralists live off their herds. European rural settlements in colonial times were concentrated in the rich coffee lands in the north and in the plateaus of Malanje, the central highlands, and the southern uplands around Lubango (former Sá da Bandeira), where the most fertile farmland is to be found. Since independence, these rural areas have become development poles for state and cooperative farming, as abandoned settler property has been taken over by the state.

Portuguese
colonists

Urban settlement. Urban settlements developed early in colonial history at ports and upriver along the Congo and Kwanza rivers as well as in the Bié Plateau on the trade route to Kasai, Katanga, and what is now Zimbabwe. Modern urban development, however, came after the start of armed struggle in 1961, when Portugal sent more than 40,000 troops to Angola and began a serious colonization program, sending unemployed or underemployed Portuguese families to Angola to populate rural development projects. Many drifted to the towns. A further rapid boom in urban development came in the 1970s after the Portuguese premier Marcelo Caetano opened Angola to foreign investors, ending Salazar's protectionism.

THE PEOPLE

Ethnic and linguistic groups. The largest ethnic group in Angola is the Ovimbundu of the central highlands, who speak Umbundu. The second largest group is the Mbundu, speaking Kimbundu, and spread over the coastal area north of Luanda down to N'gunza Kabolo (former Novo Redondo). The other major groups are the Kikongo-speaking northern Kongo; the Lunda–Chokwe from the east, speaking Chokwe and Lunda; the Nganguela, speaking Nganguela and spread over the east and southeast; the Nyaneka–Humbe from the south, speaking Nyaneka and Humbe; and the Ambo. Of non-Bantu ethnic origin are the San (Bushmen) and the Khoikhoin (Hottentots), speaking click languages. Other small separate groups are the Vatua and Xindonga. Only around one-tenth of Angolans speak Portuguese, and fewer still speak Portuguese as their mother tongue. It is the policy of the MPLA-Workers' Party to encourage development of the national languages, particularly in preliminary primary schooling and adult literacy teaching and later as part of the language course in secondary schools, while preserving Portuguese as the official language of the country. Statistics on the number of white Angolans and persons of mixed blood are not available.

Roman
Catholic
Church

Religion. The People's Republic of Angola is a lay state that guarantees, under its constitution, freedom of religion, conscience, and belief. Five hundred years of Portuguese rule had placed the Roman Catholic Church in a privileged position, constitutionally entrenched and part of the fabric of colonial government and administration. Protestant missions, concentrated in the north and in the central highlands and generally more sympathetic to Angolans' aspirations toward independence, were closely watched by the colonial government, and a number of them, suspected of harbouring subversive nationalist elements, were closed. After independence, the state took over administration of mission schools and hospitals in

MAP INDEX

Political subdivisions

Benguela	12·00s	15·00e
Bié	12·30s	17·30e
Cabinda	5·00s	12·30e
Huambo	12·30s	15·40e
Huíla	15·20s	15·00e
Kunene	16·15s	16·00e
Kwando Kubango	15·45s	20·00e
Kwanza Norte	8·50s	14·30e
Kwanza Sul	10·50s	14·50e
Luanda	8·30s	13·20e
Lunda Norte	8·15s	20·00e
Lunda Sul	10·00s	20·50e
Malanje	9·30s	16·30e
Moçâmedes	15·20s	12·30e
Moxico	13·00s	20·30e
Uíge	7·00s	15·10e
Zaire	6·40s	14·30e

The name of a political subdivision if not shown on the map is the same as that of its capital city.

Cities and towns

Ambriz	7·50s	13·06e
Ambrizete, see N'zeto		
Artur de Paiva, see Capelongo		
Baía dos Tigres	16·36s	11·43e
Baía Farta	12·40s	13·11e
Bailundo	12·12s	15·52e
Bembe	7·02s	14·18e
Benguela	12·35s	13·25e
Cacólo	10·07s	19·17e
Caconda	13·43s	15·06e
Caiundo	15·46s	17·28e
Calunda	12·06s	23·23e
Camabatela	8·11s	15·22e
Camacupa	12·03s	17·30e
Cangamba	13·40s	19·54e
Cangombe	14·24s	19·59e
Capelongo	14·28s	16·20e
Carmona, see Uíge		
Cassinga, see Kassinga		
Catete	9·06s	13·43e
Caúngula	8·25s	18·40e
Caxito	8·33s	13·36e
Cazombo	11·54s	22·52e
Cela	11·25s	15·07e
Chibía	15·11s	13·41e
Chiluage	9·30s	21·47e
Chitado	17·20s	13·54e
Chitembo	13·34s	16·40e
Cuangar	17·36s	18·39e
Cuango	6·17s	16·41e
Cubal	13·02s	14·19e
Cuimba	6·07s	14·38e
Cuíto-Cuanavale	15·10s	19·10e
Damba	6·41s	15·08e
Dirico	17·58s	20·47e
Dondo	9·38s	14·25e
Dundo	7·21s	20·50e
Duque de Bragança	9·06s	15·57e
Folgares	14·54s	15·08e
Foz do Cunene	17·16s	11·50e
Gabela	10·48s	14·20e
Gago Coutinho, see Mbundas		
Ganda	30·02s	14·40e
General Machado, see Camacupa		
Golungo Alto	9·08s	14·46e
Gove	13·27s	15·54e
Henrique de Carvalho, see Saurimo		
Huambo	12·44s	15·47e
Humbe	16·40s	14·55e
João de Almeida, see Chibía		
Kaála	12·51s	15·33e
Kassinga	15·08s	16·05e
Kipungo	14·49s	14·34e
Lândana	5·13s	12·08e
Lobito	12·20s	13·34e
Luachimo	7·20s	20·47e
Luanda	8·48s	13·14e
Luau	10·42s	22·12e
Lubango	14·55s	13·30e
Lucira	13·51s	12·31e
Luiana	17·23s	23·03e
Lukapa	8·24s	20·43e
Luremo	8·31s	17·50e
Luso, see Lwena		
Lwena	11·47s	19·52e
Malanje	9·32s	16·20e
Maquela do Zombo	6·03s	15·07e
Mariano Machado, see Ganda		
Matala	14·46s	15·04e
Mavinga	15·50s	20·21e
M'banza Kongo	6·16s	14·15e
Mbundas	14·06s	21·25e
Menongue	14·36s	17·48e
Moçâmedes	15·10s	12·09e
Munhango	12·12s	18·42e
Mussende	10·32s	16·05e
Mussuma	14·14s	21·59e
Muxima	9·31s	13·56e
Namacunde	17·18s	15·50e
N'dalatando	9·18s	14·54e
Negage	7·45s	15·16e
Neriquinha	15·58s	21·42e
N'giva	17·03s	15·47e
N'gunza Kabolo (Novo Redondo)	11·13s	13·50e
Nóqui	5·51s	13·25e
Nova Caipemba	7·26s	14·38e
Nova Chaves	10·34s	21·17e
Nova Gaia	10·09s	17·31e
Nova Lisboa, see Huambo		
Novo Redondo, see N'gunza Kabolo		
N'zeto	7·14s	12·52e
Paiva Couceiro, see Kipungo		
Pereira de Eça, see N'giva		
Porto Alexandre	15·49s	11·53e
Porto Amboim	10·44s	13·44e
Portugália, see Luachimo		
Quela	9·16s	17·02e
Quibala	10·46s	14·59e
Quilengues	14·05s	14·04e
Quimbele	6·28s	16·13e
Robert Williams, see Kaála		
Roçadas, see Xangongo		
Sá da Bandeira, see Lubango		
Salazar, see N'dalatando		
Santo António do Zaire, see Soyo		
Sanza Pombo	7·19s	15·59e
São Salvador do Congo, see M'banza Kongo		
Saurimo	9·39s	20·24e
Serpa Pinto, see Menongue		
Silva Porto	12·22s	16·56e
Soyo	6·07s	12·18e
Teixeira da Silva, see Bailundo		
Teixeira da Sousa, see Luau		
Tomboco	6·48s	13·18e
Uíge	7·37s	15·03e
Veríssimo Sarmento	8·10s	20·39e
Vila Arriaga	14·46s	13·21e
Vila Macedo de Cavaleiros	11·30s	16·45e
Xangongo	16·43s	15·01e

Physical features and points of interest

Atlantic Ocean	11·00s	12·00e
Benguela Plateau	12·30s	15·00e
Bié Plateau	12·30s	16·30e
Bikuar, Parque Nacional do, *national park*	15·18s	14·45e
Biópio Dam	12·30s	13·50e
Cambambe Dam	9·43s	14·30e
Cameia, Parque Nacional da, *national park*	12·00s	21·45e
Cassai, *river*, see Kasai		
Catumbela, *river*	12·27s	13·29e
Chela, Serra da, *mountains*	16·00s	13·10e
Chicapa, *river*	6·26s	20·47e
Chiumbe, *river*	7·00s	21·12e
Congo (Zaire), *river*	6·04s	12·24e
Cuando, *river*, see Kwando		
Cuango, *river*, see Kwango		
Cuanza, *river*, see Kwanza		
Cubango, *river*, see Kubango		
Cuilo, *river*, see Kwilo		
Cuito, *river*, see Kwito		
Cunene, *river*, see Kunene		
Cuvo, *river*, see Keve		
Dande, *river*	8·28s	13·21e
Huíla Plateau	15·30s	15·00e
Iôna, Parque Nacional do, *national park*	16·32s	12·12e
Kasai, *river*	3·02s	16·57e
Keve, *river*	10·50s	13·47e
Kubango, *river*	18·50s	22·25e
Kunene, *river*	17·20s	11·50e
Kwando, *river*	18·27s	23·32e
Kwango, *river*	3·14s	17·23e
Kwanza, *river*	9·19s	13·08e
Kwilo, *river*	3·22s	17·22e
Kwito, *river*	18·01s	20·48e
Loge, *river*	7·49s	13·06e
Lomaum Dam	12·55s	14·50e
Longa, *river*	10·15s	13·30e
Luando, *river*	10·19s	16·40e
Luangue, *river*	4·17s	20·02e
Luanguinga, *river*	15·11s	22·56e
Lucala, *river*	6·38s	12·34e
Luena, *river*	12·31s	22·34e
Lunda Divide	12·30s	20·00e
Lungué-Bungo, *river*	14·19s	28·14e
Mabubas Dam	8·28s	13·44e
Malanje Plateau	8·00s	16·00e
Milando, Reserva Parcial do	8·30s	34·06e
Moçâmedes Desert	15·30s	12·10e
Môco, Serra, *mountain*	12·28s	15·10e
Mupa ou do Girafa, Parque Nacional da *national park*	15·50s	15·40e
Quiçama, Parque Nacional da *national park*	9·45s	13·45e
Rand Plateau, see Malanje Plateau		
Ruacaná Catarata, *falls*	17·22s	14·12e
Santa Maria, Cabo de, *cape*	13·25s	12·32e
Utembo, *river*	17·06s	22·01e
Zaire, see Congo		
Zambezi, *river*	18·55s	36·04e

accordance with constitutional provisions abolishing payment of fees and private education and health services; clergy employed in these activities became paid government employees. African Christian religions, such as the Kimbanguist Church, continue to flourish, and traditional religions based on worship of ancestral spirits still prevail in the rural environment, though less so in urban areas.

The role of churches and believers in a revolutionary state led by the principles of Marxism–Leninism has been debated. The Central Committee of the MPLA-Workers' Party stated at the first party congress that "the Party will base its policy on the assumption that the struggle for a free, scientific and materialist consciousness is an integral part of the struggle to build a society in which there will be no more exploitation of man by man, a struggle in which it is vital that both believers and atheists take part," and it concluded, "following the guiding principles of Marxism–Leninism, the Party and State of the People's Republic of Angola are not going to prohibit religion." Churches and religious organizations, however, must be registered and are subject to taxation, and in March 1978 the Jehovah's Witnesses were ordered to cease functioning. Formation of new religious organizations has been prohibited.

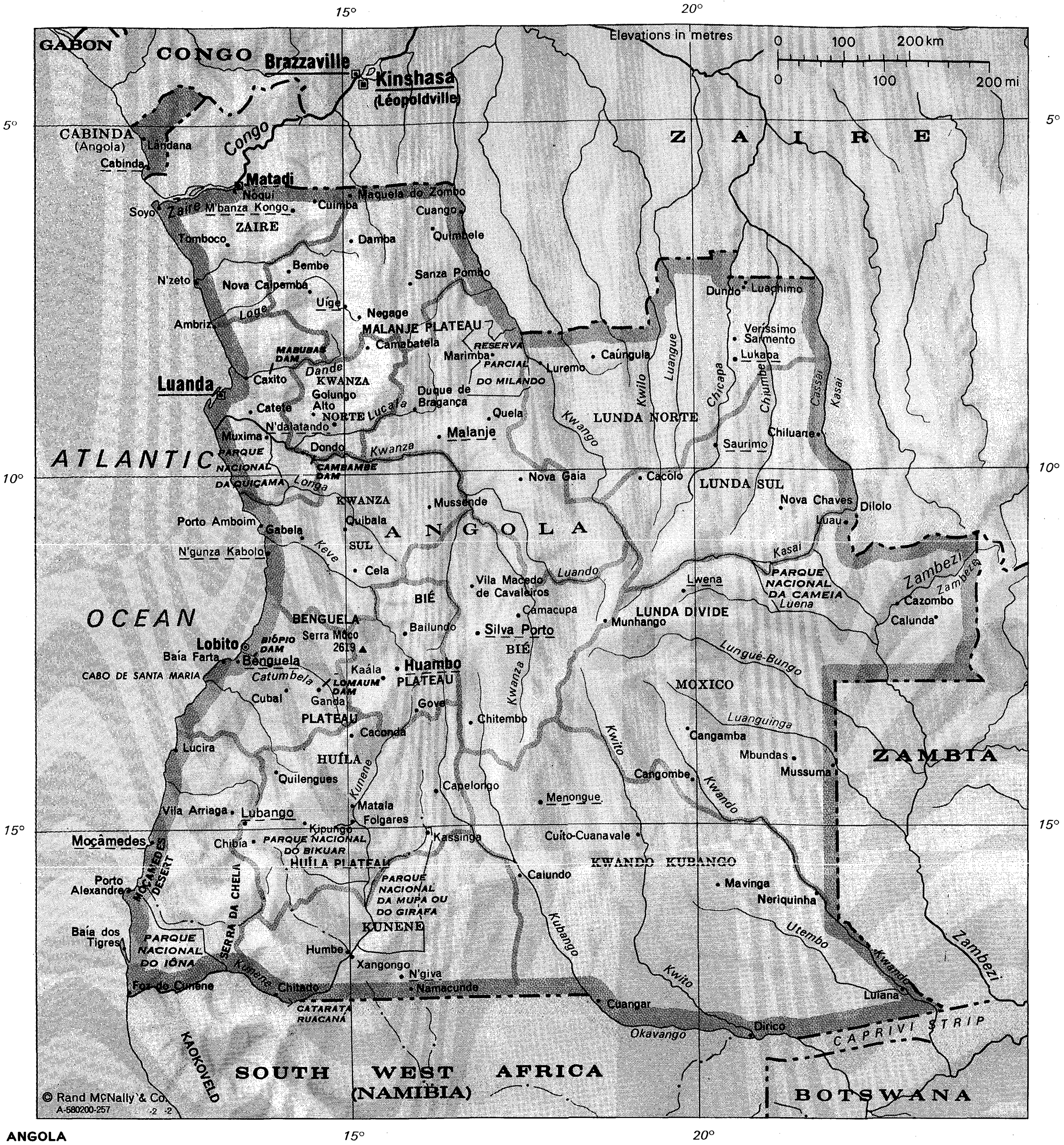
THE ECONOMY

Angola at independence found itself with a severely distorted economy, with only 12 percent of its gross national product (GNP) deriving from agriculture though 85 percent of its working population was engaged in it. The primary sector as a whole, including agriculture, forestry, and mining (prior to exploration of the Cabinda oil fields), in 1970 accounted for only 23 percent of the GNP. The secondary sector (industry and agro-industrial projects) accounted for a mere 14 percent, while the tertiary sector, mainly trade and services such as banking and insurance, accounted for a staggering 63 percent of the GNP—a typical settler-oriented economy. The country's enormous natural resources and soils and climate favourable for agriculture provide extremely promising conditions for development under an independent government. The retardant to economic growth will probably be the needed training and education of Angolans. At independence 90 percent of the population was illiterate.

Nevertheless, the large settler population and the buildup in the 1960s of the secondary and tertiary economic sectors to service it gave Angola at independence a more sophisticated economic structure than most African countries have yet achieved. Severe contrasts and imbalances that need to be corrected include the contrast between the eastern "badlands," where, in an area the size of Portugal itself, total colonial economic development (aside from the Benguela Railway line) consisted in some citrus fruit farms and a handful of sawmills, and Cabinda, where high technology petroleum industry drew industrial, commercial, and trading investments.

Angola is likely to become a leading African petroleum-producing country, along with Nigeria and Libya. Diamond extraction places Angola among the world's major producers.

Resources. Crude oil is the principal source of national income. A state oil company, Sonangol, created in 1977, derives revenues from joint ventures and production-sharing agreements with major international oil companies such as Gulf Oil in Cabinda and Petrofina and Texaco



ANGOLA

in the Congo and Kwanza fields. A law promulgated in 1978 provided for ownership of all crude oil resources by the state and set out broad guidelines for foreign participation in exploitation of oil, natural gas, and other hydrocarbons in partnership with Sonangol. Development of primarily off-shore oil fields has made Angola one of the major sub-Saharan oil producers. Recoverable reserves have been estimated at anywhere between 1,000,000,000 and 10,000,000,000 barrels. Approximately two-thirds of Angola's oil production is from the Cabinda enclave.

ıte
ʼnership

Agriculture, forestry, and fishing. Just under 3 percent of the total land area is arable and more than half of the economically active population works in agriculture. Before independence, even on more efficiently farmed plantations, use of fertilizers and irrigation was extremely rare, and intensive single-crop cultivation by the settler farmers was hard on soils.

After independence and the return to Portugal of the settler farmers, abandoned property reverted to the state and the agricultural sector was gradually restructured into large state farms or smaller cooperatives and peasant associations, formed voluntarily by subsistence farmers or former contract labourers. To improve nutrition, food crop production was extended in all provinces, and cooperative poultry and pig farming were introduced. Cash cropping presented an immediate structural problem: previously based on contract labour—wholly unacceptable in an independent state committed to raising living standards—cultivation of coffee, cotton, sisal, and bananas fell drastically in the years following independence. To compound the problem of restructuring, the infrastructure on the cash crop farms was entirely insufficient to house and provide health and education services for the families of any newly recruited labour force.

od
ɔps

Much of the essential work on the coffee plantations was carried out in the first years of independence by voluntary workers, organized through the MPLA and mass organizations into voluntary brigades from towns and cities nearby. In this way the coffee crop in 1977 reached around 70,000 metric tons. Angola produces more than enough cotton for its textile industry and exports the balance.

In the 1960s Portuguese settlers embarked on extensive ranching schemes, importing pedigreed cattle, sheep, and pigs (mainly from South Africa) and at times crossing them with local breeds to obtain better resistance to disease. Most of the country is free of the tsetse fly, and the southern part has been traditionally a livestock-raising area, particularly Kunene and Moçâmedes provinces, where Angolans live off the meat and sour milk from their cattle, cultivating small quantities of cereals as a dietary supplement. The livestock were decimated during the civil war after independence by indiscriminate slaughter and the driving of cattle over the borders into Zaire and Namibia. Rebuilding of the livestock and ranching industry is given economic priority, and government veterinary services were operating by 1977.

Timber, previously another major foreign exchange earner, was diverted wholly to fill national needs after independence. There are about 129,000 acres of forest in Angola, but the forestry industry depends mainly on Cabinda and the eastern Ganguela and Zambezi forests. Eucalyptus plantations along the Benguela Railway supply fuel for the steam locomotives. The Alto Catumbela forests provide raw material for a major paper pulp plant near Benguela.

Angola has a modern fishing industry, though it was severely affected by the war of 1975–76, when settlers sailed the majority of the fishing fleet back to Portugal or to Walvis Bay in Namibia. The refrigeration and fishmeal plants at Moçâmedes and Porto Alexandre were reactivated, and contracts were signed with foreign fishing fleets to land a proportion of their catch, while technical assistance agreements were made to build a new Angolan fleet and train Angolan crews; in 1976, however, the catch had fallen to only 10 percent of prewar figures. Before independence, industrial firms relying on modern refrigerated fishing vessels accounted for 86 percent of the catch. Main fishing centres are Moçâmedes–Porto Alexandre, Benguela–Baía Farta, Luanda–Cacuaco, and Porto Amboim–N'gunza Kabolo. Freshwater fishing is limited to traditional catching methods for home consumption. Both freshwater and saltwater fish are cured traditionally by drying or, in some areas, smoking.

Mining. Diamond mining is the principal mining activity in Angola. After a drop in production during 1975–77, during which time the state acquired a controlling 61 percent interest in Diamang, the British-dominated company that had had the monopoly on diamond mining (and Portugal retaliated by freezing assets of the company held in Lisbon), production regained the 1973 level of more than 2,100,000 carats during 1978. Black granite and fusing-grade quartz are also mined for export. Mining of manganese, copper, and iron ore had terminated by independence, after foreign mining companies had exhausted deposits that could be mined with profit. Evaluation of remaining lower grade iron ore deposits in Kassinga, however, is being carried on.

Industry. A small heavy industrial sector includes steelmaking, shipbuilding and repairs (at Lobito), paper pulp processing, and production of cement. Textile manufacture from Angolan cotton and imported synthetic fibres, tire manufacture, paints and plastics manufacture, sugar refining, and brewing and production of soft drinks are major industries, and sawmills and plywood mills process locally grown timber. The agro-industrial sector includes canning of fruit, vegetables, and meat, fish processing, production of edible oils, and production of biscuits and pasta from imported wheat, milled in local plants. Tobacco grown in the central regions supplies a cigarette industry. Following the civil war of 1975–76, the government gave priority to a national reconstruction program to recover sabotaged industrial plants, refit them, train Angolans for skilled industrial jobs previously reserved for settlers, and regain production levels of 1973.

Only a fraction of the country's population has running water or electricity at home, and a large rise in consumption is expected, entailing new investment in power generation. Angola's many rivers, as well as its domestic sources of fossil fuels, ensure plentiful supplies of cheap energy. The five most important hydroelectric plants are the Cambambe Dam on the Kwanza River near Dondo; Mabubas Dam, northeast of Luanda on the Dande; Lomaum Dam on the upper Catumbela; Biópio Dam on the Catumbela; and Matala Dam on the middle Kunene. The main Portuguese power company, Sonefe, is now under state control.

Hydroelectricity

Trade. At independence, a major aim was to diversify foreign trade, previously heavily weighted toward Portugal for all export cash crops and diamonds and for imports of general merchandise; and toward the United States, which took virtually all Angolan crude oil and most of the coffee. Sixty percent of exports of crude oil were in the hands of the state oil company, which thus had an opportunity to sell to new customers (though the low sulfur content of Angolan crude makes it ideal for the U.S. market); diamond sales had also been taken out of the hands of Portuguese company interests, though diamonds continued to be sold via the Central Selling Organisation ultimately. Coffee was sold by the state directly on the London commodity market rather than exported from Angola via private foreign brokers as it was formerly.

Imports came from a wide variety of sources. Capital equipment for existing industrial plants continued to come from Western countries.

Administration of the economy. *The public sector.* Economic management is directed by the MPLA-Workers' Party, which set out basic guidelines. Restructuring of the colonial economy proceeded rapidly, with a nationalization law, passed in March 1976, the major instrument for building a strong state-owned sector. Cooperatives were also being built up, principally in agriculture.

Long-term economic development, in the party view, must be based primarily on agriculture and the building up of heavy industry. Petroleum exploitation, building, and fishing have been singled out for development as sectors affording maximum short-term growth prospects. Medium-term objectives include investments in petrochemicals, mining, and phosphate production, the latter to provide fertilizers for agriculture.

Government revenue is derived from state-controlled enterprises and from exports, taxation, and stamp duties.

The private sector. Few non-Portuguese foreign firms were nationalized after independence, though a large number of joint ventures between the state and such firms has been established. Foreign multinational companies operating in Angola include the oil companies, banks, heavy-engineering firms, and computer, electronics, and telecommunications companies. Private manufacturers are concentrated in the clothing and food and beverage industries. The largest number of private entrepreneurs is found in the retail trade, mostly in small businesses, followed by the transport sector, with many small trucking firms.

Multinational companies

Taxation. After independence, a National Reconstruction stamp duty replaced colonial stamp duties. Direct taxation includes income tax, company tax, and building tax on property owners. Indirect taxation is under review, and one rise since independence has been in the indirect tax on gasoline. The constitution provides for progressive direct taxation.

Transportation. *Roads.* The major north–south road, built in colonial times, linked Luanda with South Africa, via South West Africa. A small stretch of unpaved road in South West Africa was the only break in a system that linked Luanda, Huambo, and Lubango (formerly Sá da Bandeira) with Windhoek (capital of Namibia), on to Cape Town in South Africa, and northward again as far as Beira in Mozambique. Coastal roads and major bridges across river estuaries were built by the Portuguese as part of the effort to control the country during the years of armed struggle between 1961 and 1974. They run from Soyo on the Congo estuary south to Luanda and on to N'gunza Kabolo, Benguela, and Lobito. Three main roads run from west to east: one from Luanda and one from Lobito to Luau (former Teixeira da Sousa) on the Zaire border and one from Moçâmedes to Luiana near the Zambian border.

Between July 1976 and July 1977 the state imported around 12,000 vehicles for state transport organizations. Private truck and bus companies continue to exist alongside the state-owned transport system. Municipal transport services, however, are all run by the state, and fares were lowered at independence for suburban areas, where the poorest people live.

Railways. The Benguela Railway, privately owned by a British company, runs 840 miles from Lobito on the Atlantic seaboard to the Zaire border at Luau, where it connects with the Zaire railway system and from there runs through to the Zambian Copperbelt. Both Zaire and

Zambia use the Benguela Railway for exports of copper and for vital imports. Other major railway lines also run east–west and were also built in colonial times for export of raw materials. The Luanda–Malanje line was built for export of cotton and coffee, the Gabela–Porto Amboim line for export of coffee, and the Moçâmedes–Menongue (former Serpa Pinto) line for export of iron ore from Kassinga.

Ports. The chief ports are Lobito, Luanda, and Moçâmedes, followed by Landana, Porto Amboim, Cabinda, Soyo, N'zeto (former Ambrizete), N'gunza Kabolo, Benguela, Porto Alexandre, and Baía dos Tigres.

Lobito has two L-shaped wharves as well as the only floating dock in Angola. The port is fully equipped with storehouses, railway lines and marshalling yards, depots for liquid fuel, and silos for grain. As terminus for the Benguela Railway, it handles Zairian and Zambian copper, Zairian manganese, agricultural products from the central highlands of Angola, and general imports for Angola, Zaire, and Zambia.

uanda
rbour

Luanda harbour is deep and easy of access, with about four square miles suitable for berthing ships. It is linked by rail with the main Luanda railway station and is equipped with up-to-date loading and unloading equipment. It is an outlet for coffee, cotton, timber, and cereals. Moçâmedes, railhead for the Moçâmedes railway, was the outlet for iron ore exports from Kassinga and has a special ore terminal. Moçâmedes harbour also serves an inland region potentially rich in agricultural and industrial products and has refrigeration plants to store meat from Huíla and Kunene. North of Lobito, Porto Amboim serves the coffee region around Gabela in Kwanza Sul province.

River traffic. Few Angolan rivers are navigable. The Kwanza can carry small craft upstream for about 140 miles. The Congo is navigable by oceangoing ships up to Noquí in Angola and to Matadi in Zaire. All Zaire-bound ships must use the Angolan territorial waters of the Congo River, which is not navigable at certain points inside the Zaire territorial water limits.

Air transport. The main international airport, at Luanda, is a port of call for airlines operating between Europe and central and southern Africa. An alternative landing site is Huambo. The Angolan state airline, Taag (Transportes Aéreas de Angola), links main cities and towns. There are airstrips for small planes. A state-owned charter company leases small aircraft and crews.

ADMINISTRATIVE AND SOCIAL CONDITIONS

Government. *Constitutional framework.* The constitution of the People's Republic of Angola, proclaimed on November 11, 1975, and amended in October 1976, provides for "construction of a prosperous and democratic state, totally free of all forms of exploitation of man by man." Leadership falls "to the MPLA-Workers' Party as the organized vanguard of the working class and a Marxist–Leninist party." State property and means of production are managed through central planning, while the right to private property and private activities is recognized in Article 10 "as long as these are useful to the country's economy and are in the interests of the Angolan people." Internationally, the constitution declares adherence to the charters of the United Nations and the Organization of African Unity and declares Angola's wish for relations of friendship and cooperation with all states on the basis of mutual respect and advantage. The highest organ of state power is the People's Assembly, which replaced, in 1980, the provisional Council of the Revolution.

ouncil of
inisters

The government executive is the Council of Ministers, presided over by the state president (who is also president of the MPLA-Workers' Party) as head of the government. Each member of the government is personally responsible for decisions and their execution. A permanent committee of the Council of Ministers acts as executive between meetings of the full assembly.

Local government. The country is divided into 18 provinces, broken down into smaller units of administration. A provincial commissioner coordinates activities of the branches of central government ministries in the provinces, under guidelines from the party. A party steering committee at provincial level, also headed by the provincial commissioner, is nominated by the Central Committee, pending elections once the new party has been fully constituted. The principle of "democratic centralism" governs relations between local government bodies and the central administration.

Foreign relations. Angola is one of the nonaligned countries, and its constitution prohibits membership in any international military organization and foreign military bases on Angolan soil. The constitution also expresses the wish for friendly relations with all countries on the basis of noninterference in each other's internal affairs, mutual respect, and mutual advantage. Angola is one of the five so-called front-line countries confronting domination by white minorities in southern Africa, and the constitution declares Angola's solidarity with peoples fighting for national independence.

The political process. The political struggle for independence from colonial rule was led by the MPLA, from its foundation in 1956. Among the several liberation movements, it alone succeeded in mobilizing a broad front of patriots from all parts of Angola and all social classes. The MPLA's national appeal was based on a conscious struggle against regionalism, tribalism, and racism. Although socialist principles were articulated in the liberation movement's program, it was not until the first party congress in December 1977 that the MPLA constituted itself into a Marxist–Leninist vanguard party of the working class (the Movimento Popular de Libertaçâo de Angola-Partido de Trabalho, or MPLA-PT, MPLA-Workers' Party) and launched a Movement of Rectification to clarify the party line and build new party cells. Members were selected through assemblies at workplaces at which all workers criticized and approved or rejected party candidates.

MPLA-
Workers'
Party

The first congress of the MPLA saw the election of a new Central Committee and Political Bureau, though provincial party committees were nominated rather than elected. The congress itself was elected from among MPLA militants, at grass roots committee level. An extraordinary congress met in 1980, after which regular congresses would be held every five years.

Mass participation in the process of party building was assured through the Movement of Rectification, involving all workers, in and outside the MPLA. As in other Communist countries, party-sponsored mass organizations have replaced former voluntary associations. The party youth organization and the women's mass organization, OMA (Organização da Mulher Angolana, or Organization of Angolan Women), involve young people and women in community organization, self-help schemes, literacy and health projects, and other mass campaigns. Local government issues are debated in mass meetings.

Justice. The Angolan judiciary is independent of the executive branch, according to article 50 of the constitution. A People's Supreme Court heads the judicial system. Collegiate courts administer justice through a system of professional judges and barristers and people's lay assistants. All private law practice has been abolished. A state public prosecutor is responsible for respect of the law by state organs, economic and social organizations and entities, and private citizens.

The colonial legal system is gradually being replaced by new laws, but colonial legislation is regarded as valid until repealed or contradicted. A law passed shortly after independence made divorce possible for the first time. Customary law is still administered in the countryside.

Education. Free education is guaranteed to all citizens under the constitution, though compulsory education extends only to primary schoolchildren.

State
aims in
education

Thorough renovation of the educational system is underway, its aims being to narrow the social gap between intellectual and manual work (by introducing practical manual work into the system), to rid the schools of colonial history and geography, to reinvigorate national cultural values, and to build a classless society. Parallel to the children's educational system, a system of adult education is being built to end illiteracy. Costs were kept low by use of volunteer teachers. The major problem facing the Ministry of Education is a shortage of teaching staff.

Foreign teachers, mainly from Cuba and Portugal, were enrolled and are both teaching in schools and centres of higher education and training Angolan teachers.

Health and welfare. A law passed in December 1975 established free medical care and a National Health Service. In 1977 all private, fee-paying medical services were abolished. Church mission hospitals are now administered by the National Health Service. Many hospitals had to be almost entirely reequipped after the devastation of 1975–76. Training of Angolans as medical and paramedical staff is under way, and foreign technical assistants, mainly from Cuba, are helping train hospital staff.

Widespread malnutrition and lack of sanitary education are principal causes of disease in both town and country. Preventive medicine is the main aim of the National Health Service, and mass immunization campaigns have been undertaken against endemic tropical and other common diseases. A high rate of infant mortality was being attacked after independence by education of pregnant and nursing mothers and establishment of mother and child care centres. African traditional medicine is still widely practiced and is being studied for its continuing values in use of herbs and other natural cures.

Housing. Most Angolans live in traditional wattle and clay or timber houses with separate kitchen buildings and outhouses for storage of food. In the towns there is sharp contrast between the modern villas and apartment and office buildings of the formerly white city centres and the sprawling peripheral shanty towns, or *musseques,* "cities of sand." A national housing program, with priority for rural areas, involves government construction of low-cost housing and a government-run self-help program, with building plans and materials made available to individual home builders.

Cities of sand

Wages and the cost of living. The wage structure inherited from colonial times was entirely reviewed after independence, and top government salaries were cut and a higher minimum guaranteed wage was established. Differences were then gradually narrowed. At the same time, privileges that under colonial rule had accrued to the civil service were either swept away or were extended to all workers (the family allowance is an example of the latter). The principle of "from each according to his ability, to each according to his work" is enshrined in the constitution, as is the right and duty of all Angolans to work. High unemployment in Luanda, a result of the flight from rural areas and the upheavals of war, was attacked by recruiting for rural jobs. To reduce the inflationary price trends that had accompanied the war and the general shortages of food and consumer goods, government price controls on essential food items and consumer goods were imposed, on the principle of a national price system.

Social and economic divisions. The overwhelming majority of Angolans are illiterate, poverty-stricken peasants. Urban workers and the privileged urban elite together represent no more than 15 percent of the entire population. The MPLA-Workers' Party has set about narrowing both social and economic divisions through national programs for literacy, education, health, and employment. A small class of private Angolan entrepreneurs, based in Luanda, Benguela, Huambo, and Lubango, has no significant influence on the national social or economic structures.

CULTURAL LIFE

A National Council for Culture is working to revive and rediscover Angolan indigenous culture while encouraging new creative forms proper to the revolutionary post-independence era. While literature and journalism were instruments of resistance to colonialism (as were music and dancing), traditional sculpture, pottery, weaving, and crafts in general declined in the final years of Portuguese rule. Suppression of Angolan culture was a general objective of the settler regime, which saw in it a threat to Portuguese hegemony and a magnet for nationalist aspirations. After independence a National Museum Directorate established an Anthropology Museum, an Archaeology Museum, a Slavery Museum, and a War Museum and breathed new life into scattered provincial museums, with greater emphasis on anthropology and Angolan history and culture. Popular music has revived, dance bands and traditional music players tour the country, and in 1978 the Luanda and Lobito *carnaval* was revived and taken to many other parts of the country.

The arts. Angolan sculpture of wood, clay, and stone is found in Lunda, Bailundo (former Teixeira da Silva), Kwango, Bié, and Cabinda. Pottery with incised designs and geometrical decoration, as well as woven textiles with geometric patterns and richly decorated basketry are seen particularly in the provinces of Lunda, Bié, and Cabinda.

Most famous is the art of the Lunda–Chokwe peoples, embracing sculpture and pottery and other objects of use, including richly carved furniture. In remote areas, shell necklaces and bracelets, tattooing, and elaborately dressed hair styles, adopted by both men and women, indicate region and tribe as well as social and economic status.

Music

Music has evolved to accompany work and recreation and, like dancing, marks principal events in the year and life cycle. Percussion instruments set the pace, and stringed and woodwind instruments, the portable finger piano or the much larger xylophone, as well as bamboo tubes accompany and provide the melody.

Literature and history are overwhelmingly oral traditions, though a school of modern Angolan writers, growing from the last part of the 19th century, became a focal part of the nationalist struggle as the Angolan intellectuals joined the MPLA in the 1950s. The first president, Agostinho Neto, was a poet of distinction in the Portuguese language.

Cultural institutions. The National Council for Culture has overall responsibility for cultural activities. The national Museum Board supervises museums and took over the most famous museum of the colonial period, a private institution run by the Diamang mining company in Lunda province; many of its treasures were transferred to the capital. The Union of Angolan Authors, formed after independence, publishes works by many Angolan writers and poets, and the National Union of Artists includes graphic artists, painters, and sculptors. Many theatre groups are being encouraged, with a view to both researching theatrical history and presenting new forms and revolutionary theatre. A National Library is based in Luanda. Cinemas exist in all towns, and a mobile cinema unit tours the provinces, under the party Department of Revolutionary Orientation, projecting films outdoors, on whitewashed house walls or other surfaces, in remote areas. Major sports include football, volleyball, basketball, swimming, gymnastics, athletics, and judo.

For statistical data on the land and people of Angola, see the *Britannica World Data* section in the BRITANNICA WORLD DATA ANNUAL. (Ja.Be.)

History

In the late Stone Age the northern part of Angola was populated by groups of hunters and gatherers who left a variety of stone tools. Later, during the 1st millennium AD, these early inhabitants were gradually absorbed by new, Bantu-speaking immigrants who introduced the revolutionary concepts of metalworking and agriculture. The Bantu communities in Angola remained for the most part small and isolated; the main exception to this was in the northwest where the Bakongo people developed a large, well-governed kingdom.

THE PORTUGUESE

About 1483 Portuguese seafarers began to arrive on the coast. They came to trade but soon discovered that Angola had no gold or other precious metals with which to pay for the new luxuries they brought. The only salable commodity was labour, and so within a few years an export trade in slaves developed. These slaves were mainly used to establish sugar plantations on the island of São Tomé and in Brazil.

Slave trade

The slave trade soon dominated events in Angola. At first it increased the wealth and power of the rulers who controlled it. In particular the king of Ndongo (Angola) greatly expanded his territory in the early 16th century. Later the trade caused internal divisions and rivalries that decayed the political fabric of the participating states and invited outside interference and conquest. In this manner in 1568 Kongo fell prey to the depredations of a warrior nation from Central Africa called the Jaga. Although the kingdom was later restored with the help of Portuguese troops it never recovered its former greatness and Portuguese interest and activity passed to Ndongo. Ndongo, however, soon clashed with the Portuguese.

For nearly a century the Portuguese who followed Diogo Cão and other explorers down the African coast were more interested in commerce than in colonial dominion. They took with them missionaries, who taught European concepts and values to their trading partners, but not soldiers. When the Kongo kingdom began to crumble, however, the Portuguese changed their policy and decided to establish in Angola a conquest colony similar to those of the Spaniards in Mexico and Peru. Luanda became their military base, and from 1576, under the leadership of Paulo Dias de Novais, they set out to subjugate the kingdom of Ndongo. In the ensuing wars the contestants proved fairly evenly matched and only after 100 years of intermittent fighting did the Portuguese prove successful. By 1680 their colony consisted of half a dozen forts in the lower Kwanza (Cuanza) Valley and the small outlying territory of Benguela on the southern coast.

Meanwhile, the slave trade continued to flourish, carrying off to the Americas more than 1,000,000 men, women, and children and killing countless others in the process. From the early 17th century, Portuguese traders were joined by those of Holland, England, and France; and during the 18th century in particular these western Europeans heightened the devastation that the slave trade was causing in Angola. The only attempt to relieve the country's decay was made by Dom Francisco de Souza Coutinho who, while governor of Luanda, unsuccessfully tried to diversify the Angolan economy.

The 19th century did not usher in a new era in Angola. The export of labour continued as before until British pressure forced the gradual suspension of the Brazilian slave trade in the 1850s. Thereafter the labourers, no longer called slaves but recruited in the same manner, were diverted to the cocoa estates of São Tomé, Portugal's original plantation island. By the end of the century Angola had suffered from four centuries of continuous slave trading and was as impoverished and underpopulated as any country in Africa.

esistance
the
ortuguese

When the colonial powers turned their attention to the partition of Africa in the 19th century the Portuguese were allotted the vast area comprising Angola and at once set about gaining possession of it on the ground. Their endeavour was fiercely resisted by the inhabitants, and sporadic wars of conquest lasted for 30 years. The first campaign was fought in the south where the Cuanhama and other pastoralists fought a long but losing battle to resist both the Portuguese from the coast and the Boers (Dutch colonists) from South Africa. The most decisive campaign was the Bailundo War of 1902 in which the Portuguese finally broke the power of the Ovimbundu kingdoms and captured the Bié Plateau, thus enabling European entrepreneurs to build the Benguela Railway and establish white settlers on the highlands.

The trials and experiments of the first 30 years of colonial rule in Angola were often hindered by political changes and unrest in Portugal. The main development projects such as the railways and diamond mines were launched with foreign capital and initiative. When in the 1930s stability returned to Portugal, new emphasis was placed on colonial planning. Progress, however, was long delayed, first by the depression of the 1930s and the conservative economic policies and then by World War II. Not until the 1950s did progress begin to take place in education, harbour and airport construction, road building, and plantation work. (D.Bi./M.A.Sa.)

INDEPENDENCE

Angolan
war of
liberation

The origins of the Angolan war of liberation lie in a series of apparently unconnected uprisings. They began in January 1961 with a revolt by the Mbundu in central Angola against forced cotton production at a time of dropping prices. Although they were bombed into submission, the Mbundu remained at the heart of Angolan resistance. The uprising was followed in February by an attempted coup in Luanda by the Movimento Popular de Libertação de Angola (MPLA), which emerged as the most important of the nationalist groups. The attempted coup ended in white hysteria, a massacre of blacks, and the suppression of the MPLA by the secret police.

Most important was the outbreak in March 1961 of violence among the poverty-stricken Kongo in the north. There the coffee boom of the 1950s had led to an influx of white settlers. Africans lost their land and were turned into wage labourers, while forced contract workers were brought in from central Angola. A dispute over wages led to the killing of several hundred settlers and contract workers and was followed by massive retaliation that spread the violence across northern Angola. Although at this stage the Kongo struggle failed to spread, the war of attrition continued for more than a dozen years. The Kongo were led by Holden Roberto's Frente Nacional de Libertação de Angola (FNLA).

By the mid-1960s warfare had also erupted in eastern Angola with the MPLA, which moved its base from the Congo to Zambia in 1965. Under its poet-president, Agostinho Neto, the MPLA also resuscitated urban political activity in Luanda. Although the MPLA was not able to extend its hold to the southern highlands, where the Ovimbundu joined Jonas Savimbi's União Nacional para a Independência Total de Angola (UNITA), the situation there, too, was tense as land was lost to white cattle ranchers. (Sh.M./Ed.)

The transition to independence in Angola was highly complex, with three nationalist movements vying for control. Their rivalries were greatly exacerbated by the involvement of Zaire, South Africa, the United States, and the Soviet Union. After Portugal recognized Angola's right to independence in 1974 the three nationalist groups made several unsuccessful attempts to form a coalition government. The exodus of Portuguese settlers left the country with a crippled infrastructure and without effective government. When independence was declared in November 1975, the MPLA and an allied FNLA-UNITA proclaimed two rival republics, and the country was plunged into civil war. The Soviet Union supported the MPLA's People's Republic of Angola, and South Africa and the United States aided the People's Democratic Republic of Angola proclaimed by FNLA-UNITA. Although the MPLA eventually established control of the government, intense factionalism continued, supported by the foreign powers. The government was unable to stabilize the country's economy and pursue its long-term objectives.

For later developments in the history of Angola, see the *Britannica Book of the Year* section in the BRITANNICA WORLD DATA ANNUAL. (Ed.)

BOTSWANA

Botswana is a landlocked republic in southern Africa with an area of about 224,600 square miles (581,700 square kilometres). The greater part of the country is covered by the Kalahari. It is bounded by South Africa to the southeast and south, Namibia to the west and northwest, Zambia to the north, and Zimbabwe to the northeast and east. The actual border with Zambia along the Zambezi River, at the east end of the thin finger of Namibian territory known as the Caprivi Strip, is only several hundred yards long. The Caprivi Strip has not been established by international treaty and has been a source of controversy among the countries of Botswana, Namibia, Zimbabwe, and South Africa (the latter having acted as trustee for Namibia until 1990).

Before its independence in 1966, Botswana was a British protectorate known as Bechuanaland; one of the first acts of the nation was to move the capital from Mafeking in South Africa (since 1980 Mafikeng, Bophuthatswana) to Gaborone (formerly Gaberones). The country is named after its chief ethnic group, the Tswana, or Batswana. The language is Tswana (Setswana or Siswana).

Botswana is a member of the United Nations, the Commonwealth, and the Organization of African Unity (OAU), and it is a signatory of the Lomé Convention. It maintains economic links with South Africa as well as diplomatic ties with black states to the north and east, including Zimbabwe, Zambia, and Mozambique. Lacking adequate water and having a small population, Botswana was once primarily a stock-raising country. Large-scale mining of substantial diamond, copper, and nickel deposits and of other less important minerals began in the 1970s, diversifying and improving the economy.

Physical and human geography

THE LAND

The plateau

Relief and drainage. Botswana extends from the Chobe River (which drains into the Zambezi) in the north, to the Molopo River, a tributary of the Orange River, in the south. Most of the nation's territory lies within the Kalahari, a semidesert in the west. A plateau, with an altitude of about 4,000 feet (1,200 metres), divides the country into two distinctive regions, each with its own drainage system. After forming the watershed between the Molopo just to the south and the Notwani River to the southeast, the plateau extends northward from a point approximately 20 miles west of Kanye, a town in the southeast, to the border of Zimbabwe in the northeast. The area that lies to the east of the plateau is crossed by ephemeral streams that flow latitudinally into the Marico, Limpopo, and Notwani rivers; the western system, which once drained the tableland into the Makgadikgadi Pans, is now covered with deep sand.

The country has a mean altitude of 3,300 feet and is almost entirely flat or gently undulating, with Kalahari sands overlying Precambrian rock (from 570,000,000 to 4,600,000,000 years old). The greater part of the south has no surface drainage, and the rest of the country's drainage, except for the Limpopo and Chobe rivers on the borders, does not reach the sea.

The main drainage system is provided by the Okavango River, which, after flowing southeastward for 90 miles from the Angola border in the northwest, drains into a vast depression to form the approximately 4,000-square-mile Okavango Swamp, as well as Lake Ngami, a 70-square-mile pan, or basin. This system seasonally drains eastward along the Boteti River to Lake Xau (Dow), which is located near the centre of the country, and finally flows into the Makgadikgadi Pans farther east. The Molopo and Ramatlabama, which form the southern boundary, are both tiny streams.

The Kalahari is a "thirstland," rather than a true desert, being covered with grass and acacia thorn scrub; its tsama melons (a species of watermelon) help to sustain the San who live there. In areas where underground water is close to the surface, the Kalahari resembles parkland. The only typical desert occurs in the southwestern corner, where there are sand dunes.

Climate. Temperatures vary from temperate to subtropical. In summer (which lasts from October to March), temperatures rise to 100° F (38° C); in winter (which lasts from April to September), there is frequent frost, and temperatures fall below 32° F (0° C). Hot August winds carry sand from the Kalahari across the entire country. The average annual rainfall is 18 inches (460 millimetres), representing a variation from 25 inches in the north to five inches or less in the southwest. Most rain falls in heavy downpours between December and the end of April. Cyclical droughts, often lasting several years, are calamitous for both crops and cattle. One such drought occurred in 1965–66.

Plant and animal life. Woodland savanna (grassy parkland), in which species of acacia predominate, covers much of the country. In the northeast the most characteristic vegetation cover consists of mopane (African ironwood), mogonono (country almond), and mokusi (Rhodesian teak), while morukuru (*Spirostachys africanus*) frequently occurs in the central Kalahari. In the abundant grazing areas the several varieties of grasses include love grass, panicum, crabgrass, and bristle grass. True forest, which occurs only on the banks of the Chobe River, consists mainly of mokwa (bloodwood) mokusi, and monato (Rhodesian ash).

The game reserves

Herds of wild game teem on the plains, and the national parks and game reserves are well stocked. National parks include the Chobe, Nxai Pan, and Gemsbok; other wildlife areas are the Okavango delta and Lake Ngami, the Moremi Wildlife Reserve, and the Makgadikgadi Pans, Khutse, Mabuasehube, and Central Kalahari game reserves. Besides many species of large and small antelopes, there are lions, elephants, leopards, hippopotamuses, giraffes, buffaloes, and crocodiles. Poisonous snakes, among them the cobra and puff adder, abound; and there are many varieties of scorpions, spiders (including tarantulas), and termites. Bird life is prolific and includes ostriches, pelicans, and bustards. The principal fish, especially in the Okavango and Chobe rivers, are tilapia, tiger fish, and catfish.

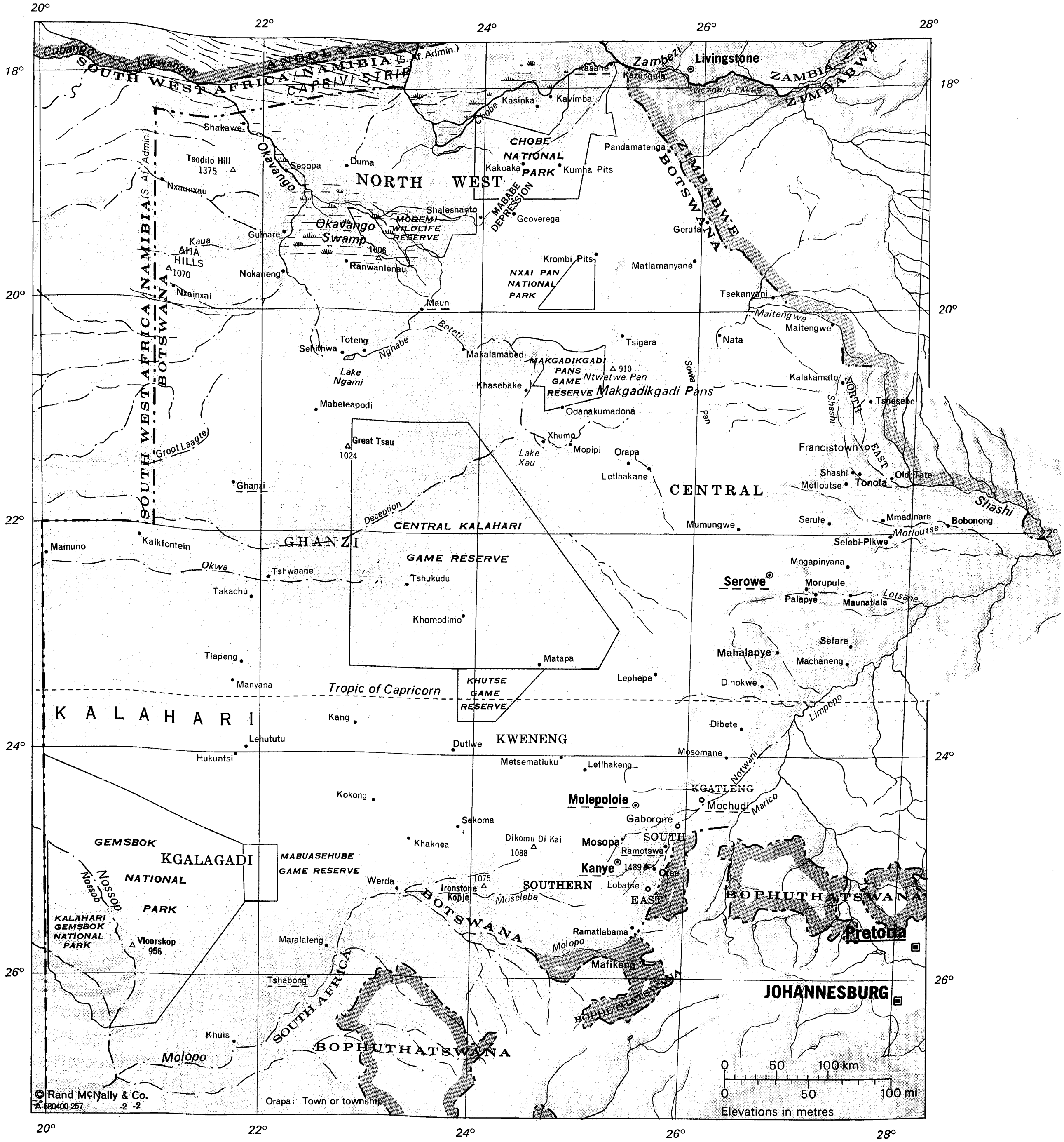
Settlement patterns. The Tswana people are divided into eight principal tribal groups, each of which occupies its own separate territory with its own traditional chiefs and retains a communal ownership over its tribal lands that is inalienable. Associated with each of the dominant tribes are smaller related or formerly subject tribes. Only the nomadic San of the Kalahari, the few thousand Europeans, and some Tswana who have moved to the urban areas are not included in this traditional system. The largest tribe is the Ngwato (Bamangwato), which comprises about one-fourth of the total population and owns one-fifth of all the land; its territory lies in the eastern part of the nation. The Ngwato capital, Serowe, is spread out beneath rocky hills and is a large circular village consisting mostly of thatched huts. The next two largest tribes are the Kwena (Bakwena) and the Ngwaketse (Bangwaketse), who live in the southeast, near Gaborone. The other smaller tribes include the Kgatla (Bakgatla), the Bamalete, and the Tlokwa (Batlokwa), all of whom live in the southeast; the Rolong (Barolong), who spill over into South Africa; and the Tawana (Batawana), who live on the South West African border.

In the 19th century, several large blocks of land were transferred from African to European ownership around Lobatse, Gaborone, and Tuli in the southeast; Ghanzi in the west; and on the banks of the Molopo River in the south. The former Tati Concession in the Francistown District (now part of North East District), for example, gave a private company the right in 1869 to exploit a potentially rich mining region having an area of more than 2,000 square miles.

The Tswana way of lif

Unlike most Africans, the Tswana have a tradition of living in large villages, which are, in effect, capital centres for each tribe but are often located at great distances from their grazing lands. Cattle posts provide shelters for herdsmen and watering places for cattle. Richer cattle farmers have a large number of cattle posts, which are sometimes spread out at distances of 100 miles or more from each other. The farmers' practice is to spend several days of each week at their cattle posts, returning to their central village and families for long weekends.

Each village has its subchief, headmen, and *kgotla* or *lekhotlo,* a court and public meeting place situated near the chief's house, usually under a large tree and always beside the communal cattle enclosures, or kraals. The *kgotla* remains the centre of daily social life, at which disputes are settled in public and matters of local importance are discussed. Most villages have a school, a store, and sometimes a dispensary. Villages, small or large, are always divided into wards comprising the extended family, each part of which lives within its own compound. In the larger villages each ward tries to have its own borehole for water. Water carrying still forms a large part of the daily routine of women and girls throughout Botswana, just as cattle herding occupies the lives of boys, often conflicting with their schooling. The outlying villages are related closely



BOTSWANA

MAP INDEX

Political subdivisions

Central	21·30s	26·00e
Chobe	18·30s	25·00e
Francistown	21·11s	27·32e
Gaborone	24·45s	25·55e
Ghanzi	22·00s	23·00e
Kgalagadi	25·00s	22·00e
Kgatleng	24·28s	26·05e
Kweneng	24·00s	24·00e
Lobatse	25·11s	25·40e
Ngamiland	19·09s	22·47e
North East	21·11s	27·32e
North West	19·00s	23·00e
Orapa	21·17s	25·22e
Selebi-Pikwe	22·00s	27·50e
South East	25·00s	25·45e
Southern	24·45s	24·00e

Cities and towns

Bobonong	21·58s	28·17e
Dibete	23·45s	25·26e
Dinokwe	23·24s	26·40e
Duma	18·45s	22·46e
Dutlwe	23·55s	23·47e
Francistown	21·11s	27·32e
Gaborone	24·45s	25·55e
Gcoverega	19·08s	24·15e
Gerufa	19·17s	26·02e
Ghanzi	21·38s	21·45e
Gumare	19·21s	22·12e
Hukuntsi	24·02s	21·48e
Kakoaka	18·40s	24·22e
Kalakamate	20·39s	27·21e
Kalkfontein	22·08s	20·53e
Kang	23·41s	22·50e
Kanye	24·59s	25·19e
Kasinka	18·13s	24·22e
Kavimba	18·02s	24·38e
Kazungula	17·45s	25·20e
Khakhea	24·51s	23·20e
Khasebake	20·41s	24·29e
Khomodimo	22·46s	23·52e
Khuis	26·37s	21·45e
Kokong	24·27s	23·03e
Krombi Pits	19·30s	25·02e
Kumha Pits	18·45s	24·45e
Lehututu	23·58s	21·51e
Lephepe	23·20s	25·50e
Letlhakane	21·27s	25·30e
Letlhakeng	24·08s	25·02e
Lobatse	25·11s	25·40e
Mabeleapodi	20·58s	22·36e
Machaneng	23·10s	27·26e
Mahalapye	23·04s	26·50e
Maitengwe	20·06s	27·13e
Makalamabedi	20·19s	23·51e
Mamuno	22·16s	20·01e
Manyana	23·23s	21·44e
Maralaleng	25·47s	22·45e
Matapa	23·11s	24·39e
Matlamanyane	19·33s	25·57e
Maun	20·00s	23·25e
Maunatlala	22·32s	27·28e
Metsematluku	24·01s	24·40e
Mmadinare	21·57s	27·52e
Mochudi	24·28s	26·05e
Mogapinyana	22·19s	27·27e
Molepolole	24·25s	25·30e
Mopipi	21·07s	24·55e
Morupule	22·36s	27·05e
Mosomane	24·04s	26·15e
Mosopa	24·50s	25·31e
Motloutse	21·28s	27·24e
Mumungwe	21·59s	26·24e
Nata	20·12s	26·12e
Nokaneng	19·40s	22·16e
Nxainxai	19·50s	21·13e
Nxaunxau	18·19s	21·04e
Odanakumadona	20·53s	24·45e
Old Tate	21·22s	27·46e
Orapa	21·17s	25·22e
Otse	25·02s	25·45e
Palapye	22·37s	27·06e
Pandamatenga	18·35s	25·42e
Ramatlabama	25·40s	25·30e
Ramotswa	25·50s	25·50e
Ranwanlenau	19·35s	22·47e
Sefare	23·02s	27·28e
Sehithwa	20·23s	22·45e
Sekoma	24·41s	23·50e
Selebi-Pikwe	22·00s	27·50e
Sepopa	18·13s	22·13e
Serowe	22·25s	26·44e
Shakawe	18·23s	21·50e
Shaleshanto	19·09s	23·58e
Shashe	21·20s	27·29e
Takachu	22·37s	21·58e
Tlapeng	23·15s	21·49e
Tonota	21·29s	27·29e
Toteng	20·22s	22·58e
Tsekanyani	19·52s	26·39e
Tshabong	26·03s	22·29e
Tshesebe	21·51s	27·35e
Tshukudu	22·30s	23·22e
Tshwaane	22·29s	22·03e
Tsigara	20·10s	25·18e
Werda	25·15s	23·16e
Xhumo	21·07s	24·42e

MAP INDEX (continued)

Physical features and points of interest

Aha Hills	19·45s	21·10e
Boteti, *river*	20·08s	23·23e
Central Kalahari Game Reserve	22·00s	24·00e
Chobe, *river*	17·50s	25·05e
Chobe National Park	18·45s	24·15e
Deception, *river*	21·16s	24·00e
Dikomu Di Kai, *mountain*	24·58s	24·31e
Gemsbok National Park	25·30s	21·10e
Groot Laagte, *river*	20·55s	21·27e
Ironstone Kopje, *mountain*	25·08s	24·05e
Kalahari, *desert*	24·00s	21·30e
Kaua, *stream*	19·24s	22·03e
Khutse Game Reserve	23·30s	24·15e
Limpopo, *river*	22·11s	29·20e
Lotsane, *river*	22·41s	28·11e
Mababe Depression	18·50s	24·15e
Mabuasehube Game Reserve	25·10s	22·00e
Maitengwe, *river*	19·59s	26·26e
Makgadikgadi Pans, *salt flat*	20·45s	25·30e
Marico, *river*	24·12s	26·52e
Molopo, *river*	26·55s	20·37e
Moremi Wildlife Reserve	19·50s	24·00e
Moselebe, *river*	25·03s	23·13e
Motloutse, *river*	22·15s	29·00e
Ngami, Lake	20·37s	22·40e
Nghabe, *river*	20·22s	22·58e
Nossop, *river*	26·55s	20·37e
Notwani, *river*	23·35s	26·58e
Ntwetwe Pan, *salt flat*	20·50s	25·15e
Nxai Pan National Park	19·55s	24·55e
Okavango, *river*	18·50s	22·25e
Okavango Swamp	18·45s	22·45e
Okwa, *river*	22·30s	23·00e
Ramatlabama, *river*	25·46s	25·14e
Shashi, *river*	22·11s	29·20e
Sowa Pan, *salt flat*	21·00s	26·00e
Tsodilo Hill	18·45s	21·45e
Xau, Lake	21·15s	24·38e
Zambezi, *river*	17·50s	25·20e

to their tribal capital, which is the seat of the paramount chief and the royal *kgotla.* This system of village democracy has, to a great extent, survived the politicizing of the nation and forms an integral part of the contemporary political structure.

Cattle, land, and rain are the most important factors in the lives of the people. Because the lives of people and cattle both depend on an unpredictable climate, rain (*pula*) is of particular significance. *Pula* is the people's familiar greeting on meeting each other and is the name of Botswana's currency.

There is considerable underemployment in Botswana, and many thousands of young people seek temporary work each year in South Africa. Outside the small tribal villages, employment within the country is mostly found in the half-dozen villages and towns along the railway line in the east—which links Botswana with South Africa and Zimbabwe—and at Selebi-Pikwe, the fast-growing centre of the copper- and nickel-mining industries. In such areas modern Tswana often break free of tribal constraints to form nationally integrated social and political groups.

Major cities

Gaborone, the capital city, has grown at the fastest rate. During the 1970s it passed Serowe, the largest of the tribal villages, in population. Francistown, the former centre of the Tati Concession area in the northeast, is the home of most of the European population. Mochudi and Lobatse have become commercial centres on the railroad line in the southeast.

THE PEOPLE

Ethnic groups. The original inhabitants were the San (known locally as the Sarwa or Masarwa), some of whom follow a nomadic way of life in the Kalahari, moving across the border of Namibia; they belong to the Naron, Auen, !Kung, and Heikum groups. Those of the San who have abandoned nomadism are mainly employed in servile positions by the dominant tribes or by the government. The Tswana group is composed of a cluster of tribes that are closely related in culture and language to the western Sotho.

The European population, mainly from South Africa, is small; it consists of farmers, traders, public officials, missionaries, and teachers. Most of these people live in European agricultural zones and in the towns along the railroad. Also living in Botswana are a few thousand Eurafricans, who are of mixed descent, and several hundred Asians.

The two official languages are Tswana and English. English is spoken only among the educated Africans or those who have worked for some time in South Africa. Afrikaans is spoken only by sections of the white community. Approximately one-third of the population is literate in Tswana and one-fourth in English.

Religious groups. The majority of people are Christians, although only a minority are active churchgoers; many are only nominal Christians in the sense that they still show some attachment to traditional religion, which is based on ancestor worship. The majority of practicing Christians are Anglicans, Congregationalists, Methodists, and Presbyterians; these four groups have united to worship as a single congregation. There are also smaller communities of the Dutch Reformed, African Methodist Episcopal, and Roman Catholic churches.

THE ECONOMY

Distribution of wealth

Most Tswana live by raising cattle. Mining, however, overtook the beef industry during the 1970s as the basis of the country's most important exports, and the mining of diamonds became the fastest growing source of revenue. A series of national economic plans after independence has given particular attention to energy needs, mining development, beef exports, and crop production. The economy has grown rapidly, but the distribution of this increased wealth has been disproportionate; wage-earning urban dwellers and large-scale livestock farmers have benefitted, while the rural majority of the population has experienced little or no gain.

Agriculture. Botswana's cattle are important to the country's economy. The animals are mostly crossbreeds; attempts have been made to introduce quality cattle based on imported herds. Except for European farming, there are few cash crops; most agriculture is grown on a subsistence basis. Periodic droughts and disease necessitate the importation of food and wreak havoc among the cattle population.

Mining. Diamond mines at Orapa, with a 50 percent share held by the government, began production in 1971; a second mine at nearby Lethakane began production in 1977. Since then diamonds have become the single most important export. A major new diamond pipe was discovered at Jwaneng in the late 1970s, and mining began there in the early 1980s.

The fastest growing area of Botswana, however, has been around Selebi-Pikwe, the location of the Shashi Project, which includes nickel and copper mines and water and power projects. Nickel and copper mining, begun in 1974, was initially hampered by technical problems but has come to play an increasingly important role in the economy of the country. There are substantial reserves of coal at Morupule and near Serowe. Prospecting for other minerals has continued, and legislation in 1977 vested all mineral rights in the state.

Industry. Almost all of Botswana's industries are associated with mining and agriculture, and there is virtually no manufacturing. The Botswana Development Corporation has assisted in the establishment of some industrial and commercial enterprises. Many of the smaller projects have been built in rural areas in an effort to decentralize employment.

Administration of the economy. The government is the largest employer of wage-earning labour and is active in promoting cooperative societies and water-development projects. In regard to private enterprise, the bulk of import and export trade is in European hands.

Transportation. During the 1970s the government began taking over the country's principal railway, which is managed by The National Railways of Zimbabwe. A main north–south road was completed in 1980, giving Botswana an all-weather road that runs the entire length of the country. Botswana has, in addition, several thousand miles of gravel and dirt roads. Air Botswana maintains domestic service and, with other airlines, provides international connections for travellers.

ADMINISTRATIVE AND SOCIAL CONDITIONS

Government. Legislative power is vested in the Parliament, which consists of the president and the National Assembly of 32 members elected by adult suffrage, four specially chosen members, and the attorney general. The president is the executive head of government; he is a member of and presides over the Cabinet, consisting of the vice president and other ministers drawn from the National Assembly. Elections for the presidency and National

Assembly are held every five years. The House of Chiefs advises the government; no legislation affecting tribal affairs may be passed without reference to this House. A code of human rights is enforceable by the High Court. A nonpolitical public service commission controls all matters relating to the public service. The local government system operates through nine district councils, as well as four town councils located in Gaborone, Francistown, Selebi-Pikwe, and Lobatse.

There are three nationalist parties, which represent more militant ideas than the ruling Botswana Democratic Party; they are the Botswana People's Party, the Botswana Independence Party, and the Botswana National Front. All four parties are national in the sense of having multitribal members; they usually attract both modern and tradition-minded supporters.

Justice. The High Court, consisting of a chief justice and puisne (associate) judges, exercises jurisdiction over all criminal and civil cases. The Court of Appeals deals with matters emanating from the High Court. There are subordinate magistrates' courts and African courts with limited jurisdiction in the local administrative districts.

Education. There are primary and secondary schools, some of which are run by mission societies and some by the government. Only a small number of pupils receive secondary education. Higher education is provided by the University College of Botswana, in Gaborone, and at teacher-training colleges. Technical training is also provided throughout the country.

Health and welfare. The most prevalent sicknesses are tuberculosis, gastroenteritis, pneumonia, and food deficiency diseases. Malaria is endemic in Chobe and Ngamiland, which is also infected by sleeping sickness. Bilharzia (a parasitic disease) occurs in the eastern area and in the Okavango Basin. There are hospitals and several health centres, dispensaries, and clinics. Care for mothers and children has been emphasized, and dental treatment was introduced in the 1970s.

CULTURAL LIFE

The life of the majority of the people is still strongly influenced by tribal institutions and cultures; there is a rich tradition of folklore, music, and dancing. The only strongly competitive cultural institutions are those provided by the churches and by some of the schools; Swaneng Hills School, an imaginative institution started by a South African expatriate, Patrick van Rensberg, may be especially mentioned. While not much vernacular literature of exceptional quality exists as yet, there are the beginnings of a modern vernacular literary tradition.

For statistical data on the land and people of Botswana, see the *Britannica World Data* section in the BRITANNICA WORLD DATA ANNUAL. (Co.L./Ed.)

History

Sotho peoples arrived in southern Africa at the time of the Bantu migrations; *i.e.,* mainly by 1600 or 1700. Some of the Tswana people only recently occupied their present homes.

EARLY EXPLORATION

A European expedition headed by J. Trüter and W. Somerville visited the Thlaping (or Tlhaping) of southern Botswana in 1801. These early visitors were followed by the traveller Martin Heinrich (Henry) Lichtenstein and the naturalist W.J. Burchell. John Campbell, having in 1813 been invited by the chief of the Thlaping to "send instructors," the London Missionary Society established a mission on the Kuruman (properly Kudumane) River. Under Robert Moffat, who took charge in 1821, Kuruman became one of the foremost Protestant missions of southern Africa. About this time the rise of the Zulu power had set up a chain of destructive movement across the continent. Thousands of refugees, fleeing from Shaka's Zulu armies, roamed the countryside pillaging and plundering and in their turn starting fresh waves of migration. In 1823 the Thlaping, threatened by one such horde, were saved only by Moffat's enlisting the help of Griqua mounted riflemen. In the same year Sebetwane started from the Vaal River and, after fighting his way through Botswana, founded the powerful but short-lived empire of the Makololo in the Zambezi Valley.

About 1826 Mzilikazi (Umsilikazi, Mosilikatze), son of Machobane, one of Shaka's captains, after breaking away from the Zulu king, settled with his people, the Ndebele, in what is now the western Transvaal. From there he raided extensively into Botswana. These depredations continued even when the Ndebele, after their defeat by the Boers in 1837, had moved to the north. The Great Trek of the Boers from the Cape Colony (1835 and after) brought the Voortrekkers across the Vaal River to the frontiers of Botswana. The Boers disliked the proximity of powerful, independent chiefs; thought that the penetration of Botswana by the missionaries of the London Missionary Society and by British traders and hunters would curb Boer expansion northward; and claimed a western boundary that included the "missionaries' road" (which ran through Botswana), the highroad from the Cape to the interior. They came sharply into conflict with David Livingstone, who had landed in Africa in 1841 and had settled among the Kwena, a senior Tswana tribe, whose chief, Sechele (Setshele), became a Christian. Livingstone sided with the Africans in their resistance to the various demands of the Boers, who further suspected him of supplying Sechele with firearms. In 1852 a raid was made on Sechele's town of Dimawe. During the raid Livingstone's mission at Kolobeng, near Dimawe, was ransacked, though by whom was never conclusively proved. The widely publicized raid on Dimawe was only one incident in a struggle that lasted for many years.

The Boers
and David
Livingstone

GOLDFIELDS

In 1867 gold was discovered near the Tati River, in country claimed both by the Ngwato and by the Ndebele. Pres. Marthinus Pretorius of the Transvaal in 1868 proclaimed the annexation of a huge area including the goldfields, but the British government refused to recognize this. At the same time Macheng (Matsheng), chief of the Ngwato, sought without success to persuade the British to occupy the goldfields. In the south the arbitration award of R.W. Keate, lieutenant governor of Natal, in 1871, purporting to define a boundary between the Rolong and the Transvaal, did nothing to abate Boer encroachment. Disorders among the tribes in 1878, quickly suppressed, led to the occupation of southern Botswana by a small police force. On the withdrawal of this force in 1881 the country relapsed into anarchy.

The Pretoria Convention of 1881 defined the boundaries of the Transvaal on all sides for the first time but still did not curb Boer encroachment. The chiefs relied on European supporters (not all of them Boers), to whom they promised land and cattle. These so-called volunteers then set up the two little republics of Stellaland and Goshen, threatening the missionaries' road. In 1883 Pres. S.J.P. Kruger of the Transvaal led a deputation to London to seek a revision of the Pretoria Convention. The London Convention of 1884 conceded to the Boers, among other things, small alterations in the western frontier; but Kruger failed to get the missionaries' road, though he tried hard to do so. About this time, moreover, German colonization in South West Africa was causing concern to the British. Humanitarians, inspired by John Mackenzie, a successor of Moffat at Kuruman and a vehement campaigner for African rights, were opposing the extension of Boer authority and advocating that the country should be taken under British protection. Cecil Rhodes meanwhile threw all his weight into a campaign for the annexation of Botswana to the Cape Colony. The British government then decided to establish a protectorate over the territory that later became British Bechuanaland and sent John Mackenzie as the first deputy commissioner. Mackenzie, without men or money, was unable to make much headway against the prevailing anarchy. He was soon superseded by Rhodes, who did no better.

Freebooters from Goshen attacked Mafeking; Kruger annexed the country of the Rolong; the republican flag was hoisted at Mafeking; and the freebooters proceeded to carve up the adjacent African-held lands. The British government, prompted by these aggressions, sent a strong force under Sir Charles Warren to set matters in order (which Warren was able to do without firing a shot) and then extended the protectorate northward to the 22nd parallel of south latitude and westward to the 20th degree of east longitude (March 1885). The territory south of the Molopo River was constituted as the crown colony of British Bechuanaland (September 1885; annexed to the Cape Colony, 1895). In 1892 another proclamation considerably enlarged the protectorate by adding to it the territory between the Shashi and Motloutsi rivers; the Tati district; those territories north of latitude 22° that were claimed by Khama (Kgama) III, chief of the Ngwato since 1875; and some vast ill-defined areas lying north and west of Khama's country.

MAINTAINING THE PROTECTORATE

In the early 1890s the protectorate, as Rhodes had intended, became the springboard for the British South Africa Company's colonization of Southern Rhodesia. But in 1895 a plan to hand the administration of the protectorate over to the company was frustrated by the action of the three chiefs Khama, Sebele, and Bathoen, who, supported by the missionaries, travelled to London and wrung from Joseph Chamberlain an assurance that their people would remain under the protection of the crown. In exchange they had to surrender strips of land on the east side of their countries for the construction of a railway. The Jameson Raid at the end of 1895 completed the ruin of any plans for the company's administration of the protectorate.

South Africa often expressed the wish that the administration of Botswana (with Lesotho [Basutoland] and Swaziland) should be handed over to it. The British government's position was made public on June 20, 1935; no transfer was to take place until the inhabitants had been consulted and until Parliament had expressed its opinion. On several occasions subsequent British governments confirmed that there had been no change in this policy.

In 1948 Seretse (later Sir Seretse Khama, who was knighted just before independence), son of Khama's son Sekgoma II and heir to the Ngwato chieftainship, married an Englishwoman, Ruth Williams, while he was studying in Britain. The British government, it is thought because of pressure from South Africa (Khama was declared a prohibited immigrant in that country), refused to recognize him as chief and he was banished from Botswana.

Until 1961 the country was administered by a resident commissioner under the direction of the British high commissioner in South Africa and with the assistance of an advisory council of chiefs. In 1961 executive and legislative councils, the latter with an unofficial majority, were introduced. In 1965 a constitution granting internal self-government was introduced, and at the first general election the Bechuanaland (later Botswana) Democratic Party (BDP) formed the government that led the new state to independence. The BDP was headed by Seretse Khama, who had been permitted to return as a private citizen in 1956. In an 80 percent poll the BDP received 113,177 votes and 28 of the Legislative Assembly's 31 seats. The Botswana People's Party received 19,664 votes and the remaining three seats.

INDEPENDENCE

The Republic of Botswana became an independent member of the Commonwealth of Nations on September 30, 1966. It was also admitted to the United Nations in that year. By terms of the constitution, the prime minister (Sir Seretse Khama) became the first president and the Legislative Assembly elected in 1965 became the first National Assembly. (Ay.Sy./S.Tr.)

Relations with Soutl Africa

South Africa changed its previous attitude to the former high commission territories (Bechuanaland, Basutoland, and Swaziland). In 1965 the prohibition on Khama's entering South Africa had been lifted. South African Prime Minister B.J. Vorster later stated that Botswana's becoming independent was not a development that caused concern and that it was in keeping with South Africa's own policy. Sir Seretse Khama, however, dismissed a suggestion, coming from unofficial quarters, that Botswana should join a southern African federation. This, he said, would prevent Botswana from having ties with black Africa and from acting as a bridge between north and south. Moreover, this would make Botswana a "Bantustan," which could not be permitted.

Sir Seretse Khama and his government therefore attempted to maintain a delicate balance between Botswana's "front-line" neighbours—the majority-ruled states bordering on South Africa—and South Africa itself. Botswana retained its close trade and transport links with South Africa, but in 1980 it also was among the founding members of the Southern African Development Coordination Conference, an economic and political union formed to increase regional cooperation and decrease the member nations' dependence on South Africa.

Botswana harboured refugees from Rhodesia, during the struggle there for majority rule, and from South Africa and was accused by South Africa of supporting guerrillas of the outlawed African National Congress (ANC) as well. The country was thus exposed to incursions by South African forces seeking to rout ANC guerrillas and supporters. At the time of independence Botswana had chosen not to establish an army but to rely on an efficient police force instead, but in 1977 a permanent defense force was created to patrol the borders and maintain order.

Sir Seretse Kharma was sworn in for his second term as president; he held office until his death on July 13, 1980, and was succeeded by the vice president, Quett Masire. The Botswana Democratic Party retained its majority representation in the national government, although some local support was lost to opposition parties in elections held in 1984.

For later development in the history of Botswana, see the *Britannica Book of the Year* section in the BRITANNICA WORLD DATA ANNUAL. (Ay.Sy./S.Tr./Ed.)

LESOTHO

The Kingdom of Lesotho in southern Africa has an area of 11,720 square miles (30,355 square kilometres). It is one of only two independent states in the world (the other is the Republic of San Marino in western Europe) completely encircled by a single country on whom it must depend exclusively for access to the outside world. It forms an enclave within the Republic of South Africa, bordering on three of the latter's provinces—Natal, the Orange Free State, and Cape Province. This physical dependence on its neighbour is further accentuated by Lesotho's own grave lack of resources, which makes it necessary for between one-fourth and one-half of the labour force to live and work, mainly as migrants, in South Africa. For this reason, it is sometimes described as a "hostage state." It is a member of the Commonwealth and of the Organization of African Unity, and it is a signatory of the Lomé Convention. Before its independence on October 4, 1966, it was one of the three British High Commission Territories—the other two being Bechuanaland (now Botswana) and Swaziland.

Lesotho is the name of the country and Sotho (or Basotho or Basuto) is the name of the people; a single individual is referred to as a Mosotho. Sotho (or Sesotho) is the language. The capital is Maseru.

Physical and human geography

THE LAND

Relief. Two-thirds of the country consists of mountains; the highest peak, Thabana Ntlenyana, is 11,425 feet (3,482 metres) above sea level. The Drakensberg Range forms the eastern boundary with Natal. The Maloti spurs,

he
ateau

running north and south, join the main range in the north, where they form a plateau between 9,000 and 10,500 feet in altitude. This plateau is the source of South Africa's two largest rivers—the eastward-flowing Tugela and the westward-flowing Orange, as well as tributaries of the Caledon. The foothills, which have average altitudes of between 6,000 and 7,000 feet, descend in undulating slopes to the west, where the lowlands bordering on the Orange Free State average 5,000 to 6,000 feet in altitude.

The mountain soils are of basaltic origin and are shallow but rich. The soils of the lowlands derive mainly from the underlying sandstone. Extensive erosion has severely damaged soils all over the country.

Climate. The rainfall, brought by the prevailing winds, occurs mostly between October and April; it is variable, averaging about 28 inches (711 millimetres) a year throughout most of the country. Although droughts are rare, their periodic occurrence is devastating. Temperatures in the lowlands vary from 90° F (32° C) in the summer to 20° F (−7° C) in the winter. In the highlands the temperature range is much wider, and readings below 0° F (−18° C) are not unusual. Frost occurs widely in the winter, when the Maloti are usually snowcapped. Hail is a frequent summer hazard.

Plant and animal life. The grasslands of the highlands are mainly covered by *Themeda triandra* or red grass. Indigenous trees and shrubs have survived only in the more remote and sheltered valleys. The most common types are wild olives, wild garlic, wild willow, brushwood, thatching grass (*Hyparrhenia hirta*), hypoxis, and aloes. Wildlife has become almost extinct, except for occasional small antelopes, hares, and reptiles.

Settlement patterns. Although not permanently inhabited, the mountain grasslands on the slopes of the high plateau as well as in the valleys provide excellent grazing for sheep and cattle, which are tended by herdsmen in isolated cattle posts. Some of the deep valleys, like the Senqunyane, produce good crops of wheat, peas, and beans. About one-third of the population lives in the highlands; another one-fifth lives in the foothills, which good grazing and cultivation have made the most productive region. One-half of the population lives in the lowlands, which form a narrow corridor, averaging only 25 miles (40 kilometres) in width, along the Caledon River.

ıe
stem of
igratory
ɔour

The Sotho combine modern and traditional ways to an unusual degree; this provides stability in a society that, possibly more than any other, is disrupted by a permanent system of migratory labour. Traditional authority is still firmly exercised through a system of chieftaincy extending from the paramount chief (the king) and his court, down through senior chiefs and subchiefs, to headmen and subheadmen at the village level. Their authority now rests largely on the fact that they are responsible for the working and distribution of all land that belongs to the nation, with the king acting as trustee.

Social cohesion is strengthened by the persistence of clan and family loyalties. Families and clans still cluster together as units in the numerous small rural villages. There are no large towns in the country. The villages range in size from one large to four or five extended families, with an average of from 30 to 50 immediate families. The villages are often picturesque, offering fine views of the rocky highlands; on the plains they are often surrounded by aloes and trees, and the walls and doors of many houses are covered with patterned designs. The villages themselves consist of clusters of circular or rectangular huts solidly built of turf, Kimberley brick (unburned clay), or dressed stone, mostly with thatched roofs; more recently, however, corrugated iron has become a popular roofing material.

The average household usually has two or three huts, the larger one being used as a living and dining area and as the parents' bedroom; the smaller ones are used for kitchen and storage purposes and as sleeping quarters for the children. The hut of the chief, or headman, is usually in the centre of the village, flanked by that of the principal wife and ringed by those of the junior wives. The *lekhotla* (open court) is in front of the chief's hut; beside it are the kraals (enclosures) for the cattle and stables for horses. Village life centres largely on the fields, the chief's court, the kraals, the school, the church, and the initiation lodge. Circumcision forms an integral part of the ritualized initiation ceremonies that train boys to take their place as full members of the family, clan, and nation—the three centres of social cohesion. Young boys still spend a large part of their lives as herdsmen, while women and young girls do much of the hard work in the fields, a practical necessity because the able men are usually absent, although they contrive to return home briefly for the plowing or the harvest.

Initiation
ceremonies

Two distinctive features of social life are the people's love for horses and for blankets. The small, sturdy Sotho pony is renowned for its surefootedness in rough country; it is often the only means of rural transport. Because of the sharp variations in climate, both men and women invariably wear blankets, which they use as cloaks; a great deal of care is taken in choosing a blanket, which, especially for the men, is usually multicoloured. Men and women also wear the typical Sotho hat, which is woven from reed into conical shapes, with an unusual topknot.

Maseru, the capital, in the northwest on the banks of the Caledon River, is a small rural town with civil servants, professional people, and traders. On most days of the week, however, it is a bustling town crowded with people who are either in from the surrounding lowlands or down from the mountains, or with migratory workers passing to and from South Africa.

THE PEOPLE

Ethnic groups. The Sotho comprise a cluster of tribes of the Southern Sotho linguistic stock, who are united by a common loyalty to the royal house of Mshweshwe (or Moshesh or Moshoeshoe). The group that formed the nucleus of the nation is that of the Kwena, who are made up of the Molibeli, Monaheng, Hlakwana, Kxwakxwa, and Fokeng tribes. They still regard themselves as the true nation and tend to look down on the other two main groups that were politically absorbed into their society but were not always culturally assimilated, although they speak the same language. The first of these groups is formed by the Natal (North) Nguni, who include the Phetla, Polane, and Phuti; they are largely indistinguishable from the Kwena. The second group comprises the Mahlape tribe of the Natal Nguni and the Cape (South) Nguni (Thembu).

There are a few thousand Europeans, Asians, and Coloureds (of mixed racial origin). The Europeans are mainly traders, businessmen, technicians, government officials, missionaries, and teachers, while the Asians are mainly professionals or traders.

Emigration

Because the country is severely overpopulated, both temporary and permanent emigration has taken place. No exact figures exist for the number of Sotho working or living in South Africa. The great majority of the migrant workers are men under the age of 40.

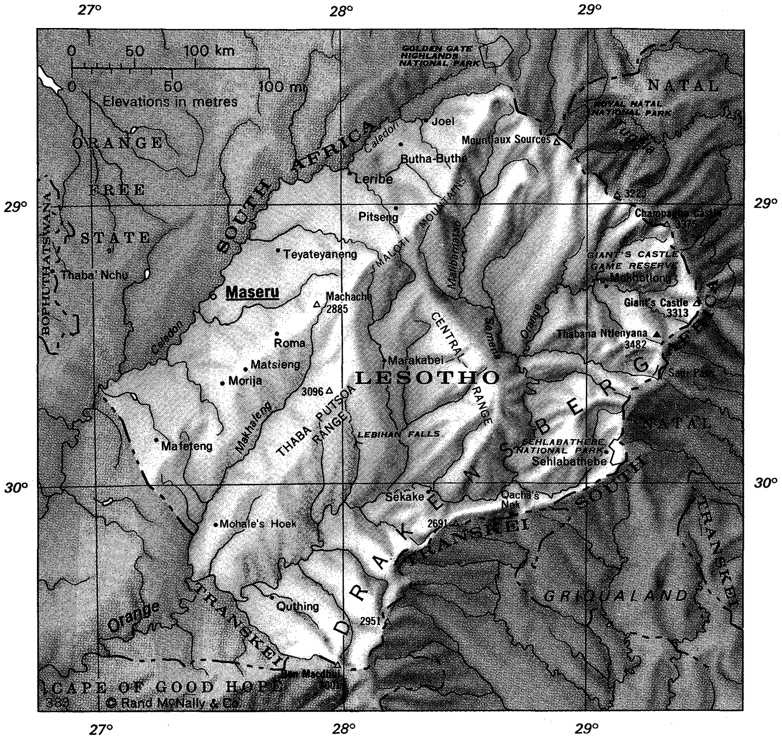
Religious groups. Christianity and traditional religions (which are based on a belief in ancestral spirits) coexist uneasily in a country where the great majority of people are active members of Christian churches. The Christians are about evenly divided between Roman Catholics and Protestants (mainly French Protestants, Anglicans, Baptists, and members of the Dutch Reformed Church).

THE ECONOMY

Economic
links to
South
Africa

Lesotho is a poor country with few natural resources; these are certainly insufficient to sustain even the present population. Its economy could not be sustained at all without such benefits as it derives from South Africa, with which Lesotho forms part of a customs union and shares an integrated communications system. Lesotho also depends wholly on South Africa for the export of its surplus working population.

Agriculture. Only about one-tenth of the land is arable; this amounts to less than one acre (one-half hectare) per head of the resident population. Agriculture provides about one-third of the gross domestic product. The crop yields are low; subsistence crops such as corn (maize), sorghum, wheat, peas, beans, and barley are the most important. A large percentage of the cattle die each year, and so the cattle herds have little commercial value. By



LESOTHO

contrast, the sheep and goat herds are valuable assets. Wool is exported, and Lesotho is one of the world's largest mohair producers.

Industry. Geological surveys have so far shown little promise of mineral wealth, although small quantities of diamonds have been found. Other minerals discovered, such as iron ore, gypsum, quartz, and calcite (limestone chalk), are commercially uneconomic.

Secondary industry is of recent growth and is confined to such small-scale manufacturing as milling, brickmaking, and furniture making.

Tourism. Tourism is promising for a country that has justifiably been called the Switzerland of southern Africa. Modern hotels have been built, mountain roads and pony trails developed, and trout streams stocked with fish. There are casinos and large hotels in Maseru and a ski resort at Oxbow.

Trade. Lesotho has a large and growing trade deficit. Most trade is with South Africa. The large deficit is made up by the remittances of Lesotho's migrant workers, by external aid, and by receipts from the Customs Union in agreement with South Africa, Swaziland, and Botswana.

Administration of the economy. The government is the largest employer of labour in the country, more than one-third of its annual budget being made up of payments to its public employees. Three forms of direct taxation are levied: a basic tax on all adult males; a graded tax on all income earners; and an income tax.

Transportation. A main road runs along the western and southern boundary, and a mountain road from Maseru reaches into the interior. These two main arteries are served by short-distance feeder roads. Villages in the mountains are served by bridle paths. A one-mile railway line links the capital to the South African transport network.

Although considerable use is made of light aircraft for passengers and for transporting mail and freight to the interior, only Maseru has a regular airport suitable for commercial planes. There are also rough airstrips. A national airline was formed in 1977.

ADMINISTRATIVE AND SOCIAL CONDITIONS

The country's constitution provides for a bicameral parliamentary system with the king as hereditary head of state. Under the constitution the National Assembly has 60 members, each elected by universal adult suffrage to represent a single constituency; the Senate has 33 members, of whom 22 are principal chiefs and 11 are appointed.

The constitution was suspended following the contested results of the 1970 general election, in which the opposition claimed to have defeated the ruling Basotho National Party (BNP). The legislature was replaced by an appointed 93-member interim assembly that included the 33 members of the Senate. In 1973 a new National Assembly, consisting of the 22 principal chiefs and 71 members from the BNP and other parties, was appointed, but the assembly was dissolved in 1986.

Justice. The legal system is based on Roman and Dutch law. There are magistrates' courts, a Court of Appeals, and a High Court. Appeals of certain cases of a constitutional nature may be taken to the Privy Council in the United Kingdom.

Literacy

Education. The country has one of the highest literacy rates in Africa. More girls than boys attend school, due to the latter's employment as herdsmen. Although about nine-tenths of all school-age children are in primary schools, fewer than one-tenth attend beyond the lower primary level, and fewer than a third of these attend secondary schools. Most of the schools are run by the Christian churches. There are primary and secondary schools, one agricultural school, and teacher-training colleges. The National University of Lesotho is based at Roma, 22 miles from Maseru.

Health and welfare. The country's healthy climate contributes largely to the comparatively low rate of sickness. The main incidence of ill health is the result of food-deficiency diseases, venereal disease, chronic rheumatism, infections of the respiratory tract, and dyspepsia. There are several hospitals, about half of which are operated by the government, and a number of clinics for maternity, child health, and venereal diseases, as well as a few health centres and dispensaries.

The country offers few opportunities for earning a salary. Some teachers and corporate employees are paid by the state. Of those working outside the agricultural sector, most are employed by the government, by schools or missions, in commerce, or by cooperative societies.

CULTURAL LIFE

Traditional institutions remain strong despite the modernization and politicizing of Sotho society. The paramount chief and the system of chieftaincy remain a strong focus

nportance
'the
ieftains

of loyalty, especially since some of its more venal aspects were reformed in the 1960s; there still is, however, considerable criticism of the power of chiefs and headmen, especially of their manner of administering the distribution of land and of the proliferation of officeholders within the chieftaincy system. Historical traditions and the philosophy of the founder of the nation, Mshweshwe (who was born at the end of the 18th century and died in 1870), exercise a profound influence in sustaining the Sotho nation's sense of independence and national pride. There is a strong literary tradition, which has been stimulated by modern education. Folklore flourishes, both in oral and written forms.

The initiation schools continue to play an important role in the villages and are a source of keeping alive a sense of the traditional values of the clans and the nation.

Village crafts, such as the decoration of walls, woodwork, clay making, weaving, and basketwork, flourish. Choir singing is popular.

For statistical data on the land and people of Lesotho, see the *Britannica World Data* section in the BRITANNICA WORLD DATA ANNUAL. (Co.L./Ed.)

History

The San (Bushmen) were the earliest inhabitants of what is now Lesotho, but they died out many years ago. The first Bantu were Nguni (Zulu-speaking) tribes who crossed the Drakensberg in the 18th century and settled along the banks of the Caledon River. Later they were joined by Sotho-speaking tribes. They all lived at peace, disturbed only by intermittent cattle raids, until early in the 19th century, when that part of Africa was set in turmoil by the Zulu Shaka. Many tribes fled over the Drakensberg, pursued by his regiments until the social fabric of the people was almost destroyed. Eventually a young chief, Mshweshwe (also Moshoeshoe or Moshesh), born about 1786, gathered the remnants of the tribes about him and, using impregnable mountain fortresses such as Thaba Bosiu as his bases, beat back the invaders, and established his people securely against the Zulus.

No sooner was the Zulu menace from the east contained than the Basuto people had to face the Europeans who were trekking up from the south. With the unerring skill of a diplomat, Mshweshwe played off the rulers of the Boer states and the colonial government of the Cape Colony against each other. He welcomed missionaries and traders to Lesotho, but with both he showed himself well able to hold his own. As tension increased between the Basuto and the Boers, who were passing through or settling in his country in increasing numbers, Mshweshwe's conviction grew that the only hope of "existence and peace" lay in British protection. He laboured incessantly to make this hope a reality, and with the help of Sir George Napier, governor of the Cape Colony, he achieved it.

Protection was first granted in 1843. Then in 1848 British sovereignty was proclaimed over the Orange River country, and in 1849 the boundaries of Mshweshwe's territory were reduced. But in 1854 the British renounced sovereignty over the Boers north of the Orange, and Mshweshwe was left to battle on through repeated disappointments and almost continuous strife with his Bantu, Boer, and British neighbours, until at last, in 1868, Lesotho was proclaimed British territory. The Boers, who had overrun practically the whole country up to the foothills, protested in vain against this interference and had to be content with a 30-mile strip of "conquered territory" along the Caledon. Mshweshwe died two years later, happy that his people had at last become British subjects. The Basuto owe their existence as a nation to the genius of Mshweshwe, who must be regarded as one of the dominant figures of South African history in the 19th century.

ınex-
on to
pe
lony

In 1871, without adequate consultation, Lesotho was annexed to the Cape Colony and a small administrative staff established. Then tactless attempts at direct rule, ignoring Basuto customs and loyalties, coupled with precipitate doubling of the hut tax, led to minor revolts and increasing restlessness. This finally flared into open revolt (the Gun War, 1880) when the government attempted to disarm the people. Peace was reestablished in 1883. The Cape Colony, weary of attempting to administer Lesotho, asked the British government to be relieved of the responsibility, while the Basuto chiefs for their part preferred British protection to the possibility of renewed encroachment from the Free State. In 1884 Lesotho ceased to be a portion of the Cape Colony and became a crown colony.

The new governor's agent was instructed that "nothing more should be attempted at first than the protection of life and property and the maintenance of order on the border," and that the Basuto "were to be encouraged to establish internal self-government sufficient to suppress crimes and settle intertribal disputes." These instructions were faithfully obeyed. The chiefs were left to rule their people according to traditional law and custom, and the authority of Mshweshwe's senior son and of his son Lerotholi after him was supported. By 1898, after a few bloody affrays, Lerotholi enjoyed undisputed supremacy, and British control was accepted. Credit for this was due to the early resident commissioners, Sir Marshall Clarke (1884), Sir Godfrey Lagden (1893), and Sir Herbert Sloley (1902), and to Lerotholi, who was paramount chief from 1891 to 1905 during the difficult transitional period. In 1905 Lerotholi died, and the national council with the imperial government concurring elected his son Letsie chief.

Basuto fears for their future were never entirely allayed, however. During the South African War, they were uneasily neutral. In 1906 the extension of the South African railway system to Maseru, indicative of the growing importance of the country's agriculture, trade, and labour supplies, was strenuously opposed by the younger chiefs as a potential threat to their independence, although only one mile of the new line lay inside the border. These fears were revived again in 1909 by the unification of the four South African colonies and discussion about the transfer of the administration of the High Commission Territories to the government of the new Union of South Africa. A deputation petitioned the king, asking that the country remain "outside [the Union] as far as possible, independent as it is now" and was reassured.

Thereafter the matter remained quiescent until 1933, when informal discussions were opened between the Union and British governments. In 1935 an *aide-mémoire* recalled the pledges made by the British government both to the British Parliament during the passage of the South Africa Act (1909) and to the inhabitants of the Territories, and recorded that the two governments now agreed that in the ensuing years their policies toward Africans "should be directed to bringing about a situation in which if transfer were to become a matter of practical politics it could be effected with the full acquiescence of the populations concerned." The South African government continued until 1955 to assume that it would ultimately incorporate Lesotho. The Basuto, however, never expressed anything but hostility to this assumption, and their determination not to be taken over by South Africa was strengthened in mid-20th century by the rise of a nationalist movement seeking independence and by the support given the movement by member states of the United Nations and the Organization of African Unity. The South African government initially showed great anxiety about the possibility of an independent African state in its heartland, but after the electoral victory in 1965 of the conservative and traditionalist Basotho National Party, led by Chief Leabua Jonathan, it agreed to its small and completely dependent neighbour's mapping out its own political future.

Nationalist
movement

After discussions in September 1966 Hendrik Verwoerd, prime minister of the Republic of South Africa, and Chief Jonathan issued a statement outlining their objectives. These were "to establish how good neighbourly relations and cooperation could be arranged. . . . There is no desire for our states to interfere in one another's domestic affairs, but that friendly relations between these two independent neighbouring states will be preserved." (S.Tr.)

Lesotho became an independent kingdom within the Commonwealth of Nations on October 4, 1966. The first head of state was the paramount chief (Motlotlehi) Moshoeshoe II; Chief Jonathan was the first prime minister. Moshoeshoe's demand for powers wider than those of a constitutional monarch was denied. In January 1970, following elections in which the opposition Basotho Congress Party claimed to have gained a majority, Chief Jonathan proclaimed a state of emergency, suspended the constitution, and placed the King under house arrest. Moshoeshoe was later allowed to go to The Netherlands. A reconciliation was effected between Chief Jonathan and the King, who returned to Lesotho in December 1970 to again become head of state, with the understanding that he would take no part in politics. (S.Tr./Ed.)

Lesotho's relations with South Africa fluctuated under Chief Jonathan and deteriorated in the early 1980s. Seeking friendship and aid elsewhere, Chief Jonathan opened diplomatic relations with China, North Korea, and the Soviet Union. Within Lesotho, political opposition to Chief Jonathan's Basotho National Party increased. In January 1986 troops of the country's paramilitary forces overthrew the government. A six-man military council dissolved the National Assembly and returned executive and legislative powers to the King, who was to act on the advice of the council.

For later developments in the history of Lesotho, see the *Britannica Book of the Year* section in the BRITANNICA WORLD DATA ANNUAL. (Ed.)

MALAŴI

A landlocked southeast African country of dramatic highlands and extensive lakes, the Republic of Malaŵi occupies a narrow, curving strip of land along the East African Rift Valley. Stretching about 520 miles (837 kilometres) from north to south, it has a width varying from five to 100 miles and is bordered by Tanzania to the north, Mozambique to the east and south, and Zambia to the west. Its total area of 45,747 square miles (118,484 square kilometres) includes some 9,400 square miles of lake surface dominated by the 8,900 square miles of Lake Nyasa (known in Malaŵi as Lake Malaŵi). In 1975, the capital was moved from Zomba to Lilongwe, in the Central Region, and the new capital city became the focal point of the nation's development.

Most of Malaŵi's population engages in cash-crop and subsistence agriculture. The country's exports consist of the produce of both small landholdings and large tea and tobacco estates. Malaŵi has successfully attracted capital investment and has made great strides in the exploitation of its natural resources. The country is almost totally dependent upon Mozambique for access to the sea.

Physical and human geography

THE LAND

The four basic regions

Relief. While Malaŵi's landscape is highly varied, there are four basic regions—the East African (or Great) Rift Valley, the central plateaus, the highlands, and the isolated mountains. The Great Rift Valley—by far the dominant feature of the country—is a gigantic troughlike depression running throughout the country from north to south and containing Lake Nyasa and the Shire River valley. The lake's littoral, situated along the western and southern shores and ranging from 5 to 15 miles in width, covers about 8 percent of the total land area and is spotted with swamps and lagoons. The Shire Valley stretches some 250 miles from the southern end of Lake Nyasa at Mangochi to Nsanje at the Mozambique border and contains Lake Malombe at its northern end. The Central Region plateaus rise to an altitude of between 2,500 and 4,500 feet (760 and 1,370 metres) and lie beyond the littoral to the west; the plateaus cover about three-quarters of the total land area. The highland areas are mainly isolated tracts that rise as much as 8,000 feet above sea level. They comprise the Nyika, Viphya, and Dowa highlands and Dedza-Kirk Mountain range in the north and west and the Shire Highlands in the south. The isolated massifs of Mulanje (10,000 feet) and Zomba (7,000 feet) represent the fourth physical region. Surmounting the Shire Highlands, they fall away in the east, running down to the Lake Chilwa–Phalombe Plains.

Drainage and soils. The major drainage system is that of Lake Nyasa, which covers some 11,430 square miles and extends beyond the Malaŵi border. It is fed by the North and South Rukuru, Dwangwa, Lilongwe, and Bua rivers. The Shire River, the lake's only outlet, flows through adjacent Lake Malombe and receives several tributaries before joining the Zambezi River in Mozambique. A second drainage system is that of Lake Chilwa, the rivers of which flow from the Lake Chilwa–Phalombe Plains and the adjacent highlands.

Soils, composed primarily of red earths, with brown soils and yellow gritty clays on the plateaus, are distributed in a complex pattern. Alluvial soils occur on the lakeshores and in the Shire Valley, while other soil types include hydromorphic (excessively moist) soils, black clays, and sandy dunes on the lakeshore.

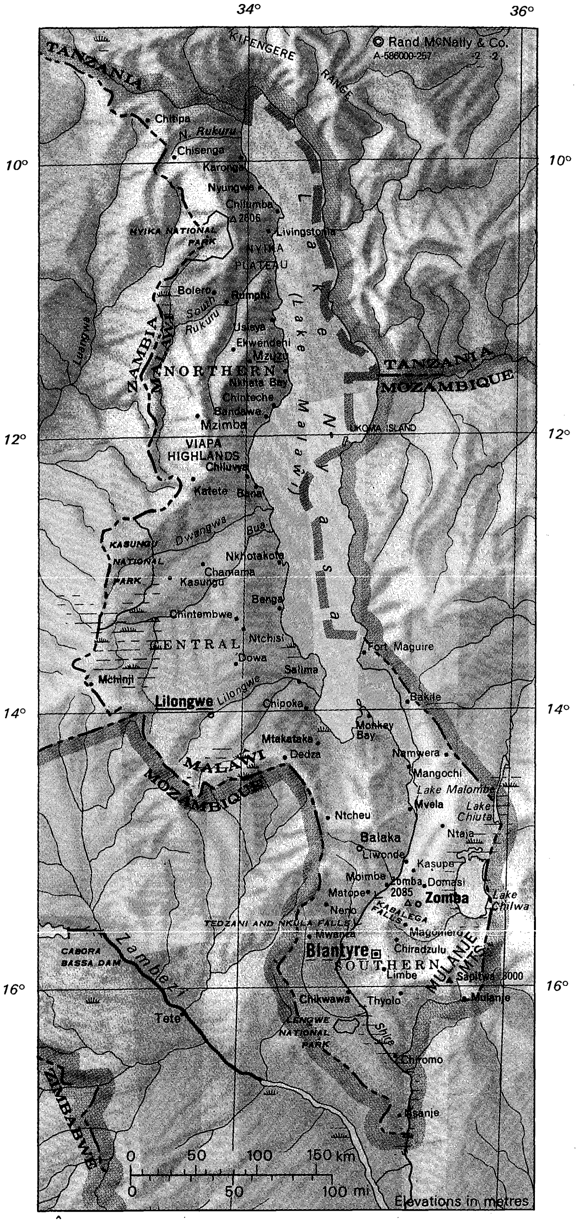
Climate. There are two main seasons—the dry season from May to October and the wet season from November to April. Altitude has an important effect upon temperature. Nsanje (Port Herald), in the Shire River plain, has a mean July temperature of 69° F (21° C) and an October mean of 84° F (29° C), while Dedza, which lies at an altitude of more than 5,000 feet, has a July mean of 57° F (14° C) and an October mean of almost 69° F (21° C). On the Nyika Plateau and on the upper levels of the Mulanje Massif, frosts are not uncommon in July. Annual rainfall is highest over parts of the northern highlands and on the Sapitwa peak of Mulanje Mountain, where it is about 90 inches (2,300 millimetres); it is lowest in the lower Shire Valley, where it ranges from 25 to 35 inches (650 to 900 millimetres).

Diversity of vegetation

Plant and animal life. The natural vegetation pattern reflects diversities in altitude, soils, and climate. Savanna (grassy parkland) occurs in the dry lowland areas. Open woodland with bark cloth trees, or *miombo* (leguminous trees unsuitable for timber), is widespread on the infertile plateaus and escarpments. Woodland, with species of acacia tree, covers isolated, fertile plateau sites and river margins; grass-covered, broad depressions, called *madambo* (singular, *dambo*), dot the plateaus; grassland and evergreen forest are found in conjunction on the highlands and on the Mulanje and Zomba massifs. Swamp vegetation is, however, being greatly altered by human settlement. Much of the original woodland has been cleared, and, at the same time, forests of softwoods are being planted in the highland areas. High population density and intensive cultivation of the Shire Highlands hinder the natural regeneration process; wells have been sunk and rivers dammed to irrigate the dry grasslands for agriculture.

Game animals abound only in the game reserves, where antelope, buffalo, elephants, leopards, lions, rhinoceroses, and zebras occur; hippopotamuses live in Lake Nyasa. The lakes and rivers contain more than 200 species and 13 families of fish. The most common and commercially significant fish include the endemic tilapia, or *chambo* (nest-building freshwater fish); catfish, or *mlamba;* and minnows, or *matemba.*

Settlement patterns. A rural village—called a *mudzi*—is usually small. Organized around the extended family, it is limited by the amount of water and arable land available in the vicinity. On the plateaus, which support the bulk of the population, the most common village sites are at the margins of *madambo,* which are usually contiguous with streams or rivers and are characterized by woodland, grassland, and fertile alluvial soils. In highland areas, scattered villages are located near perennial mountain streams and pockets of thin but arable land. The larger settlements of the Lake Nyasa littoral originated in the 19th century as collection points for slaves and later developed as lakeside ports. Improvements in communication and the sinking of wells in semi-arid areas have permitted the establishment of new settlements in previously uninhabited areas. Architecture is also changing; the traditional round, mud-



MALAŴI

MAP INDEX

Political subdivisions

Central	13·00s	34·00e
Northern	11·00s	34·00e
Southern	15·30s	35·00e

Cities and towns

Bakile	13·58s	35·15e
Balaka	14·59s	34·57e
Bana	12·25s	34·08e
Bandawe	11·57s	34·10e
Benga	13·19s	34·16e
Blantyre	15·47s	35·00e
Bolero	10·59s	33·45e
Chamama	12·55s	33·43e
Chikwawa	16·03s	34·48e
Chilumba	10·28s	34·12e
Chiluvya	12·20s	34·03e
Chinteche	11·52s	34·09e
Chintembwe	13·25s	33·59e
Chipoka	14·00s	34·31e
Chiradzulu	15·42s	35·10e
Chiromo	16·33s	35·08e
Chisenga	9·56s	33·26e
Chitipa	9·43s	33·15e
Dedza	14·22s	34·20e
Domasi	15·18s	35·20e
Dowa	13·40s	33·58e
Ekwendeni	11·23s	33·50e
Fort Maguire	13·08s	34·52e
Karonga	9·56s	33·56e
Kasungu	13·01s	33·30e
Kasupe	15·10s	35·15e
Katete	12·17s	33·39e
Lilongwe	13·59s	33·44e
Limbe	15·49s	35·03e
Livingstonia	10·36s	34·07e
Liwonde	14·52s	35·28e
Magomero	15·36s	35·15e
Mangochi	14·28s	35·16e
Matope	15·20s	34·59e
Mchinji	13·41s	32·55e
Moimbe	15·18s	35·04e
Monkey Bay	14·05s	34·55e
Mtakataka	14·14s	34·31e
Mulanje	16·02s	35·30e
Mvela	14·46s	35·16e
Mwanza	15·37s	34·31e
Mzimba	11·52s	33·34e
Mzuzu	11·27s	33·55e
Namwera	14·22s	35·30e
Neno	15·24s	34·39e
Nkhata Bay	11·33s	34·18e
Nkhotakota	12·57s	34·17e
Nsanje	16·55s	35·12e
Ntaja	14·52s	35·32e
Ntcheu	14·49s	34·38e
Ntchisi	13·22s	34·00e
Nyungwe	10·16s	34·07e
Rumphi	11·01s	33·52e
Salima	13·47s	34·26e
Thyolo	16·04s	35·08e
Usisya	11·09s	34·11e
Zomba	15·23s	35·18e

Physical features and points of interest

Bua, *river*	12·42s	34·13e
Cabora Bassa Dam, reservoir	15·34s	32·38e
Chilwa, Lake	15·12s	35·50e
Chiuta, Lake	14·55s	35·50e
Dwangwa, *river*	12·33s	34·12e
Kabalega Falls	15·23s	34·54e
Kasungu National Park	12·55s	33·15e
Lengwe National Park	16·18s	34·46e
Likoma Island	12·05s	34·45e
Lilongwe, *river*	14·37s	34·25e
Malaŵi, Lake, see Nyasa, Lake		
Malombe, Lake	14·38s	35·12e
Mulanje Mountains	15·58s	35·38e
North Rukuru, *river*	9·53s	33·56e
Nyasa, Lake (Lake Malaŵi)	12·08s	34·30e
Nyika National Park	10·35s	33·45e
Nyika Plateau	10·40s	33·50e
Sapitwa, *peak*	15·58s	35·38e
Shire, *river*	17·42s	35·19e
South Rukuru, *river*	10·46s	34·14e
Tedzani and Nkula Falls	15·32s	34·49e
Viapa Highlands	11·30s	33·50e
Zomba, *peak*	15·17s	35·17e

walled, grass-roofed hut is yielding its place to rectangular brick buildings with corrugated iron roofs.

Urban development began in the colonial era, with the arrival of missionaries, traders, and administrators, and was further stimulated by the construction of the railway. The only true urban centres are Blantyre-Limbe, Zomba, Mzuzu, and Lilongwe. Although some district centres and missionary stations have an urban appearance, they are closely associated with the rural settlements surrounding them. Blantyre, Malaŵi's industrial and commercial centre, is situated in a depression on the Shire Highlands at an altitude of about 3,400 feet. Zomba, seat of the University of Malaŵi, lies at the foot of Zomba Mountain and is purely of administrative origin. Farther north is Lilongwe, Malaŵi's new capital, which is developing agricultural industries.

THE PEOPLE

edomi-
ınt
ıguages

Nine main groups are historically associated with modern Malaŵi—the Chewa, Nyanja, Lomwe, Yao, Tumbuka, Sena, Tonga, Ngoni, and Ngonde (Nkonde). All the African languages spoken belong to the Bantu language family. Although Chichewa and English are the official languages, English was understood in 1966 by less than one-fifth of the population, while Chichewa was spoken by more than half and understood by about three-quarters. Other important languages are Chilomwe, Chiyao, and Chitumbuka.

The major religious groups are traditionalist, Christian, and Muslim. Although these groups appear to be about equal in number, their exact distribution is not known.

The establishment of new rural and urban growth centres, and the opening up of previously inaccessible fertile agricultural areas, has resulted in a significant reversal of emigration. The majority of persons who were working in other parts of southern Africa in 1966 had returned by 1977.

THE ECONOMY

The backbone of the Malaŵi economy is agriculture. The most significant evidence of Malaŵi's economic performance after independence was that the government recurrent budget moved from a significant deficit to a small annual surplus in the 1970s.

Resources. Most of Malaŵi's mineral deposits are neither extensive enough for commercial exploitation nor easily accessible. Exploration and assessment studies continue, however, on minerals such as apatite, located south of Lake Chilwa; bauxite, on the Mulanje Massif; kyanite, on the Dedza-Kirk range; coal, in Karonga District; vermiculite, in Ntcheu District; and rare-earth minerals at Kangankunde Hill in Machinga District.

More than half of Malaŵi's total land area is suitable for cultivation. Some 3,400,000 acres (1,400,000 hectares) are under field crops, with a further 1,600,000 acres (640,000 hectares) lying fallow and about the same area potentially arable. Forests and woodlands cover nearly 9,000 square miles, of which almost 4,000 square miles are state-controlled forest reserves.

The lakes and rivers of Malaŵi provide a rich harvest of fish.

Malaŵi has a vast water supply potential, although some areas have an inadequate amount of water. Treated water for the major cities of Blantyre and Lilongwe is supplied by the Walker's Ferry Scheme and the Kamuzu Dam, respectively. Most of the rivers are seasonal, but a few large ones, particularly the Shire River, have a considerable irrigation and generating potential. The total hydroelectric potential of the country is about 1,200 megawatts, of which more than 500 megawatts can be generated on the Shire River alone. Present power demands, which represent about 8 percent of potential capacity, are met by the Nkula Falls and Tedzani Falls hydroelectric schemes and diesel plants.

Agriculture, fishing, and forestry. The most important agricultural export products are tobacco, tea, sugar, and peanuts (groundnuts). Tea is grown on plantations on the Shire Highlands by the largest proportion of the country's salaried labour force. Tobacco is raised largely on the Central Region plateau. Other important crops are corn (maize; the major staple food), cotton, cassava, coffee, and rice. Between 1966 and 1978, the contribution of cash crops to the gross domestic product averaged about 15 percent a year, while that of subsistence agriculture averaged about 33 percent. The principal cash crops are marketed by the Agricultural Development and Marketing Corporation; a few cooperative societies purchase and market produce.

The amount of fish landed—mainly from Lake Nyasa—is likely to increase with the recent rise in the water level of Lake Chilwa, which had previously shown signs of drying up. Sales of softwood forestry products greatly increased, and Malaŵi is no longer a net importer of wood and wood products. Sawn poles, posts, and manufactured wooden items are produced largely for the domestic market, and some forest products are exported.

Industry. Malaŵi now satisfies its domestic need for products such as cotton textiles, canned foodstuffs, beer, edible oils, soaps, sugar, radios, hoes, and shoes, all of which previously had to be imported. The main demand for electric power is in the industrial areas of the Southern Region, where electricity consumption more than trebled after 1964; the industrial area of Lilongwe; the vast sugar estates of Sucoma and Dwangwa; and the pulpwood scheme of Vipya.

Manufacturing—together with mining and quarrying of clay, granite, limestone, and marble for building and construction—contributes about 6 percent of the annual gross domestic product. Mining may expand as a result of surveys of the bauxite deposits on the Mulanje Massif and of other minerals in promising areas. Exploitation of bauxite will depend on an increased hydroelectric capacity to meet the demand of bauxite smelting for abundant cheap electric power. In addition, it is expected that the Vipya pulpwood afforestation scheme will form the basis for a pulp and paper industry.

Trade and finance. About 95 percent of Malaŵi's exports are derived from agriculture. In the 1970s, however, exports of manufactured products to some neighbouring countries began to increase.

There are two commercial banks—the National Bank of Malaŵi and the Commercial Bank of Malaŵi. The Reserve Bank of Malaŵi is the central bank of the country. Other financial institutions include the Post Office Savings Bank, the New Building Society, and finance houses. Among the several insurance companies, only one is locally based.

The public and private sectors. In the early 1970s the government's economic policy was anti-inflationary, arising from the need to reduce the deficit in public expenditure and to maintain the level of foreign-exchange reserves. In budgetary policies maximum restraint, consistent with development needs and planned reduction of grants-in-aid from the United Kingdom, was exercised.

The main emphasis continues to be directed toward agricultural export production and the completion of investment projects, while at the same time maintaining a satisfactory external balance. The Malaŵi development program is concerned with the public sector only insofar as it does not interfere with the private sector. Priority is given to transport, agriculture, education, and housing.

Taxation, trade unions, and employer associations. A minimum annual tax of kwacha (K) 3.50 is payable by all men over 18 years of age unless they are liable to other taxes. Employed men and women who earn up to K 900 a year pay a graduated tax, and those with higher incomes pay an income tax. Local companies pay taxes at the rate of 40 percent of chargeable income, and companies incorporated outside of Malaŵi pay an additional 5 percent.

There is no sizable industrial labour force. Some 20 trade unions and employer associations are connected with such enterprises as the tea plantations, the building and construction industry, road transport, and railways. The Ministry of Labour plays a significant role in maintaining good relations between employers and employees.

Promoting higher crop yields

Administration of the economy. The government seeks to strengthen the agricultural sector by encouraging integrated land use, higher crop yields, and irrigation schemes. In pursuit of these goals, four large-scale integrated rural development programs, covering one-fifth of the country's land area, were in operation by 1980. They were the Shire Valley Agricultural Development Project, the Lilongwe Land Development Project, the Central Region Lakeshore Development Project, and the Karonga Rural Development Project. These projects include extension services; credit and marketing facilities; physical infrastructures such as roads, buildings, water supplies, health centres, afforestation units, and crop storage and protection facilities. Outside the main program areas, advisory services and educational programs are available and the Malaŵi Young Pioneers, a national youth movement, trains over 2,000 young men and women yearly in techniques of rural development.

Both higher incomes in the rural areas and continued public expenditure are likely to increase the purchasing power of the public as a whole, thus providing a stimulus for further industrial development. The government continues to promote the establishment of manufacturing industries, thus reducing reliance on expensive imported goods, and so strengthening the balance of payments situation, at the same time increasing employment opportunities.

Transportation. *Road and rail network.* Malaŵi has road connections to Chipata on the Zambian border; to Harare (formerly Salisbury), Zimbabwe, via Mwanza and Tete; and to several points on the Mozambique border. The backbone of the road system is represented by a road running from Blantyre in the south to Lilongwe in the west. A lakeshore highway runs roughly parallel to the inland highway from Mangochi to Karonga.

Of Malaŵi's two railway links to the sea, the first stretches more than 570 miles from Lilongwe eastward to the port of Beira on the Mozambique coast; an extension from Lilongwe to Mchinji, on the Zambia border was completed in 1980. The second railroad joins the Salima–Blantyre line at Nkaya Junction to the south of Balaka and travels due east to link with the Mozambique Railways system at Cuamba, from where it continues to the port of Nacala.

Lake transport. Of the rivers, only the Shire is partially navigable, all other streams being broken by rapids and cataracts. Lake Nyasa has long been used as a means of inexpensive transportation. A passenger and cargo service that operates on the lake is linked to Chipoka railway junction about 17 miles south of Salima. The main ports on the lake are Monkey Bay, Nkhotakota, Nkhata Bay, and Likoma Island.

Air services. Air Malaŵi, the national airline, operates services from the main airport at Chileka, 11 miles from Blantyre, to several foreign countries and neighbouring African capitals.

ADMINISTRATIVE AND SOCIAL CONDITIONS

Under the republican constitution of Malaŵi promulgated in July 1966, Parliament is composed of a president, who is head of state and government and of the National Assembly. The cabinet is appointed by the president. The original number of 50 elected members of the assembly was raised to 60 in 1969 and 87 in 1979. In addition, the president can appoint no more than 15 nominated members.

The country is divided into 24 administrative districts. The local government system consists of district councils, the city councils of Blantyre and Lilongwe, the municipality of Zomba, and seven town councils.

Malaŵi was a de facto one-party state from August 1961, when the first general elections were held, until 1966, when the constitution formally recognized the Malaŵi

Congress Party—led by Pres. H. Kamuzu Banda—as the sole political organization. According to the constitution, elections for the presidency and the assembly are to be held every five years; general elections, however, have been held only in 1971 and 1978. The presidential candidate was nominated by an electoral college composed of party officials at the national, regional, and district levels; the League of Malaŵi Women; the League of Malaŵi Youth; members of Parliament; recognized chiefs; and all chairmen of district councils. In 1971, Banda was elected president for life. Candidates for the National Assembly may stand for election only after their nomination by the district party conferences.

:lections

Justice. The judiciary is based upon the system prevailing in the British colonial era and Malaŵi traditional law. It consists of a Supreme Court of Appeal, a High Court, magistrates' courts, and traditional courts. Since 1969, criminal cases involving witchcraft or local superstition, for which the death penalty can be imposed, have been tried in the traditional courts instead of the High Court. The minister of justice has the power to direct a particular case or group of cases to a particular court; cases tried in the traditional courts can be appealed to the National Traditional Court of Appeal.

Health and welfare. Health facilities include one central hospital, general hospitals, district hospitals and clinics, Zomba Mental Hospital and Kochira Leprosarium. The most common diseases are malaria, bilharzia, and trachoma. The main emphasis of the health services is unavoidably on curative medicine.

An acute shortage of housing has existed for several years in urban areas. The Malaŵi Housing Corporation has, however, launched several projects to build houses and develop traditional housing areas.

ducation

Education. Elementary education in the primary schools is provided by local education authorities. Post-primary education comprises a four-year secondary school course that can lead to a university education. There are also institutions for teacher training and for technical and vocational training. Because of limited resources, only about a third of the school-age population is enrolled in schools. The Malaŵi Correspondence College is available to many students. The Kamuzu Academy at Mtunthama is a secondary school for gifted children. The University of Malaŵi, founded in 1965, has three constituent colleges.

CULTURAL LIFE

Though under the impact of modernization, Malaŵi's traditional culture is characterized by continuity as well as change, and the traditional life of the village has remained largely intact. One of the most distinctive features of Malaŵi culture is the enormous variety of traditional songs and dances that use the drum as the major musical instrument. Among the most notable of these dances are *ingoma* and *gule wa mkulu* for men, and *chimtali* and *visekese* for women. There are various traditional arts and crafts, including sculpture in wood and ivory. There are two museums—the Museum of Malaŵi in Blantyre and a smaller one in Mangochi. While various cultural activities are organized by the Ministry of Youth and Culture, the University of Malaŵi Travelling Theatre, and other groups in Blantyre, the radio from Zomba and Lilongwe has proved to be the most effective means of bringing traditional and modern plays to the rural population.

For statistical data on the land and people of Malaŵi, see the *Britannica World Data* section in the BRITANNICA WORLD DATA ANNUAL. (Z.D.K.)

History

The paleontological record of human cultural artifacts in Malaŵi dates back more than 50,000 years, although known fossil remains of early *Homo sapiens* belong to the period between 8000 and 2000 BC. These prehistoric forebears have affinities to the San (Bushmen) of southern Africa and were probably ancestral to the Twa and Fula, whom Bantu-speaking peoples claimed to have found when they invaded the Malaŵi region between the 1st and 4th centuries AD. From then to about AD 1200, Bantu settlement patterns spread, as did ironworking and the slash-and-burn method of cultivation. The identity of these early Bantu-speaking inhabitants is uncertain. According to oral tradition, names such as Kalimanjira, Katanga, and Zimba are associated with them.

With the arrival of another wave of Bantu-speaking peoples between the 13th and 15th centuries AD, the recorded history of the Malaŵi region began. These peoples migrated into the region from the north and they interacted with and assimilated the earlier pre-Bantu and Bantu inhabitants. The descendants of these peoples maintained a rich oral history, and from 1500 written records were kept in Portuguese and English.

Early political units

Among the notable accomplishments of the last group of Bantu immigrants was the creation of political states or the introduction of centralized systems of government. They established the Maravi Confederacy around 1480. During the 16th century, the confederacy encompassed the greater part of what is now central and southern Malaŵi, and, at the height of its influence in the 17th century, its system of government affected peoples in the adjacent areas of modern Zambia and Mozambique. North of the Maravi territory, the Ngonde founded a kingdom around 1600; in the 18th century, a group of immigrants from the eastern side of Lake Nyasa created the Chikulamayembe state to the south of the Ngonde.

The pre-colonial period witnessed other important developments. In the 18th and 19th centuries, better and more productive agricultural practices were adopted. In some parts of the Malaŵi region, shifting cultivation of indigenous varieties of millet and sorghum began to give way to more intensive cultivation of crops with a higher carbohydrate content, such as corn (maize), cassava, and rice.

The independent growth of indigenous governments and improved economic systems was severely disturbed by the development of the slave trade in the late 18th century and by the arrival of foreign intruders in the late 19th century. The slave trade in Malaŵi increased dramatically between 1790 and 1860 because of the growing demand for slaves on Africa's east coast. Swahili-speaking people from the east coast and the Ngoni and Yao peoples entered the Malaŵi region between 1830 and 1860 as traders or as armed refugees fleeing the Zulu Empire to the south. All of them eventually created spheres of influence within which they became the dominant ruling class. The Swahili speakers and the Yao also played a major role in the slave trade.

Islām and Christianity

Islām spread into Malaŵi from the east coast. It was first introduced at Nkhotakota by the ruling Swahili-speaking slave trader, the Jumbe, in the 1860s. Traders returning from the coast in the 1870s and 1880s brought Islām to the Yao of the Shire Highlands. Christianity was introduced in the 1860s by David Livingstone and by other Scottish missionaries who came to Malaŵi after his death in 1873. Missionaries of the Dutch Reformed Church of South Africa and the White Fathers of the Roman Catholic Church arrived in the 1880s and 1890s.

Christianity owed its success to the protection given to the missionaries by the colonial government, which the British established after occupying the Malaŵi region in the 1880s and 1890s. British colonial authority was welcomed by the missionaries and some African societies, but was strongly resisted by the Yao, Chewa, and others. In 1891 the British established the Nyasaland Districts Protectorate, which was called the British Central Africa Protectorate from 1893 and Nyasaland from 1907.

Many changes for the better took place under the colonial regime. Roads and railways were built, the cultivation of cash crops by European settlers was introduced, and inhumane practices were suppressed. On the other hand, the colonial administration did little to enhance the welfare of the African majority because of its commitment to the interests of the European settlers. It failed to develop African agriculture, and many able-bodied men migrated to neighbouring countries to seek employment. Furthermore, between 1951 and 1953 the colonial government decided to join the colonies of Southern and Northern Rhodesia and Nyasaland into a federation, against bitter opposition from their African inhabitants.

These negative features of colonial rule prompted the rise of a nationalist movement. From its humble beginnings during the period between the world wars, African nationalism gathered momentum in the early 1950s. Of special impetus was the imposition of the federation, which nationalists feared as an extension of colonial power. The full force of nationalism as an instrument of change became evident after 1958 under the leadership of Hastings Kamuzu Banda. The federation was dissolved in 1963 and Malaŵi became independent as a member of the Commonwealth of Nations on July 6, 1964. (Z.D.K./K.M.G.P.)

Soon after independence, a serious dispute arose between Banda, the prime minister, and other ministers. In September 1964 three ministers were dismissed and three others resigned in sympathy. Henry Chipembere, one of the dismissed ministers, escaped from house arrest and defied attempts to recapture him, becoming thenceforth the focus for antigovernment opinion. On July 6, 1966, Malaŵi became a republic, and Banda was elected president. He was made president for life in 1971.

In external relations Banda pursued a policy at variance with that of most of the recently independent countries of tropical Africa. He accused the Organization of African Unity (OAU) of adopting attitudes that it was powerless to support, established friendly trading relations with the Republic of South Africa, and appealed to other African leaders to be more realistic in their attitude toward racial problems. (J.C.Mi./K.In./Ed.)

Malaŵi joined the Southern African Development Coordination Conference, a union of majority-ruled nations neighbouring South Africa, in 1980 but refused to sever its formal diplomatic links with South Africa.

For later developments in the history of Malaŵi, see the *Britannica Book of the Year* section in the BRITANNICA WORLD DATA ANNUAL. (Ed.)

MOZAMBIQUE

The People's Republic of Mozambique is located on the southeastern coast of Africa between latitudes 10°27′ and 26°52′ south. Its elongated territory of 308,642 square miles (799,380 square kilometres) has an Indian Ocean coastline of 1,737 miles (2,795 kilometres). Mozambique is bounded on the north by Tanzania, on the west by Malaŵi, Zambia, Zimbabwe, and the Transvaal province of South Africa, and on the south by Swaziland and the Natal province of South Africa. To the east lies the Mozambique Channel, which separates the African mainland from the island-state of Madagascar (formerly Malagasy). The capital of Mozambique is Maputo (formerly Lourenço Marques).

Physical and human geography

THE LAND

Plain, plateau, mountains

Relief. Mozambique occupies southern Africa's largest coastal plain; nearly half of the total area lies below 750 feet (230 metres). This coastal lowland is widest in the south, where it covers virtually the entire width of the territory, but northward it becomes narrower. In central and northern Mozambique an interior plateau develops, with elevations ranging from 500 to 2,000 feet. Toward the western boundaries with Malaŵi, Zambia, and Zimbabwe, this plateau rises still higher, and mountainous terrain occurs. Highest elevations along the Zambia–Malaŵi border range from 5,741 feet extreme north, near Lake Nyasa (Malaŵi) in Niassa Province, to 7,241 feet in the northwest, near the intersection of the Zambia–Malaŵi boundary in Tete Province.

The lengthy coastline is marked by several distinct features. In the extreme south lies Delagoa Bay, the best natural harbour on the East African coast, and on its shore the capital of Mozambique, Maputo, has grown. Northward, from Delagoa Bay to the delta of the Zambezi River, the coast is sandy and swampy, with mangroves and offshore bars. North of the Zambezi Delta, however, sandy stretches of coast are interrupted by rocky cliffs and headlands, as well as islands, many of which are of coral formation. In addition to the Zambezi, some 50 other rivers reach the coast of Mozambique, including the Limpopo, the Save, the Búzi, the Lúrio, and the Messalo. In the far north the Rovuma River forms the boundary between Mozambique and Tanzania.

The Zambezi River

The pivotal geographical feature of Mozambique is the Zambezi River. Almost all the higher and more rugged terrain lies to the north of the river, while to the south the gently undulating coastal lowland prevails. Apart from the higher elevations near the Rhodesian Plateau, there is only one marked feature of positive relief located south of the Zambezi: the Serra da Gorongosa in Sofala Province. Elsewhere, southern Mozambique presents a gently rolling, rather featureless landscape, with an occasional group of low hills, broken by the wide alluvial valleys of the rivers.

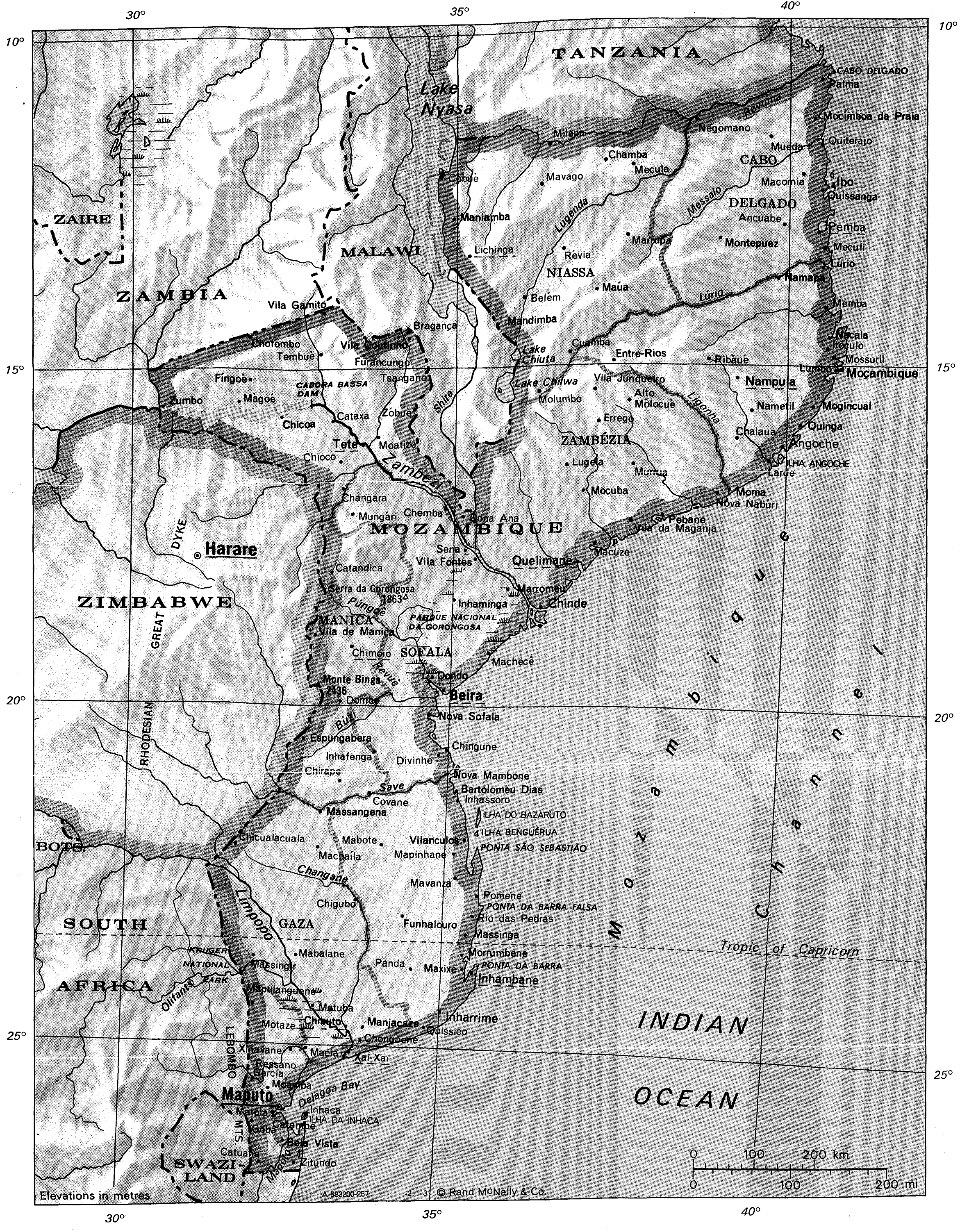
Drainage. Mozambique's topography generally declines in elevation from west to east, so that, from the Rovuma River in the north to the Limpopo in the south, streams have an eastward orientation. The higher parts of the territory are sustained by the crystalline rocks of the ancient African basement complex, or shield. In the Zambezi Basin and on the coastal plain, dipping toward the Mozambique Channel, lie more recent sedimentary rocks deposited during the Triassic Period (about 260,000,000 to 200,000,000 years ago). Still more recent sediments, of Late Cretaceous and Eocene age, lie in the eastern parts of the coastal plain. A notable feature is the rhyolite-sustained Lebombo Mountains, which coincide with the political boundary between Mozambique and Swaziland and also extend northward along the Transvaal border.

With the exception of northern Mozambique east of Lake Nyasa, Mozambique's major rivers rise to the west of the territory, on the African Highveld. The headwaters of the Limpopo lie on the margins of the Kalahari; the Save River rises on the Great Dyke of Zimbabwe, and the Zambezi rises in the Angola–Zambia borderland. Since the interior plateaus receive moderate rainfall at best, with marked seasonal variation, water levels alter considerably. The largest drainage basin is that of the Zambezi, which, with its Shire tributary (the outlet of Lake Nyasa), drains most of the centrally situated provinces of Tete, Manica, Sofala, and Zambézia.

Soils. Good soils are the exception rather than the rule in Mozambique. The soils of the plateau areas, derived from underlying granites, gneisses, and schists, are leached and rich in iron and aluminum and are generally rather infertile. In the dry interior to the north of the Limpopo Valley, sandy, iron-bearing soils prevail, their high permeability increasing the area's dryness. Along the coastal plain there are waterlogged organic soils, with some poorly developed alluvials. In the alluvial lowlands of several of the river basins, and in the *machongas* (patches of humus-rich, moist lowland between the dunes of the coastal plain) lie Mozambique's most productive soils.

Rainfall

Climate. As Mozambique lies on the tropical East African coast, apparently in the path of the southeast trade winds, and as the warm Mozambique Current flows southward offshore, a prevailingly rainy, moist climate would be expected, especially since the mountains of the interior cause air currents to rise, increasing the probability of rain. Much of Mozambique, however, is subhumid, and drought is common in many areas. Maputo, for example, receives a mere 30 inches (750 millimetres) of rain in an average year, Inhambane 36, Beira 60, Angoche 42, and Tete 23. With the prevailing tropical temperatures (Maputo has an annual average of 73° F [23° C], while Quelimane's warmest month averages 86° F [30° C] and its coolest month 68° F [20° C]), much of this rainfall is lost by evaporation or by plant transpiration. Mozambique north of the Zambezi is a region of greater moisture and lesser drought incidence (although the Zambezi Basin around Tete is itself deficient in moisture); the south suf-



MOZAMBIQUE

MAP INDEX

Political subdivisions

Cabo Delgado	12 00 S 40 00 E
Gaza	23 00 S 33 00 E
Inhambane	23 00 S 35 00 E
Manica	19 00 S 33 30 E
Maputo	26 00 S 32 30 E
Nampula	15 00 S 39 00 E
Niassa	13 00 S 36 30 E
Sofala	19 00 S 35 00 E
Tete	15 30 S 33 00 E
Zambézia	17 00 S 37 00 E

Cities and towns

Alto Molócuè	15 38 S 37 42 E
Ancuabe	12 58 S 39 54 E
Angoche	16 14 S 39 58 E
António Enes, see Angoche	
Bartolomeu Dias	21 11 S 35 07 E
Beira	19 49 S 34 52 E
Bela Vista	26 20 S 32 40 E
Belém	14 13 S 35 58 E
Bragança	14 23 S 34 20 E
Catandica	18 03 S 33 11 E
Cataxa	15 58 S 33 12 E
Catembe	26 00 S 32 33 E
Catuane	26 48 S 32 18 E
Chalaua	16 06 S 39 11 E
Chamba	12 07 S 36 57 E
Changara	16 54 S 33 14 E
Chemba	17 08 S 34 52 E
Chibuto	24 44 S 33 33 E
Chicoa	15 37 S 32 24 E
Chicualacuala	22 05 S 31 40 E
Chigubo	22 50 S 33 34 E
Chimoio	19 08 S 33 29 E
Chinde	18 37 S 36 24 E
Chingune	20 38 S 34 55 E
Chioco	16 25 S 32 50 E
Chirape	21 20 S 33 33 E
Chofombo	14 35 S 31 50 E
Chongoene	25 00 S 33 47 E
Cóbuè	12 04 S 34 50 E
Covane	21 22 S 33 56 E
Cuamba	14 49 S 36 33 E
Divinhe	20 40 S 34 49 E
Dombe	19 59 S 33 25 E
Dona Ana	17 25 S 35 07 E
Dondo	19 36 S 34 44 E
Entre Rios	14 57 S 37 20 E
Errego	16 02 S 37 14 E
Espungabera	20 29 S 32 48 E
Fíngoè	15 12 S 31 50 E
Funhalouro	23 03 S 34 25 E
Furancungo	14 55 S 33 35 E
Goba	26 12 S 32 08 E
Ibo	12 20 S 40 35 E
Inhaca	26 01 S 32 58 E
Inhafenga	20 35 S 33 53 E
Inhambane	23 51 S 35 29 E
Inhaminga	18 24 S 35 00 E
Inharrime	24 29 S 35 01 E
Inhassoro	21 33 S 35 11 E
Itoculo	14 42 S 40 18 E
João Belo, see Xai-Xai	
Larde	16 28 S 39 43 E
Lourenço Marques, see Maputo	
Lugela	16 25 S 36 43 E
Lumbo	15 00 S 40 44 E
Lúrio	13 35 S 40 30 E
Mabalane	23 37 S 32 31 E

Mabote	22 03 s	34 09 E
Machaíla	22 15 s	32 55 E
Machecé	19 17 s	35 33 E
Macia	25 03 s	33 10 E
Macomia	12 15 s	40 08 E
Macuze	17 42 s	37 11 E
Mágoé	15 48 s	31 43 E
Malvérnia, see Chicualacuala		
Mandimba	14 21 s	35 39 E
Maniamba	12 43 s	35 00 E
Manjacaze	24 44 s	33 53 E
Mapinhane	22 19 s	35 03 E
Mapulanguene	24 29 s	32 06 E
Maputo	25 58 s	32 35 E
Marromeu	18 20 s	35 56 E
Marrupa	13 08 s	37 30 E
Massangena	21 32 s	32 57 E
Massinga	23 20 s	35 25 E
Massingir	23 51 s	32 04 E
Matola	25 49 s	32 27 E
Matuba	24 27 s	32 55 E
Maúa	13 51 s	37 10 E
Mavago	12 22 s	36 10 E
Mavanza	22 43 s	35 08 E
Maxixe	23 51 s	35 21 E
Mecúfi	13 17 s	40 30 E
Mecula	12 04 s	37 40 E
Memba	14 11 s	40 30 E
Milepa	11 43 s	36 20 E
Moamba	25 36 s	32 15 E
Moatize	16 08 s	33 45 E
Moçambique	15 03 s	40 42 E
Mocímboa da Praia	11 20 s	40 21 E
Mocuba	16 50 s	36 59 E
Mogincual	15 35 s	40 25 E
Molumbo	15 27 s	30 15 E
Moma	16 44 s	39 14 E
Montepuez	13 07 s	39 00 E
Morrumbene	23 39 s	35 20 E
Mossuril	14 58 s	40 42 E
Motaze	24 48 s	32 52 E
Mueda	11 39 s	39 33 E
Mungári	17 12 s	33 31 E
Murrua	16 20 s	37 56 E
Nacala	14 34 s	40 41 E
Namapa	13 43 s	39 50 E
Nametil	15 43 s	39 21 E
Nampula	15 07 s	39 15 E
Negomano	11 27 s	38 31 E
Nova Freixo, see Cuamba		
Nova Mambone	20 59 s	35 01 E
Nova Nabúri	16 46 s	38 57 E
Nova Sofala	20 09 s	34 42 E
Pafúri	22 27 s	31 21 E
Palma	10 46 s	40 29 E
Panda	24 02 s	34 45 E
Pebane	17 10 s	38 08 E
Pemba	12 58 s	40 30 E
Pomene	22 53 s	35 33 E
Porto Amélia, see Pemba		
Quelimane	17 53 s	36 51 E
Quinga	15 49 s	40 15 E
Quissanga	12 25 s	40 29 E
Quissico	24 42 s	34 44 E
Quiterajo	11 48 s	40 25 E
Ressano Garcia	25 27 s	32 00 E
Révia	13 23 s	36 31 E
Ribauè	14 57 s	38 17 E
Rio das Pedras	23 12 s	35 23 E
Sena	17 27 s	35 00 E
Tembuè	14 52 s	32 58 E
Tete	16 13 s	33 35 E
Tsangano	15 08 s	34 32 E
Valadim, see Mavago		
Vila Coutinho	14 37 s	34 19 E
Vila da Maganja	17 18 s	37 30 E
Vila de Manica	18 56 s	32 53 E
Vila Fontes	17 50 s	35 21 E
Vila Gamito	14 12 s	33 00 E
Vila Gouveia, see Catandica		
Vila Junqueiro	15 25 s	36 58 E
Vilanculos	22 01 s	35 19 E
Vila Pery, see Chimoio		
Xai-Xai	25 02 s	33 34 E
Xinavane	25 02 s	32 47 E
Zitundo	26 45 s	32 50 E
Zóbuè	15 38 s	34 26 E
Zumbo	15 36 s	30 25 E

Physical features and points of interest

Barra, Ponta da, *point*	23 45 s	35 30 E
Barra Falsa, Ponta da, *point*	22 55 s	35 37 E
Bazaruto, Ilha do, *island*	21 40 s	35 28 E
Benguérua, Ilha, *island*	21 58 s	35 28 E
Binga, Monte, *mountain*	19 45 s	33 04 E
Búzi, *river*	19 52 s	34 46 E
Cabora Bassa Dam	15 34 s	32 38 E
Changane, *river*	24 43 s	33 32 E
Chilwa, Lake	15 12 s	35 50 E
Chiúta, Lake	14 55 s	35 50 E
Delagoa Bay	25 48 s	32 51 E
Delgado, Cabo, *cape*	10 40 s	40 35 E
Gorongosa, Parque Nacional *da, national park*	18 45 s	34 15 E
Gorongosa, Serra da, *inselberg*	18 30 s	34 03 E
Inhaca, Ilha da, *island*	26 03 s	32 57 E
Lebombo Mountains	26 15 s	32 00 E
Ligonha, *river*	16 54 s	39 09 E
Limpopo, *river*	25 15 s	33 30 E
Lourenço Marques, Baíe de, see Delagoa Bay		
Lugenda, *river*	11 25 s	38 33 E
Lúrio, *river*	13 35 s	40 32 E
Maputo, *river*	26 11 s	32 42 E
Messalo, *river*	11 40 s	40 26 E
Mozambique Channel	20 09 s	40 00 E
Nyasa, Lake	12 00 s	34 30 E
Púngoè, *river*	19 50 s	34 48 E
Revuè, *river*	19 50 s	34 02 E
Rovuma, *river*	10 29 s	40 28 E
São Sebastião, Ponta, *point*	22 07 s	35 30 E
Save, *river*	21 00 s	35 02 E
Shire, *river*	17 42 s	35 19 E
Zambezi, *river*	18 55 s	36 04 E

fers from drought and low rainfall, except along the littoral and against the Rhodesian Plateau slopes. Only north central and northwestern Mozambique generally maintain a favourable moisture balance. (H.J. de B./A.C.G.B.)

There is a distinct rainy season from October (or November) until April. The dry season usually lasts from April into September. It is altitude that moderates temperatures. Although Maputo lies nearly 1,000 miles south of Pemba, its annual average temperature is only slightly lower and its annual range only slightly greater. Chillier nights and cooler winter months are recorded in the higher zones along the western borders.

Palm trees

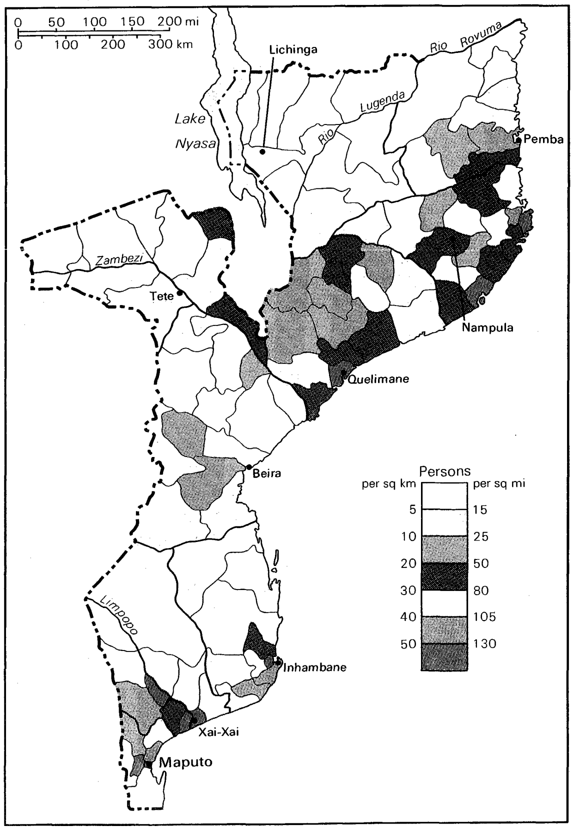
Plant and animal life. Mozambique's vegetation is mostly tropical forest and savanna. Along the coast, especially between 17° and 20° south, the coconut palm is common. Several other types of palm, including the date palm, also occur. Patches of forest, which include stands of ironwood and ebony, are generally found on better drained slopes amid savanna country. In the narrow ravines, tree growth is dense, including the large Senegal khaya, resembling mahogany and used for timber. Along the upper reaches of the major rivers, notably the Zambezi and the Limpopo, the mopani tree—a type of ironwood producing hard, durable timber—prevails on the savanna, where the baobab is also a characteristic feature. Areas of mangrove exist along the coast, especially in the Zambezi Delta. Bamboo and spear grass occur abundantly along the river banks and in the marshy areas.

Despite excessive hunting, Mozambique still has a rich fauna. Lions are numerous. The cheetah, spotted hyena, jackal, serval, civet, mongoose, and wild dog occur in several areas. The elephant is common, as is the black rhinoceros; the white rhinoceros occurs, although more rarely. The lower courses of many of Mozambique's rivers are inhabited by the hippopotamus. Among the herd animals, antelope, buffalo, and zebra are often seen. Giraffes, warthogs, and monkeys are common, as are snakes, especially the python, cobra, puff adder, and viper. Crocodiles abound in the rivers.

Mozambique's birdlife is varied. Flamingoes inhabit the northeast, and species of cranes, storks, herons, pelicans, and ibises occur. The eagle, buzzard, vulture, and crow are omnipresent. Guinea fowl, partridge, bustard, quail, wild goose, and wild duck are all found in large numbers. (H.J. de B./A.C.G.B./Ed.)

Settlement patterns. The Zambezi River forms a significant dividing line between northern and southern Mozambique. South of the river the colonial influence is stronger: there Portugal established its most profitable agricultural settlements, which yielded cash crops whose primary markets lay in Europe. There also Portugal invested most heavily in roads and railroads to link the coastal ports of Maputo (then Lourenço Marques) and Beira with the prosperous mines, cities, and farms of South Africa and what was then Rhodesia. It was from southern Mozambique that tens of thousands of African workers were recruited each year to supply South Africa's mines and farms with labour under a system that was tantamount to involuntary servitude to the state.

Whereas southern Mozambique contained the majority



Population density of Mozambique.

of the approximately 250,000 Europeans just before independence, some of the country's most densely settled African regions were in the north. The northern provinces of Nampula and Zambêzia together contain more than 40 percent of the country's total population. With its larger African population, its remoteness from the capital, its rudimentary transport system, and its lengthy boundary with Tanzania, northern Mozambique became the staging ground for the revolutionary struggle against Portuguese colonialism. There are other factors that distinguish the north from the south, among them the lack of economic opportunity inland, where Malaŵi and Zambia generate far less employment and trade than South Africa and Zimbabwe do, the scarcity of alternatives to subsistence agriculture, the imprint of Islām on some of the African peoples, and the generally stronger resistance to change.

Rural population

Approximately nine-tenths of Mozambique's population is rural, the vast majority of the people living in dispersed homesteads or small hamlets or villages. Most villages comprise clusters of houses grouped around either a cattle kraal or a central meeting place such as a school or agricultural cooperative building. Around each village are the cultivated lands, which by tradition were at the disposal of the community and after independence became the property of the state, which encouraged the peasantry to farm collectively. Manioc is the main food staple, especially along the coastal plains, while corn (maize) is widely grown. Yields of both crops tend to be low and the annual production highly variable. Famine remains a local threat.

Much of rural Mozambique is thinly populated bush country, but pressure on the land occurs in the coastal areas of Nampula and Zambêzia provinces and between Inhambane and Maputo. Because of past slave raiding, endemic sleeping sickness, and low soil fertility, Niassa Province in the north remains the country's area of lowest population density.

Considerable agglomerations of people exist in the hinterlands of the coastal towns, as a result of both greater soil fertility and the drawing power of the towns and cities such as Maputo, Xai-Xai (former João Belo), Inhambane, Beira, Quelimane, Angoche (former António Enes), and Lumbo, near Nampula. In these areas the traditional rural homesteads or hamlets are replaced by more intensive vegetable farms, fields of rice and maize, and some sugar and citrus plantations. The plantations and large-scale agricultural schemes, established during the colonial era, were nationalized as state farms in 1977.

Maputo nd Beira

Mozambique's largest cities are Maputo and Beira. Both are modern, attractive, and well-planned cities, with wide avenues, terraced public gardens, elaborate public buildings, and large apartment buildings. In the suburbs are villas and smaller houses in the characteristic Portuguese style. Many apartments—abandoned by the Portuguese when the government nationalized housing, abolished private rents, and prohibited individual ownership of more than two homes—were taken over by squatters who had lived in the crowded ghettos on the outskirts of the cities. The shanty towns continue to grow, swollen by Mozambicans in search of urban opportunities.

THE PEOPLE

Ethnic groups. Three major historical events shaped the peopling of Mozambique: the great Bantu migration from the north and west; the northward movement that followed the rise of the Zulu Empire in Natal to the south; and the penetration of Arabs, Persians, Portuguese, and other non-Africans from the east. The result is an ethnically distinct north and south, the Zambezi River being the boundary. North of the river the people are predominantly matrilineal in reckoning descent and are accustomed to a pattern of shifting agriculture; south of the Zambezi the people are predominantly sedentary farmers and pastoralists who trace their origins patrilineally.

frican thnic roups

Almost all of Mozambique's population is African. The balance is composed of Europeans, Asians, and persons of mixed racial origin. The Makua–Lomwe is the largest of Mozambique's 10 major ethnic groups. Makua and Lomwe are presently concentrated between Malaŵi and the east coast, having migrated there from the mountainous upper reaches of the Zambezi River two centuries ago. Most are subsistence farmers, although some produce cash crops and others are coastal fishermen. North of them and extending across the Rovuma River into Tanzania are the Yao and Makonde peoples. The Makonde are traditionally conservative and have resisted the Islāmization that is common among the coastal peoples. Makonde art, especially wooden masks and statuary, is acknowledged to be the finest in eastern Africa. The Yao, an Islāmized agricultural people who live in the remote northwest corner of the country, were the principal intermediaries in the slave and ivory trade between Arabs and interior African tribes during the 18th and 19th centuries. The Zambezi Valley is settled by many small fragmented groups.

Southern Mozambique is inhabited by the Tsonga, Karanga, Shona, Chopi, and Nguni people, who belong to the East African cattle-raising complex. For them, cattle traditionally symbolized wealth and prestige. The Tsonga, Mozambique's second largest population group, live between the Save (Sabi) River and Delagoa Bay. They repeatedly were the largest component of Mozambique's migrant labour force to South Africa. The Karanga Shona peoples, living between the Save and Zambezi rivers, are descendants of the Mwene Matapa Empire that thrived in Mozambique and adjoining Zimbabwe five centuries ago.

Europeans

Mozambique's Europeans, Asians, and mestizos are few but economically influential. In 1930 the territory had only some 17,000 whites (mainly Portuguese) and in 1950 about 48,000. During the 1950s, the European population nearly doubled, and on the eve of independence it had reached an estimated 250,000. In late 1978, fewer than 15,000 Portuguese remained. White immigration was encouraged by colonial Portugal in order to exploit the agricultural and mineral resources and to gain more effective control over the black population. Portuguese settlers had been offered land and loans by the colonial authorities, and elaborate settlement schemes were promoted in the valleys of the Limpopo, Revué, Zambezi, and other rivers.

Most Portuguese, however, were urban dwellers; almost half of them resided in and around the capital. They held a virtual monopoly not only of professional and skilled but also of semiskilled jobs. Their sudden withdrawal in 1974–75, accompanied by the willful destruction of property and equipment, threw Mozambique's economy into chaos.

In addition to the European minority, there are Euro-Africans and persons of Chinese, Goan, and Pakistani origin. Almost all of them are urban dwellers, employed in commerce and light industry.

Languages. Several distinct Bantu languages are spoken: Yao and Makua predominate in the north, Tsonga in the south, and Nyanja in the Zambezi Valley. None is spoken widely enough to be adopted as the national language. Swahili comes closest to serving as a lingua franca; it is spoken in the towns and villages along the coast, especially north of the Zambezi. Portuguese is the official language, but it is rarely used beyond government circles, the urban areas, and places where Portugal's influence was strongest. Less than one-tenth of the population is literate.

Swahili

Religion. A similar variation occurs in religious affiliations. In the north, especially among the Yao of Niassa Province, Islāmic traditions are superimposed on essentially animist societies. One-tenth of the entire population may be Muslim. In the south, Christianity, especially Roman Catholicism, has made deeper inroads. The constitution "guarantees the freedom of citizens to practice or not to practice a religion."

Society. During the colonial era, Mozambique's society was highly stratified. In 1927 the population was divided by law into two categories: *indígenas* (indigenous or unassimilated Africans) and *nao indígenas* (Europeans, Asians, *mestiços,* and *assimilados,* or "assimilated" Africans). The *indígena* had few rights, was required to pay taxes and to carry a passbook, and was subject to labour regulations and curfews. He could be conscripted into labour gangs for both public and private enterprise or to farm as directed by the authorities. The *nao indígena* enjoyed full rights of Portuguese citizenship.

Africans could become *assimilados* if they met stringent standards: they had to speak Portuguese, abandon their traditional way of life, be legally employed in commerce or industry, be financially solvent, and supply testimonials of their good character to the authorities. Failure to maintain these "responsibilities" could mean the revocation of the privilege. In 1961 the *indígena* system was abolished, and full citizenship was granted to all Mozambicans.

Mozambique's Frelimo (Frente de Libertação de Moçambique) government is committed to building a nonracist and nontribal socialist society. Frelimo's leadership is explicitly atheist and hostile to the Roman Catholic Church, which it considers to have been a bulwark of the colonial regime. Given the polyglot nature of the people, the institutional diversity, the traditionalism that prevails in many regions, and the paucity of both transport and communications, the government's goals have been difficult to achieve. Fundamental to the indoctrination of socialism is the establishment of "dynamizing groups" (*grupos dinamizadores*) in the communal villages, collective farms, and reorganized businesses and industries. These groups are the basic political links between the people and the Frelimo party. Their function is to involve all the people in discussion of national and local policy issues and programs and of possible solutions to the country's social and economic problems.

There has always been considerable internal migration in Mozambique in addition to large outmigrations from the southern provinces to South Africa and Zimbabwe. Domestic labour shortages created by these international movements were often filled by recruits from the north, especially from Zambézia and Nampula provinces. A pattern of male migration gave the cities an overly male and ethnically heterogeneous character. Like most cities in the Third World, those of Mozambique have a high degree of unemployment and inadequate housing, schools, and social services for the poor.

During the revolutionary war of independence, tens of thousands of Mozambicans fled across the borders to Tanzania, Malaŵi, and Zambia. Between late 1974 and mid-1976 the office of the UN High Commissioner for Refugees assisted in the repatriation of about 35,000 Mozambican refugees from Malaŵi and Zimbabwe and another 80,000 from Tanzania.

THE ECONOMY

Economic policy pursued during the colonial era neglected domestic development, notably in the rural African sector, and it denied most Africans access to technical skills and managerial experience. To protect manufacturing interests in Portugal, industrial development was discouraged. Little was done to exploit the country's mineral resources or to create an infrastructure that would support a self-sufficient, integrated, and growing economic system. The transport system, especially that of railroads, was designed for transit trade with South Africa and Rhodesia more than for developing the local economy and peoples.

Cash crops

Agriculture. Cash crops, primarily cashew nuts, cotton, sugar, tea, copra, and sisal, represented a critical sector of the colonial economy. Agricultural products during the last two decades of colonial rule accounted for at least three-quarters of Mozambique's foreign trade. To stimulate production, the colonial regime introduced a labour system known as *shibalo*, a word derived from the Swahili word *shiba,* meaning "serf," by which African peasants were compelled to cultivate specific cash crops. In the northern areas, for example, more than 500,000 peasants were forced to grow cotton instead of staple food crops. Local famines frequently occurred. Those who refused to cooperate were beaten, imprisoned, or otherwise punished by the local authorities.

Colonial Portugal was more interested in aiding European settlers than in promoting African-based agriculture. It offered choice tracts of land, technical assistance, long-term, low-interest loans, and cash bonuses to those Europeans willing to settle the land. One of the largest settlement projects (*colonatos*) was the Limpopo Valley Scheme, where 250,000 acres (100,000 hectares) were designated for irrigation and occupation by 10,000 families. In the hinterland of Beira another major scheme, involving 75,000 acres, was laid out.

In 1965 the colonial government relaxed its investment policy and permitted multinational agricultural firms, such as the largely British-owned Sena Sugar Estates, Ltd., to expand their operations. By the late 1960s, almost half of Mozambique's cultivated land belonged to these estates, representing less than 1 percent of all farming units, while the other half was cultivated by subsistence farmers.

Mining. Before independence, mining and minerals contributed only a very small proportion of the gross domestic product. The coal mine at Moatize, near Tete, was the only significant mining operation, and its production was directed to the domestic railroads, cement plants, sugar estates, and local shipping lines. Total reserves are estimated at 700,000,000 tons, minimally. Other minerals mined in small quantities include tantalite (Mozambique reportedly possesses the world's largest deposit of this rare mineral) at Murrua in the north, gold at Bragança, and manganese and asbestos in Catuane in the southwest corner of the country. Even smaller amounts of graphite, ilmenite, mica, and bismuth have been mined.

Coal

Natural gas was tapped near Moamba (northwest of Maputo) and near Pande (60 miles south of Beira). Just before independence, some potentially significant mineral discoveries were made. Near Namapa in Nampula Province, for example, an iron ore deposit was found with sufficient promise to attract Japanese capital for development.

Industry. Industrial activity during the colonial era was largely restricted to processing agricultural raw materials such as sugar, cashew nuts, and tobacco and was oriented either to the small local market (mainly European) or to distant Portugal and foreign markets. Manufacturing, long discouraged by Portugal and adversely affected by the close proximity of South Africa and Zimbabwe (Rhodesia), did not really begin in earnest until the mid-1960s. Most manufactures were consumer items and substituted for imports; they included textiles, clothing, beer, bicycles, soap, and pharmaceuticals. The largest capital-intensive projects operating before independence were the oil refinery, fertilizer factory, and chemical plants in greater Maputo and the steel works at Beira.

Migrant labour. Mozambique has supplied thousands of migrant workers to South Africa each year, beginning with the discovery of gold on the Witwatersrand in 1886. Migrant labour soon became one of Mozambique's major sources of revenue. The flow of labour was initially regulated by the Witwatersrand Native Labour Association, formed in 1901 by South African mine owners. Agreements were subsequently signed that provided between 100,000 and 150,000 contract workers each year. Apart from taxation on the earned income, withheld and paid directly to the Portuguese administration in Mozambique, South Africa paid the administration a fee for every worker who entered South Africa for contract work. Under these labour agreements, Mozambique received 60 percent of the miners' wages in gold at the official gold rate. It then resold the gold at much higher free market prices, thus making substantial profits that helped balance the foreign exchange position. In 1975, for example, some $150,000,000 was earned this way on a labour pool of 103,000 workers.

The right to recruit workers in Mozambique was linked by the Mozambique conventions (1909 and 1928) to the right of the Mozambique railroads to handle 47.5 percent of the southern Transvaal's transit traffic. Thus Maputo became the single most important outlet for the Witwatersrand, and through separate agreements Beira became the chief outlet for Rhodesia and Malaŵi. Transit trade often accounted for 30 percent of Mozambique's foreign exchange earnings.

Transit trade

Tourism. Tourism was another important source of foreign exchange for colonial Mozambique. Tourists, primarily from South Africa and Rhodesia, came to enjoy the "continental" atmosphere of Maputo and Beira, the beaches, and the Gorongosa National Park, famous for its lions. Tourism fell during the final years of Mozambique's struggle for independence and after the closure of the Mozambique–Rhodesia border in 1976.

Throughout the colonial period, Portugal and South Africa were Mozambique's principal trading partners.

During the 1960s Portugal bought, on average, about 35 percent of the territory's exports (cotton, sugar, vegetable oils, tea, and other agricultural products) and supplied more than a third of its imports (primarily machinery, metals, and textiles). South Africa's share of imports and exports just before independence averaged about 20 and 10 percent, respectively.

Administration of the economy. Mozambique experienced an acute economic crisis for the first three years of its independence. Several factors caused this crisis: the mass and rapid departure of skilled European workers with their capital, technology, administrative expertise, and consumer demands; the deliberate and widespread destruction of property, machinery, and equipment by the departing white settlers; the closure of the border with Rhodesia (1976) and the inadequacy of international compensatory assistance; the fall in foreign exchange earnings under the Mozambique Convention and a drastic cut in the number of Mozambican workers in the South African mines; a further drop in foreign exchange because of severe cuts in world sugar prices; a large drop in the production of agricultural export crops caused by the white exodus, droughts in the north, and floods in the south; a complete halt to the tourist industry since 1974; the inexperience of the new government in handling and reorganizing the economic system; and the high cost of rehabilitating war-torn areas in the north.

Marxist–Leninist program

When Mozambique's independent government came to power, it adopted a Marxist–Leninist program of development in which special emphasis was given to developing the country's industrial base as well as its agricultural sector. Agriculture is viewed as the base of development, industry as the driving force. Economic development is seen as inseparable from the building of a socialist society.

The first national plan for economic recovery and development stressed the need to establish a network of communal villages and state farms to increase agricultural production. The first stage emphasized increased food production and the need to restore the country's agricultural exports. The second phase stressed the growth of industrial production.

State ownership

Under the constitution, Mozambique's land and subsoil resources are the property of the state, which is to determine their development and use. Although private property is still recognized and guaranteed by the constitution, and foreign capital will be encouraged (especially in industry), the state-owned sector is considered paramount. Private property will eventually be replaced by collective, socialist ownership. By the end of 1978 the large plantations, settlement schemes, coal industry, oil refinery, some metal industries, rental properties, and banking and insurance had been nationalized, as were all private banks except one; their assets and liabilities were transferred to the Bank of Mozambique and a newly created state-owned People's Development Bank. Education and health services and the Cabora Bassa Hydroelectric Project also were nationalized.

Despite its commitment to the establishment of a Marxist state, Frelimo has adopted a pragmatic approach to solving its economic difficulties. Foreign-owned businesses are allowed to operate provided they contribute to the well-being of the country and do not exploit its workers. The country relies heavily upon foreign aid in the form of cash grants, pledges, loans, technical assistance, and supplies of food, medicines, and other goods.

Transportation. Transport is insufficiently developed, having been constructed more to serve the transit trade of South Africa and Zimbabwe than the country's own needs. Both roads and railroads focus on three ports: Maputo, Beira, and Nacala.

Maputo–Witwatersrand line

Railways. There are three main systems. The Maputo system serves the Witwatersrand of South Africa, Swaziland via a short line through Goba, and the Limpopo Valley as far as Chicualacuala (former Malvêrnia) on the Zimbabwe border. The Beira system was built to serve Zimbabwe, Malaŵi, and Zambia. It is V-shaped, the western line running to Zimbabwe and the northern lines serving Tete and Malaŵi. The Nacala system extends due west to Malaŵi, with a branch line to Lichinga. Three secondary rail systems focussing on Quelimane, Inhambane, and Xai-Xai serve small agricultural areas near the coast.

Roads. The major all-weather roads are between Maputo and Ressano Garcia on the Transvaal border; Beira and Umtali in Zimbabwe; and Chimoio and Tete and the borders of Malaŵi and Zambia. There is also a coastal road from Maputo through Inhambane to Beira, which has been extended to Quelimane, Lumbo, Nacala, Pemba, and Mocímboa da Praia in the far north. During the 1960s the road network was improved and expanded, mainly for military purposes, but vast areas of the country are still virtually without roads.

Seaports. Mozambique's major ports are Maputo, Beira, and Nacala. The deepwater port of Maputo, situated on the well-protected Delagoa Bay, has special coal- and ore-loading facilities. It can accommodate 20 ocean-going vessels at a time and handle more than 16,000,000 tons of goods each year. Four miles away, at Matola, are a petroleum refinery, lumber wharves, and a mechanical loader for iron ore from Swaziland.

Beira, located at the confluence of the Púngoè and Búzi rivers, 15 miles from the open sea, is less favoured. Its harbour is shallow, sand and silt accumulations necessitate constant dredging, and a sandbar constricts the mouth of the Púngoè River. Its capacity is about half that of Maputo. Nacala, the principal port in the north, has a sheltered, modern harbour, one of the finest in East Africa. Other smaller ports include Mocímboa da Praia, Pemba, Lumbo, Angoche, Quelimane, Chinde, and Inhambane.

Air transport. Domestic air transport is operated by the state-owned DETA (Direção e Exploração dos Transportes Aêreas), which also has service to Portugal, South Africa, and Tanzania. There are international airports (Maputo, Beira, and Nampula), smaller domestic airports, and numerous landing strips in isolated regions.

Cabora Bassa dam

Power. Mozambique possesses a large hydroelectric potential and has many small and medium-sized power stations, mainly thermal, burning coal, wood, diesel oil, and gasoline. Several large dams are in operation, the largest of which is Cabora Bassa, located on the Zambezi River about 80 miles upstream from Tete. Begun in 1969 and built and financed by an international consortium of companies including South African, Cabora Bassa became operational (sending some electricity to South Africa) in 1976–77. During the war of independence, Frelimo forces repeatedly tried to destroy the dam.

Cabora Bassa's principal markets lie in South Africa, but there are plans to use its energy to develop mineral resources in Tete, Zambézia, and Nampula provinces; to irrigate parts of the Zambezi Valley; and to promote forestry, the livestock industry, and general manufacturing. River transport is now possible between the dam site and Tete. Backed up behind the dam, and extending 150 miles (240 kilometres) west, is a 600-square-mile lake that is being stocked with fish. The new lake required the resettlement of about 25,000 people.

ADMINISTRATIVE AND SOCIAL CONDITIONS

At independence, Mozambique's constitution (1975) created a single-party state in which all power was vested in the leadership of Frelimo. It declared the president of Frelimo to be president of the state, and it provided for the creation of the People's Assembly (later Popular Assembly) as the supreme legislative organ of the state. Samora Moïsés Machel, leader of Frelimo, became the first president.

Political Committee

At its third national congress, held in February 1977, Frelimo was officially converted into a party of the Marxist–Leninist ideological elite, which is committed to the destruction of capitalism and racism and to the creation of a truly revolutionary socialist (*i.e.*, Communist) state. The most powerful body is the 10-member Political Committee, which is composed of the President and his Cabinet (Council of Ministers). Frelimo insists on a very strong degree of centralization. Open discussion and criticism are encouraged at the lower levels of government provided they do not question the party line. At all levels, decisions rest with a very small inner group. Outside the organ of Frelimo, no political discussion is tolerated. Most members of the first Cabinet and the Political Committee were southerners and non-black Africans, a situation resented by the northerners, who provided the vanguard of the independence struggle.

For administrative purposes, Mozambique is divided into 10 provinces, about 94 districts, many town and city districts, and localities, or *localidades*. Each level has assemblies, but election to membership is not by direct universal suffrage. Above the *localidade* level, elections are indirect, each tier of assemblies electing those in the tier above. Mozambique's first general elections were completed in December 1977.

Police

Justice. Local law and order is in the hands of the Mozambique Police Corps and the National Service of Popular Security, which is responsible for the detection and combat of subversion and sabotage. The criminal code inherited from the Portuguese has been modified; the justice system suffers from lack of facilities and experienced administrators. Vigilance groups have been formed to counter reactionary activities and to fight alcoholism. Rehabilitation camps instead of prisons deal with minor criminals, prostitutes, and alcoholics.

Health and welfare. Health services collapsed with the flight of medical personnel at the end of colonial rule. In 1975 Frelimo nationalized all private clinics and prohibited private medical practice. It also introduced a program to train paramedics and began a vaccination campaign against smallpox and measles. The campaign began in the northern provinces of Cabo Delgado, Niassa, and Tete. Fragmentary data suggest that malaria, tuberculosis, bacterial and amoebic dysentery and other gastroenteric infections, plus pneumonia, measles, and infectious hepatitis are the most common diseases.

Rural health clinics are being expanded, and most medical care is free; preventive medicine is emphasized. The Maputo Central Hospital is the country's largest.

Education. Education was also nationalized in 1975, and in the next two years primary school enrollment rose. Primary education is now obligatory, and the government is attempting to encourage all young girls to attend school. Adult literacy classes are also available. Education is closely linked to the requirements of economic development and emphasizes reading, vocational training, and political education. Collective work in the fields is an integral part of the education process.

Housing. Under Frelimo's socialist–collectivist directives, the state largely controls the development and construction of housing. Large-scale urban renewal is underway in the major cities, in an attempt to house the thousands of Mozambicans who lived in the crowded shantytowns and slums. Rental properties belong to the state, which sets rental prices according to family income, family size, and type of residence. In the fully developed communal villages, there are nurseries, schools, health stations, and consumer cooperatives.

CULTURAL LIFE

During the colonial era there was no attempt to systematically introduce the Africans to Portuguese language and culture. In contemporary Mozambique, the state is trying to preserve and promote those indigenous values and customs that strengthen the socialist cause. Cultural centres have been established in urban neighbourhoods and communal villages throughout the country by the National Institute of Culture, which aims to collect indigenous literature, music, crafts, and mythology. During the revolutionary years, much of the poetry and literature had an immediate emotional and ideological purpose.

Cultural centres

For statistical data on the land and people of Mozambique, see the *Britannica World Data* section in the BRITANNICA WORLD DATA ANNUAL. (A.C.G.B.)

History

EARLY EXPLORATION AND TRADE

Between the 11th and the 15th centuries the Mozambique coast as far south as Sofala was linked politically and commercially to the coastal city-states north of Cape Delgado. Arab and Swahili traders penetrated inland from Sofala and up the Zambezi into what later became Manica e Sofala District. This was the situation the Portuguese found when Vasco da Gama put in at Mozambique Island early in March 1498. Just before this he had discovered and landed at what he called the Rio do Cobre (the Inharrime) and the Rio dos Bons Sinais (the Quelimane, the northern outlet of the Zambezi). Da Gama left Mozambique for India at the end of March, accompanied by a pilot lent by the sultan. In July 1500 Pedro Álvares Cabral, on his way to India, saw Sofala, and on his return it was visited by a squadron of his ships under Sancho de Toar, who made inquiries about its legendary gold trade. Between 1500 and 1502 either Cabral or João da Nova visited Lourenço Marques Bay. A Portuguese settlement was made at Sofala to exploit the fabled gold trade from the empire of Mwene Matapa, which was centred in what is now Zimbabwe but extended into Mozambique. The Portuguese in 1507 occupied Mozambique Island to serve as a way station to India.

Portuguese settlement at Sofala

Attempts to reach the Mwene Matapa were disappointing, though a convict named António Fernandes explored much of his territory before 1514. The Portuguese soon realized that they could not control the trade of the interior from Sofala alone, as Arab and Swahili merchants were making increasing use of the Zambezi route. Accordingly, after 1531 the Portuguese established settlements on the Zambezi banks at Sena and Tete, respectively 160 miles and 260 miles upstream. In 1544 a post was founded at Quelimane to guard the river delta. Simultaneously, the Portuguese began to trade at Inhambane and Lourenço Marques Bay, though no permanent settlement was set up at either place for some time. Portuguese efforts at Mozambique centred on the gold of the Mwene Matapa; it was only there that they involved themselves seriously in African affairs. In 1561 the Jesuit Gonçalo da Silveira reached the head village of the Mwene Matapa, whom he baptized. The Arabs and Swahili, seeing their commercial position threatened by the Portuguese, convinced the chief that Silveira's intentions were evil, and he had the Jesuit strangled.

The death of Silveira was followed by the dispatch (1569) of an expedition from Lisbon to seize the Mwene Matapa's mines. The expedition ended unsuccessfully in 1575, but despite this setback the Zambezi Valley and the Mwene Matapa's country remained the focus of Portuguese attention until the late 19th century. Elsewhere, the Portuguese limited themselves, or were limited by African opposition, to scattered settlements along the coast. Nevertheless, the arrival of the Portuguese in East Africa had a vital impact on African trade to the north of the Zambezi.

Taking advantage of factional strife the Portuguese in 1629 negotiated a vassalage treaty with Mwene Matapa Mavura, who thereby recognized Portuguese sovereignty over his kingdom. Portuguese ascendancy was short-lived, however, and at the end of the 17th century the expansion of the Rozvi kingdom of Guruhuswa, ruled by Changamire, eclipsed the power of both the Mwene Matapa and the Portuguese to the south of the Zambezi.

In 1752 Mozambique, after being subordinated to the Portuguese administration at Goa in India, was separated and given an independent governor. In the 18th century the slave trade became an important factor in Mozambique for the first time. The demand came first from French slavers from the Mascarene Islands and later from the Brazilians. Slave trading by Europeans continued into the 1850s, and an active slave trade to Madagascar was carried on by Arabs and Swahili for several decades more.

COLONIZATION

Until the mid-19th century the most serious attempt to colonize Mozambique was through the *prazo* system. This system of crown grant estates, which had its roots in the late 16th century, was formalized by the Portuguese government in the middle of the 17th. It failed as a colonization scheme, because the estate holders (*prazeros*) amassed large personal fiefdoms. In the 19th century a series of wars was waged by the government against these virtually autonomous rulers. A similar though less ambitious attempt was made to colonize the Kerimba Islands, which extend from Cape Delgado almost to Porto Amélia. There, too, the estate holders became independent lords.

No further colonization plans were contemplated until the 20th century. Several pilot projects failed in the 1930s and '40s, but since the 1950s an attempt to settle immigrant Portuguese peasant cultivators, either in cooperatives or on small individual family holdings, was pursued in the south. The most ambitious of these was centred in the lower Limpopo Valley, another was proposed for the upper reaches of the Revué River. The Mozambique government, unlike its counterpart in Angola, kept to a minimum the influx of poor unskilled white immigrants.

Portugal and other European powers in Africa. At the beginning of the 17th century the Portuguese beat off three attempts by the Dutch to seize the port of Mozambique. The Austrians seized Lourenço Marques in 1777, but Portuguese sovereignty was reestablished in 1782. During the Napoleonic Wars the French attacked Lourenço Marques, Inhambane, Mozambique, and the Kerimba Islands. During the same period these islands suffered more seriously at the hands of Sakalava raiders from Madagascar.

In the 19th-century partition of Africa among the European powers, Portugal maintained rights based on the discovery of most of the African coast, in the face of the doctrine of effective occupation. Until the Berlin Conference (1884) Portugal hoped to preserve a continuous belt of territory from coast to coast, uniting modern Angola with Mozambique. The acquisition of British rights in the area frustrated this design. The Portuguese proposed to submit the question to arbitration, but Britain refused and delivered an ultimatum to Lisbon (1890). A treaty in 1891 defined British and Portuguese possessions.

Colonial rule. The establishment of effective colonial rule in the interior began in the 1890s and was not completed in the northwest until 1912. The Makombe Rebellion (1917), which embraced much of the country south of the Zambezi, was the last traditional manifestation of African opposition to Portuguese rule.

The Companhia de Moçambique was incorporated by royal charter in 1891 for a term of 50 years. Much of its capital was foreign, and it exercised sovereign rights over the territories of Manica and Sofala until they were entrusted to direct Portuguese administration in 1942. The Companhia do Niassa acquired a charter for the administration of lands north of the Lurio River in 1894. Development was comparatively slow, and in 1929 the Portuguese government took over administration of the area.

Under the Treaty of Versailles (1919) the German East African territory between the mouth of the Ruvuma River and Cape Delgado was added to Mozambique.

Economic expansion

Railway building led to a rapid expansion of Mozambique's economy, and conventions with the countries of the hinterland were signed covering the question of transit trade and the recruitment of Africans from Mozambique to work in the South African mines. In 1901 a modus vivendi governing these matters was concluded with the Transvaal government. This was replaced by the Transvaal-Mozambique Convention in 1909, which lapsed in 1923 and was replaced in 1934 by a new treaty, which ran until 1939; it was then prolonged for five years and later remained in force subject to 12 months' notice of termination by either party.

A measure of autonomy was granted to the colony under the Portuguese republic, and this was extended in 1920. After the Portuguese revolution of 1926 this system was brought to an end, and within the provisions of the Colonial Act (1930) a much more centralized system was adopted. Mozambique was designated an overseas province of Portugal in 1951. (A.A.G.P./E.A.A.)

NATIONALISM AND INDEPENDENCE

Nationalist organizations became active in the 1950s, for the most part among the educated urban elite. In 1960 some 500 Makonde peasants were shot during a peaceful demonstration against forced cotton growing, and several thousand fled into Tanzania; there they were joined by a group of urban exiles committed to national revolution and led by Eduardo Mondlane. In 1962 they formed the Frente de Libertação de Moçambique (Frelimo) as an underground political and military movement. Within three years they claimed to have an 8,000-strong army and to have established their own administrative, educational, and marketing networks in Cabo Delgado and Niassa districts. Despite the assassination of Mondlane in 1969, a new phase of the war opened in 1971 in Tete province, under the leadership of Samora Machel. By 1974 Frelimo had won victories in central Mozambique.

Portugal's initial response to the outbreak of revolt was all-out war; by the mid-1960s there were some 70,000 Portuguese troops in each territory. Large numbers of black troops were recruited, and villagers supporting the guerrillas were subjected to savage reprisals, as at Wiriyamu in Mozambique in December 1972, where hundreds of people were burned to death. In 1963, in a bid to attract international support, Portugal opened the colonies to foreign investment, and the building of the Cabora Bassa Dam was partly designed with the same object. By the late 1960s the regime also turned to economic and educational reform, in part to meet the rising demands for a semiskilled work force, in part to preempt the nationalists. But the reforms were too few and too late, and in April 1974 the sheer cost of the wars—together with rising dissatisfaction with the government in Portugal—led to a revolt of Portuguese army officers.

Independence

The collapse of the regime in Portugal led directly to Mozambique's independence in June 1975 under a Frelimo government headed by Machel. The flight of skilled expatriates and Mozambique's geopolitical relationship to South Africa and Rhodesia were the immediate problems confronting the new government. Machel's Frelimo movement was restructured as a self-proclaimed Marxist-Leninist party and set about revising the constitution, providing for a system of elections, reforming the military, and reviving the shattered economy. At the same time, Mozambique was constantly raided by Rhodesian and South African air and land commandos in reprisal for obeying UN sanctions against Rhodesia and for providing a guerrilla base for the Zimbabwe African National Union (ZANU). (Sh.M./Ed.)

Mozambique was a founding member of the Southern African Development Coordination Conference, and Machel made diplomatic visits to several Western nations as well as to Mozambique's traditional allies, the Soviet Union and East Germany. South Africa continued to finance and support the dissident Mozambique National Resistance Movement (known as Renamo, or MNR), and Mozambique continued to harbour supporters of the African National Congress. In 1984 the two nations signed the Nkomati Accord, which promised mutual nonaggression, but each side later accused the other of violating the treaty.

For later developments in the history of Mozambique, see the *Britannica Book of the Year* section in the BRITANNICA WORLD DATA ANNUAL. (Ed.)

NAMIBIA

The Republic of Namibia, formerly known as South West Africa, lies on the Atlantic coast of Africa between latitudes 17° and 29° south. In the north it is bounded by the Kunene and Okavango rivers, which separate it from Angola. Here a narrow strip of land—the Caprivi Strip (Caprivi Zipfel)—extends eastward for nearly 300 miles (485 kilometres), bringing the territory to the borders of Zambia. In the south the Orange River marks the boundary with South Africa. In the east Namibia adjoins both Botswana and South Africa. Extending inland some 600 miles in the north, without the Caprivi Strip, and about 220 miles in the south, the total area of Namibia, including the Walvis Bay enclave (administered by South Africa), is about 318,000 square miles (824,000 square kilometres).

Physical and human geography

THE LAND

Relief. Namibia may be described as a highland between two desert areas. Three main topographical regions may be distinguished—the Namib coastal belt, the Central Highland of the Central Plateau, and the northern and eastern parts of the plateau known as the Kalahari Belt. The Namib Desert extends from the Atlantic coast in the west until, at an altitude of 3,000 to 3,500 feet (900 to 1,100 metres), it meets the foot of the plateau edge, or Great Escarpment, in the east. A desolate strip of sand desert, high shifting dunes, salt pans, and bare rock outcrops, the desert is bounded by the Kunene River in the north and by the Orange River in the south; for much of its length it is less than 100 miles wide. To the north it becomes more mountainous, with spectacular granite masses rising abruptly from the plain. The Brandberg, with an altitude of 8,461 feet (2,579 metres), is the highest point in Namibia. The Namib constitutes about one-seventh of the total area of the territory.

East of the Great Escarpment lies the Central Plateau, which has two major surface divisions—the plateau hardveld (or rocky plain), known as the Central Highland, and the Kalahari sandveld (desert). The Central Highland extends from the southern to the northern frontier and varies in altitude from 3,200 to 6,500 feet. Its diverse surface characteristics include broken veld, rugged mountains, sand-filled valleys, gentle plains, and isolated peaks. A mountainous central part in the vicinity of Windhoek divides the Central Highland naturally into Damaraland to the north and Great Namaland to the south. Comprising such ranges as the Khomas Highland, the Auasberge (where the Moltkeblick at 8,143 feet is the second highest peak), and the Eros and Onyati to the east, it forms the territory's highest mountain tract, constituting a central watershed from which the various headwaters drain north, south, east, and west.

The Kalahari sandveld extends along the whole of the eastern frontier of Namibia northward to the Okavango River. Its thick layers of terrestrial sands and limestone vary in depth from more than 1,000 feet to a thin covering at the hardveld edge. The width of the sandveld varies from 50 miles in the south, where it covers most of the Nossob-Auob Basin, to 250 miles in the north, where it comprises part of the Etosha Basin and the middle Okavango Basin. It has a level, sandy surface, which may rise to as much as 300 feet. In the south the sandveld is crossed by long sand dunes and grass and is covered with thorn trees.

The scarcity of water

The water shortage, or the uncertainty of obtaining water, touches all aspects of life and of development. Of the larger rivers only the Orange, Kunene, and Okavango are perennial, and they constitute boundaries of the territory. All rise outside Namibia and—as with the Okavango—the Kwando and the Zambezi in the north flow only through small sections of the territory. The potential of these rivers for irrigation, however vital, is thus limited to the southern and northern margins. The remaining rivers, whether the large westward-flowing deep-channel streams such as the Swakop and the Kuiseb or the eastward-flowing shallow watercourses (*omuramba*s) such as the Epukiro and the Omuramba Omatako, usually run only intermittently. They rarely contain surface water, except after rains or floods, although they constitute storage places for groundwater. Surface water collects mainly in shallow clay or limestone pans, especially in the sandveld region—the most famous of which is the giant Etosha Pan measuring some 68 by 31 miles—as well as in small lakes. There are also a number of springs. The general scarcity of water thus makes the well and the borehole the basis for land utilization. The conservation of floodwater through the construction of dams is, however, also of importance. It is estimated that there are some 10,000 dams, with a total storage capacity of 875,000,000 cubic yards (670,000,000 cubic metres). The largest of these is the Hardap Dam on the southward-flowing Fish River, located northwest of Mariental.

Climate. The climate is generally hot and dry. The territory is partly situated in the tropical high pressure belt of the Southern Hemisphere, located between latitudes 25° and 35° south, which governs the wind system of the plateau. The coastal region, however, experiences the air circulation that occurs as a result of the interaction between the eastern sector of the South Atlantic High pressure system, and the upwelling of the cold Benguela Current offshore. The mean temperature is, as a result, lower than the latitude might suggest. The temperature varies from 66° to 72° F (19° to 22° C) in the Central Highland to about 60° F (16° C) on the coast. During the summer months it may reach as high as 100° F (38° C) in the north and 120° F (49° C) in the Namib and the south. In July, which is usually the coldest month of the year, frost may occur, although only lightly in the north. In the south, however, temperatures have been known to drop to 20° F (−7° C) or lower.

Variations in temperature

Rainfall in the territory increases from south to north. In the arid Namib and the Orange River Valley, the mean annual rainfall is less than four inches (100 millimetres); in the north, and along the Caprivi Strip, the rainfall increases to between 24 and 26 inches. The mean annual rainfall for the territory as a whole is about 10 inches. Most rainfall occurs in the summer months (December to February). Only the extreme southwest receives any rain (up to one-half inch) in winter. The territory is occasionally subject not only to periods of flood, when the precipitation may be more than double the normal expectation, but also to periods of drought. The rains are the most reliable in the wet northeast, and the most variable in the dry south.

Plant and animal life. There are two categories of vegetation. In the west is found the desert and semidesert type, with its desert grasses and shrub; in the northeast the woodland savanna type, with its grasses, bushes, and trees. In the Namib Desert region there is little vegetation except for succulents and the narra, the nut of a gourdlike plant. Farther inland, and in the south, there are varieties of grasses, of succulents, such as the large aloe, of euphorbias, and of dwarf trees such as those belonging to the *Acacia* and *Boscia* genera. From south to north the veld changes to grass country, increasingly studded with thorn trees, of which most are acacia. Many others are also found, however, such as raffirboom, witgat, bush willow, and wild olive. Farther north the annual grasses become more abundant, the "bush" vegetation and acacia scrub grows thicker, and the variety of thorn trees increases. This pattern is not, however, uniform. The areas to the east and west of the palm-tree belt in southern Ovamboland are, for example, sparsely wooded. In the north the growing density of the trees, the appearance of the palm-tree belt, and, to the west, the baobab tree, all are signs marking the proximity of the well-wooded northern forest region, which stretches from the Kaokoveld across Ovamboland in the west to the greater part of the Okavango and Caprivi Strip areas in the east. It is in this region that most of the indigenous timber trees of the territory are found—including Rhodesian teak and mahogany, mopane, marula, manketti, and kiaat among others. In the west, however, as one approaches the dry steppe of the Kaokoveld, the trees and the bushes thin out; in this region the grass spreads more abundantly, and the mopane trees and bushes diminish in size.

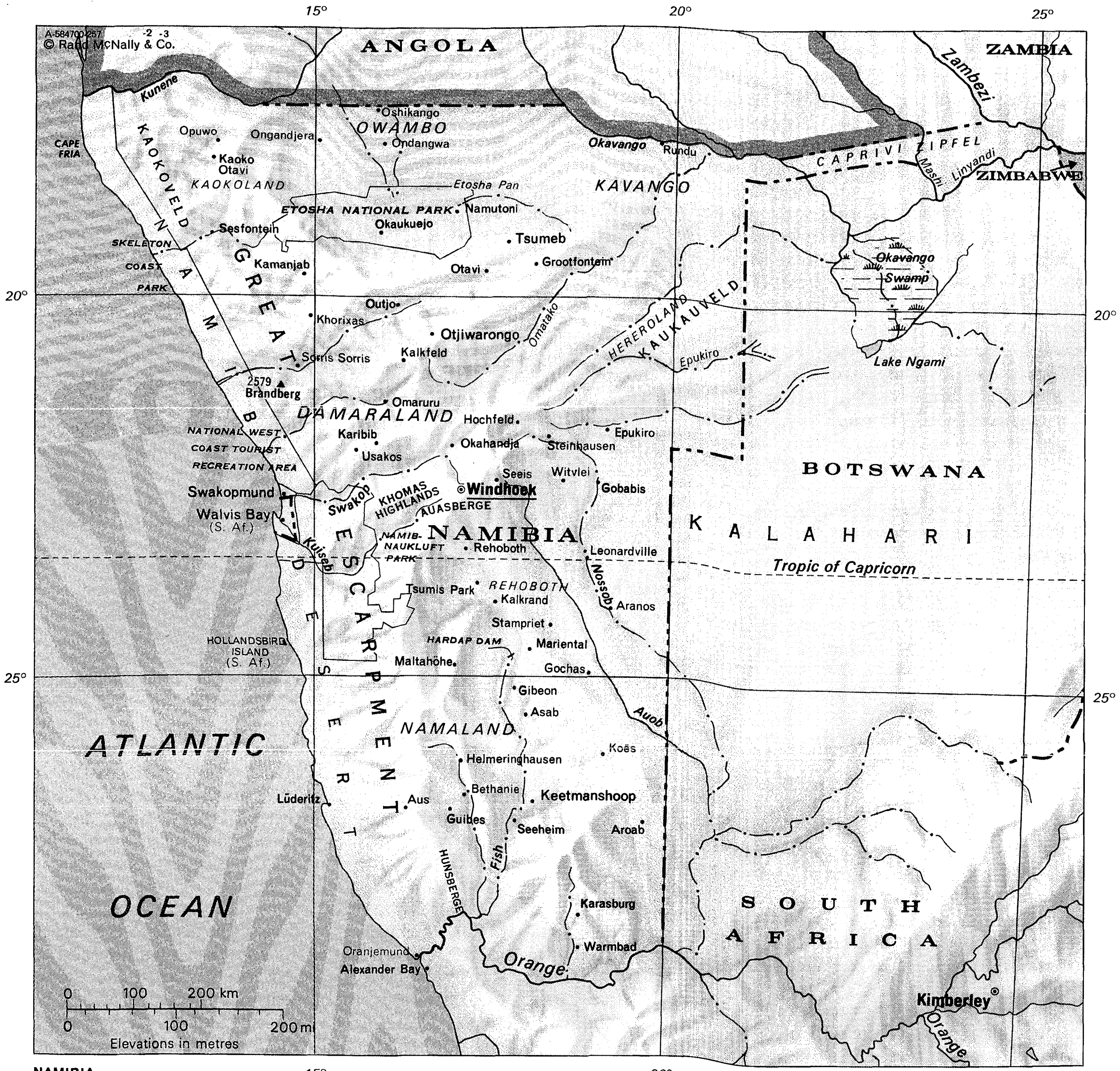
The northern forest region

Big game is relatively abundant and is found in many parts of the territory. The wildlife includes lion, elephant, rhinoceros, giraffe, leopard, buffalo, eland, oryx, hartebeest, kudu, and zebra, as well as many smaller species of game and of predatory animals. The most renowned of the game reserves is the Etosha National Park, which covers an area of 8,500 square miles and is one of the largest game parks in the world.

Settlement patterns. The ecological relationship between man and environment is a complex one. Among the elements affecting it are frequent drought, the low protein and low phosphorous content of grasses, and the high salt content and humus deficiency of the soil. The best grazing and the most groundwater are found in the plateau hardveld region, most of which is now available to all inhabitants of the region since the repeal of apartheid laws in the late 1970s. Despite this, however, the white-owned farms have continued to predominate.

The many factors affecting the land settlement include the large, uninhabitable desert regions, the German wars against the Herero and Nama, the encouragement of white settlement by South Africa, the setting aside of the Northern Sector for the blacks living there, and the allocation of additional Reserves in the Southern Sector.

Windhoek is the administrative centre and principal town of Namibia. Walvis Bay, administered by South Africa, is Namibia's chief port and only natural deep harbour. The other port, Lüderitz, is shallow and used mainly by small



NAMIBIA

MAP INDEX

Cities and towns

Aranos	24.09s	19.09e
Aroab	26.47s	19.40e
Asab	25.29s	17.59e
Aus	26.40s	16.15e
Bethanie	26.32s	17.11e
Epukiro	21.41s	19.08e
Gibeon	25.09s	17.43e
Gobabis	22.30s	18.58e
Gochas	24.55s	18.55e
Grootfontein	19.32s	18.05e
Guibes	26.41s	16.42e
Helmeringhausen	25.54s	16.57e
Hochfeld	21.28s	17.58e
Kalkfeld	20.53s	16.11e
Kalkrand	24.03s	17.33e
Kamanjab	19.35s	14.51e
Kaoko Otavi	18.14s	13.40e
Karasburg	28.00s	18.43e
Karibib	21.58s	15.51e
Keetmanshoop	26.36s	18.08e
Khorixas	20.22s	14.58e
Koës	25.59s	19.08e
Leonardville	23.29s	18.49e
Lüderitz	26.38s	15.10e
Maltahöhe	24.50s	17.00e
Mariental	24.36s	17.59e
Namutoni	18.49s	16.55e
Okahandja	21.59s	16.58e
Okaukuejo	19.10s	15.54e
Omaruru	21.28s	15.56e
Ondangwa	17.55s	16.00e
Ongandjera	17.54s	15.05e
Opuwo	18.03s	13.45e
Oranjemund	28.38s	16.24e
Oshikango	17.24s	15.53e
Otavi	19.39s	17.20e
Otjiwarongo	20.29s	16.36e
Outjo	20.08s	16.08e
Rehoboth	23.18s	17.03e
Rundu	17.52s	19.43e
Seeheim	26.50s	17.45e
Seeis	22.29s	17.39e
Sesfontein	19.07s	13.39e
Sorris Sorris	20.57s	14.50e
Stampriet	24.20s	18.28e
Steinhausen	21.49s	18.20e
Swakopmund	22.41s	14.34e
Tsumeb	19.13s	17.42e
Tsumis Park	23.43s	17.28e
Usakos	22.01s	15.32e
Walvis Bay (S. Af.)	22.59s	14.31e
Warmbad	28.29s	18.41e
Windhoek	22.34s	17.06e
Witvlei	22.23s	18.32e

Physical features and points of interest

Atlantic Ocean	25.00s	13.00e
Auasberge, *mountains*	22.45s	17.22e
Auob, *river*	25.43s	20.00e
Brandberg, *mountain*	21.10s	14.33e
Caprivi Zipfel, *physical region*	17.59s	23.00e
Damaraland, *historic region*	20.30s	14.30e
Diamond Area 1	27.15s	15.45e
Diamond Area 2	25.00s	15.15e
Epukiro, *watercourse*	21.28s	19.59e
Etosha National Park	19.00s	16.00e
Etosha Pan, *salt flat*	18.45s	16.15e
Fish, *river*	28.07s	17.45e
Fria, *cape*	18.30s	12.01e
Great Escarpment	23.00s	15.00e
Hardap Dam	24.28s	17.48e
Hereroland, *historic region*	20.45s	19.30e
Hollandsbird Island	24.45s	14.34e
Hunsberge, *mountains*	27.45s	17.12e
Kaokoland, *historic region*	18.15s	13.00e
Kaokoveld, *plateau*	19.00s	13.00e
Kaukauveld, *plateau*	20.00s	20.30e
Kavango, *historic region*	18.30s	19.30e
Khomas Highlands, *mountains*	22.30s	16.30e
Kuiseb, *river*	22.59s	14.31e
Kunene, *river*	17.20s	11.50e
Linyandi, *river*	17.50s	25.05e
Mashi, *river*	18.27s	23.32e
Namaland, *historic region*	25.45s	18.00e
Namib Desert	23.00s	15.00e
Namib-Naukluft Park	23.30s	15.30e
National West Coast Tourist Recreation Area	22.00s	14.15e
Nossob, *river*	24.52s	20.00e
Okavango, *river*	18.17s	21.45e
Omatako, *watercourse*	17.57s	20.25e
Orange, *river*	28.41s	16.28e
Owambo, *historic region*	18.00s	16.00e
Rehoboth, *historic region*	23.30s	17.00e
Skeleton Coast Park	19.30s	13.00e
Swakop, *river*	22.38s	14.36e
Zambezi, *river*	18.55s	36.04e

coastal craft or lighters. Tsumeb, the principal copper and lead-mining centre, is some 250 miles north of Windhoek.

THE PEOPLE

Linguistic groups

Of the white population approximately two-thirds are Afrikaans speaking, one-quarter German speaking, and one-tenth English speaking. The nonwhite population may be divided into two main types, the Southwestern Bantu and the Khoisan types. The San (Bushman) and Nama peoples belong to the Khoisan peoples, and their language contains the characteristic Khoisan clicks. The Bergdama (Damara), a black African people of unknown origin, have adopted the Khoisan language. The Hereros, who speak their own language, belong to the Bantu language family, as do the Ovambo, the Okavango, and the Kaokovelders. The latter are closely related to the Hereros in origin, language, and culture. The East Caprivians also belong to the Bantu-speaking peoples but are related to the Lozi and Makololo groups of Barotseland (western Zambia), rather than to the Bantu of Namibia. The Tswana and other groups include a few thousand people originally from Angola. They are mostly Bantu-speaking but are not related to the other ethnic groups in the territory. Apart from the established group of Tswana, who are related to the Tswana of Botswana, most of these peoples are dispersed throughout the rural and urban areas of the territory. The Basters of Rehoboth (or Rehobothers) are of mixed origin, being descended in part from unions between Afrikaner trekkers and Nama women. They moved from the Cape Colony in the 1860s and settled at Rehoboth. Their language, like that of the Coloured people, is predominantly Afrikaans, and their way of life is Western. The Coloured people, more recent immigrants, were originally a part of the Cape Coloured population of South Africa.

THE ECONOMY

Namibia's economy is determined, in part, by the harsh and arid environment. Its three key industries—mining, farming, and fishing—together account for a large portion of the gross domestic product (GDP). Its internal market is small; traditional subsistence farming persists over large areas, and its economic dependence upon South Africa is virtually complete. Cattle and sheep (mostly Karakul) account for most of the total gross output of commercial agriculture. Namibia obtains its fruits and vegetables, as well as the bulk of its grains for the Southern Sector, from South Africa. In bad seasons, the Northern Sector, too, must rely on South Africa for these commodities. Mining of diamonds and base metals contributes more than one-fourth of the GDP. Manufacturing is severely restricted by high costs and by a low aggregate demand from a small, poor, and widely scattered population. Mineral products account for more than four-fifths of the total value of exports, and fishery products for most of the remainder.

The transportation network primarily serves the Southern Sector. Transport development has been aided by capital and loan funds from the South African government. Tarred roads link Windhoek to Tsumeb in the north, to Keetmanshoop in the south, and to Swakopmund and Walvis Bay in the west—all of which are in the Southern Sector. A road construction program for the Northern Sector is still in a rudimentary phase.

The railroad system

Namibia's railway system was amalgamated with that of the South African Railways and Harbours Administration, and the track and sidings are all in the Southern Sector. The railroad runs from South Africa to Windhoek and north to Tsumeb. This north–south central spine is linked, via east–west lines, with the two ports of Walvis Bay and Lüderitz. Another line links Windhoek with Gobabis.

The high cost of road building adds importance to air services. Windhoek has an international airport, and there are a number of other airfields and landing strips. There are flights from Cape Town and Johannesburg to Windhoek. South West Airways, a subsidized private airline, maintains internal services. (R.B.Ba.)

ADMINISTRATIVE AND SOCIAL CONDITIONS

Administration. Prior to its independence in 1990, Namibia was administered by the South African government, which gained possession of the territory in 1915. Namibia's constitution, which took effect on March 21, 1990, establishes a multiparty system, outlaws the practice of apartheid that had been imposed by the South African government, explicitly protects a large number of civil rights, including freedom of speech, and abolishes the death penalty. It provides for a president, who is the head of state and who, along with the cabinet, has broad executive powers; a bicameral Parliament; and an independent judiciary, which rules on the constitutionality of legislation.

The president is directly elected and limited to two five-year terms. The cabinet, headed by the president, directs the ministries of government and is composed of the prime minister and other members of the National Assembly appointed by the president. The president has the power to declare martial law and to dissolve the National Assembly.

The National Assembly, the lower house of Parliament, is composed of 72 directly elected members and no more than six nonvoting members appointed by the president because of special expertise. Legislation may be passed by a simple majority, but the constitution requires a two-thirds majority for legislation to become law without both the assent of the president and the confirmation of the National Council, the upper house of Parliament. The National Council is also responsible for recommending legislation concerning regional matters; its members are elected by the regional councils, discussed below, each of which choose two from among their members. A two-thirds majority in both the National Assembly and the National Council is required to impeach the president.

Appointments to the Supreme Court and to the High Court are made by the president. Decisions by the Supreme Court, which rules on appeals brought from the High Court, are binding on all Namibian courts. The High Court hears civil disputes and criminal prosecutions, as well as appeals from the Lower Courts, which are established and regulated by the National Assembly.

The boundaries for local and regional governments are delimited by a national commission. The constitution requires that each regional and local government have a freely elected council as its principal governing body.

Education. About one-quarter of adults in Namibia are illiterate. The 1990 constitution, however, requires all children under 16 who have not completed their primary education to attend school. There is no university in the territory; Namibians who attend university usually do so in South Africa.

Health and Welfare. While Namibia has one of the highest per capita incomes in Africa, a great majority of the nonwhite population is extremely poor. Many Namibians live in overcrowded housing and have a water supply that is either inadequate or unclean. Health problems linked to poor sanitation, such as tuberculosis and gastrointestinal diseases, are common, as is malnutrition. Much of the population lacks even basic health care services.

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

History

As a geographically defined area, South West Africa came into being during the last two decades of the 19th century. It was the result of agreements and boundary settlements between Britain, Portugal, and Germany following Germany's decision in 1884 to acquire a colony in southwestern Africa. One important enclave, however, did not become part of the German acquisition. The largest port, Walvis Bay, and a coastal strip comprising in all an enclave of 434 square miles had already been annexed by Britain in 1878 and became a part of Cape Colony in 1884. At the time of the formation of the Union of South Africa in 1910, it became South African property.

Apart from this enclave, South West Africa remained under German rule until shortly after the outbreak of World War I, when it was successfully invaded by troops from South Africa. The colony was then placed under South African military administration until January 1, 1921, when it became a Class C mandated territory, administered by South Africa on behalf of the League of Nations.

The Class C mandates explicitly authorized the administering governments to exercise full powers over the territories as integral parts of their own states, subject to the international supervisory authority of the Council of the League of Nations. Upon the formal dissolution of the League in April 1946, South Africa, after unsuccessfully seeking UN approval for the incorporation of the territory, refused to place it under the UN trusteeship system, and a dispute between South Africa and the United Nations began on the status of South West Africa. The position of successive South African governments was essentially that, with the demise of the League and its supervisory machinery, all of their international obligations concerning the mandate, save one, had lapsed. The acknowledged exception was the clause providing that the territory should be administered as a "sacred trust of civilization." The initial position of the United Nations was that the mandate was still in force, that the supervisory authority of the League was to be exercised by the United Nations, and that, both by its refusal to sign a trusteeship agreement for South West Africa and by the conduct of its administration, South Africa had violated the mandate. (R.B.Ba.)

)ispute
etween
outh
ıfrica and
ıe UN

The nature of South African rule came before the United Nations annually, and in 1966 the UN terminated the mandate and called for complete South African withdrawal. This was upheld in an advisory opinion from the International Court of Justice at The Hague in 1971. In 1973 the United Nations appointed its own commissioner for Namibia (as the territory became known in the 1970s).

In 1958 a mass-based nationalist organization, the Ovamboland People's Organization, was formed to protest against South Africa's policies. In 1960 it became the South West Africa People's Organization (SWAPO), which came to represent the opposition of all sectors of the population to political incorporation by South Africa, the alienation of African land, the contract labour system, grossly inadequate education, the proposed introduction of black states (formerly called Bantu Homelands), and racial inequalities even more glaring than in the republic.

.rmed
:ruggle
nd the
'urnhalle
'onference

After years of unsuccessful peaceful tactics, SWAPO decided on armed struggle, and in 1966 a major conflict at Ongulumbashe led to the replacement of the police with the South African Defence Force. With the independence of Angola in 1975, the war in the northern part of the country intensified. In the light of this and pressure from the United Nations and the Western powers in the UN Security Council, South Africa initiated the Turnhalle Conference, held in 1976 at Windhoek. It was unrepresentative, and its constitutional proposals called for the division of the country into 11 semiautonomous ethnic governments. SWAPO refused to attend the conference and, along with the United Nations and the Organization of African Unity, denounced its proposals. In 1977 a South African administrator general was appointed to carry out a form of decolonization, and a constituent assembly was formed in 1978 in accordance with the Turnhalle plans; neither gained international recognition. (Sh.M./Ed.)

Protracted negotiations among South Africa, SWAPO, the United Nations, and several Western powers broke down in the early 1980s over the implementation of the UN Security Council's Resolution 435, which contained a settlement plan that had been conditionally accepted in 1978 by both Namibia and South Africa. The plan involved a UN-supervised election for a national assembly that would draft Namibia's constitution. With the support of the United States, South Africa made acceptance of the plan contingent upon Cuban troop withdrawal from Angola.

South Africa renewed direct control over Namibia in 1983. Under the auspices of the interim administration, several Namibian political parties formed the Multi-Party Conference (MPC), denounced by SWAPO as an instrument of South Africa's interests. The South African government accepted the MPC's proposals in 1985 for a "transitional government of national unity," which was established to the condemnation of SWAPO and the United Nations.

In 1988 Cuba for the first time negotiated with South Africa and Angola on its presence in the region. Mediated by the United States, the talks resulted in a timetable for the withdrawal of South African and Cuban troops from Angola, as well as for the implementation of UN Security Council Resolution 435, which began on April 1, 1989. In November, elections for the new National Assembly gave SWAPO 57 percent of the vote, far short of the two-thirds required to control the passage of the constitution. Therefore, SWAPO was forced to work with other parties in writing the constitution, passed in February 1990, which established a multiparty system and a directly elected executive president. The assembly, which gave itself the power to elect the first president, unanimously chose the veteran SWAPO leader Sam Nujoma, who took office on March 21, the day of Namibia's independence.

For later developments in the history of Namibia, see the *Britannica Book of the Year* section in the BRITANNICA WORLD DATA ANNUAL. (Ed.)

SWAZILAND

Swaziland is a landlocked, independent kingdom wedged between South Africa to the north, west, and south and Mozambique to the east. It has a total area of 6,704 square miles (17,364 square kilometres). Small and compact, the oval-shaped country is about 110 miles from north to south and about 85 miles from east to west. It is named for the Swazi nation, which comprises almost all of the country's people. The national capital is at Mbabane.

Formerly a British High Commission Territory, Swaziland gained its independence in 1968 under the leadership of its king, Sobhuza II. The small black nation is faced with delicate problems of economic development. It is rich in agricultural land and mineral resources, but its mines are controlled by foreign interests, and its commercial agricultural operations remain in the hands of Europeans. Economic progress has, however, been steady.

Physical and human geography

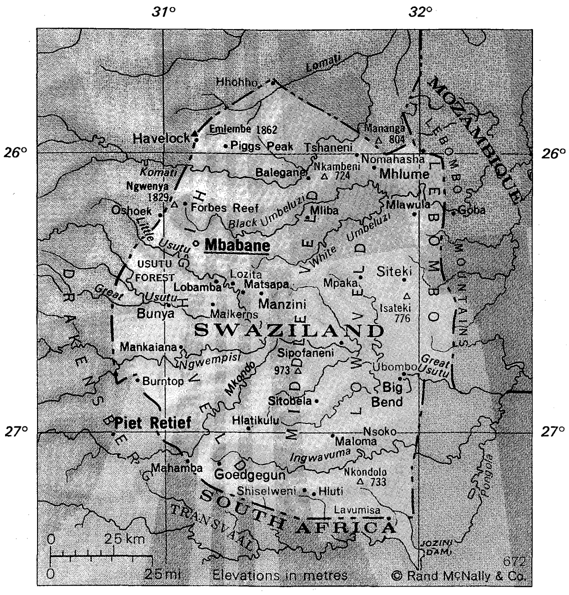
THE LAND

Relief and soils. Despite its small size, Swaziland has a rich variety of relief features and soils. There are four well-defined topographical regions, extending longitudinally from north to south in roughly parallel belts. From west to east they are Highveld, the Middle Veld, and Lowveld, and the Lebombo (Lubombo) escarpment.

The Highveld, which covers an area of 2,000 square miles, is a continuation of the Drakensberg range of South Africa. Locally known as Inkangala (meaning "a cold treeless place"), it is a granite massif (mountainous mass) with quartzite ridges that are broken up into rugged terrain. The average elevation is between 3,500 and 4,500 feet (1,100 and 1,400 metres); the highest elevations are the summits of Emlembe (6,100 feet) and Ngwenya (6,000 feet). The rocks on gentle gradients are covered by deep red, orange, and yellow acidic soils of medium texture.

The Middle Veld covers an area of 1,900 square miles and has an average altitude of 2,000 to 2,500 feet. It is a region of hilly country and well-watered valleys. Its foundation is mainly dolomite (limestone rich in magnesium carbonate) and gneiss (coarse-grained rocks containing ponds of minerals). Dolerite (coarse basalts) and quartzite also occur. The principal soils are deep and friable red and clay loams; gray-brown sands and sandy loams rest on mottled, sandy clay or iron concretions.

The Lowveld, or Bushveld (locally known as Lihlanze, a Swazi name meaning " a warm place with trees") covers 2,200 square miles. It is a gently undulating region with hills that rise from an average of 500 to 1,000 feet to a height of 2,300 feet reached by the isolated hill of Nkambeni and the isolated ridge of Nkondolo. The complex geology of the region ranges from the acidic rocks of the



SWAZILAND

west to the basalt and dolerite of the east. The soils are similar to those of the Middle Veld.

The Lebombo, covering 600 square miles, is an escarpment along the eastern fringe of the Lowveld. It rises from an average height of 2,000 feet to the two highest points of Isateki hill (2,500 feet) and Mt. Mananga, known as the Mananga Beacon (2,640 feet), which is on the border with South Africa. The scarp is intersected by the gorges of the Ingwavuma, Usutu, and Black Umbeluzi rivers. The plateau's foundations are composed of acidic and intermediate volcanic lavas, such as rhyolite, lava-formed granite. Soil is virtually absent in some areas, while in others it is fairly deep, reddish, and medium to heavy in texture.

The rivers

Drainage. The country is one of the best watered in southern Africa; major rivers flow through all four regions. From the Highveld the Lomati, Komati, Black Umbeluzi, Usushwane (Little Usutu), Usutu, Ngwempisi, and Mkhondvo rivers flow in a generally easterly direction toward the Indian Ocean. The Usutu, which carries the greatest volume of water, flows into the Pongola to form the Maputo River of Mozambique.

Climate. The climate varies from almost tropical and subhumid conditions in the Lowveld, where the average annual rainfall amounts to between 20 and 35 inches, to the humid, near temperate climate in the Highveld, which receives between 40 and 90 inches of rainfall. The Middle Veld and Lebombo region are subtropical and drier, receiving only 30 to 50 inches of rain. Rains occur during the summer from October to March. The mean annual temperature is 59° F (15° C) in the Highveld and 72° F (22° C) in the Lowveld.

Plant and animal life. More than 2,600 species of ferns and flowering plants have been recorded. Among these are 25 species of aloe (members of the lily family), arum lilies, Cape primroses, begonias, gladioli, and a rich variety of orchids. Cape heath (*Erica*), a large sturdy shrub, with red, pink, green, yellow, and white flowers, grows in the mountains. Trees include the crimson-flowered Hottentot's bean, the scarlet-flowered umsini, umvangata or kiaat (a leguminous tree that yields durable wood), and umbrella trees.

Although the animal population has been greatly reduced (earlier by hunting, but now by the expansion of agriculture), there are still numbers of hippopotamus, blue wildebeest, zebra, and antelope (including the kudu, impala, reedbuck, duiker, klipspringer, and waterbuck). Lions, baboons, monkeys, and galagos (small African lemurs known as bush babies) are less numerous. There are crocodiles in the Lowveld rivers. Birdlife is abundant and includes a few rare species with northern affinities, such as the European stork, the sacred ibis and hadedah. Other more conspicuous species are the hammerhead, the gray heron, the lisakabuli (widow bird), the hornbill, and the lilac-breasted roller.

Settlement patterns. Nine-tenths of all Swazis live in villages that are divided into family homesteads, or *umuti.* Each village is led by a headman, who is the village patriarch. A unique feature of Swazi life is the role of the headman's mother, who acts as village "mother" and is in control of everything affecting the homestead. Each such homestead, at the minimum, consists of a great hut, the cattle kraal (enclosure), and the bachelors' quarters. Flask-shaped pits for storing grain are built into the wooded palisades surrounding the kraal.

Rural life

The pattern of village life is strictly set by the seasons. With the onset of the rains in spring (August to October), women plant small gardens along the riverbanks; later when the heavy rains come in summer (November to January), with some help from the men, they hoe and sow corn (maize) and millet in larger fields; at this time all able women and children abandon their homesteads for the fields, and the men also join in the planting and weeding. The summer months are, on the whole, the hungry months.

Autumn (February to March) is the time of harvest; in winter (April to July) the last of the crops are brought in. Activity then moves to the homesteads, where women and men thresh the grain, the best of which is stored and the remainder consumed at once. Winter is a time for relaxation; the men hunt and entertain from their full food bins and the women visit their parents.

The central points of Swazi political life are the royal villages of the king at Lozita, and the queen mother at Lobamba. The two principal commercial centres are Mbabane and Manzini. Mbanane, attractively situated on wooded hills, contains both modern and colonial-style buildings. It is the country's largest town. Manzini is a growing industrial and agricultural service centre.

THE PEOPLE

The Swazi nation

The Swazi nation is composed of more than 70 clans, of which 70 percent are Nguni and 30 percent Sotho. The Dlamini, which is the dominant clan, forms the nucleus of the nation and constitutes the royal family. Its customs, laws of property and person, and its dialect, known as Siswati, prevail in all matters.

The Swazis compose about nine-tenths of the total population. Other Africans, mainly Shangane from Mozambique, Europeans, largely English- and Afrikaans-speaking South Africans, Eurafricans (persons of mixed European and African descent), and Asians make up the rest of the population. Eurafricans and Asians are mainly in commerce and industry as shopkeepers, clerks, and artisans. Most Europeans are either large-scale farmers, miners, technicians, teachers, public officials, or traders. The majority of the Europeans speak English, which is one of the official languages, together with Siswati.

A majority of Swazis are Christians. The remainder hold strongly to their traditional animist religious beliefs.

THE ECONOMY

Swaziland is principally an agricultural country, with nine-tenths of the population engaged in subsistence agriculture. Mixed farming is carried out on the Middle Veld, where rice and fruits, such as bananas and citrus, are grown under irrigation; corn, cotton, beans, tobacco, and pineapples are raised on dry lands. The Lowveld is ideally suited to cattle raising, which is little developed commercially; this region is under irrigation, and there is intensive cultivation of cash crops such as sugarcane, rice, tobacco, and citrus fruits, which are mainly grown by European farmers for export.

Resources. The nation's main export wealth derives from forestry and mining. The largest man-made forest, the Usutu afforestation estate, covers 100,000 acres of pine and eucalyptus trees planted in the Highveld. It supplies the wood pulp factory, near the town of Bunya, the products of which are a major export earner. Although not extensive, mineral resources are rich. Iron ore and asbestos deposits are located in the Highveld region, and there are also important coal deposits near Mpaka.

Mining operations

The most important mine is the iron-ore mine at Ngwenya, owned by the Swaziland Iron Ore Development Corporation and controlled by South African and British interests. The British-owned Havelock Asbestos Mine exports asbestos fibre. Third in importance are the Mpaka coal mines, mainly owned by a South African firm. Smaller mines, also foreign owned, produce kaolin, barite, and pyrophyllite.

Industry. Industry is mainly confined to the processing of agricultural products and to providing small-scale services. Manufactured products include clothing, edible oils, soap, canned fruit, and ginned cotton. There are also a slaughterhouse and a small engineering industry. Handmade woven and wooden articles are also produced.

Less than one-third of the working population is engaged in paid employment. Of this group, more than a third is employed in agriculture and related services. It was formerly traditional for Swazi labourers to engage in migratory labour in South Africa's mines and farms, but by the early 1970s the proportion that still did so had dwindled to about one-quarter of all male workers.

Trade. The bulk of all wholesale and retail trade is traditionally controlled by white South Africans. All commercial banking is foreign owned.

Swaziland depends on foreign aid—mostly from the United Kingdom—for capital and development needs. Imports of machinery, fertilizers, clothing, fuel, and vehicles largely come from South Africa. Exports—including iron ore, sugar, wood pulp, asbestos, and agricultural products—are sold to the United Kingdom, Japan, and South Africa.

The main sources of national income are taxes, mining royalties, and customs revenues. Income tax is payable on all incomes derived inside the country. A personal tax is paid by all non-Swazi adult males, and Swazi adult males pay a graduated tax based on the number of their wives. Company taxes are paid at a fixed rate, and mining taxes are on a variable scale.

Transportation. Good all-weather roads link the main population centres. The transterritorial highway runs from Oshoek, on the western frontier, to Nomahash, on the eastern border.

A railway line running from Kadake to the port of Maputo, in Mozambique, is linked to the Mozambique Railways network at Goba across Swaziland's northeast border. There is no rail link with South Africa.

The only airport is at Matsapa, five miles from Manzini. South African Airways operates scheduled services to Johannesburg and Durban, and Mozambique Airways flies to Maputo. There are also small landing strips.

ADMINISTRATIVE AND SOCIAL CONDITIONS

he overnmental ructure

Executive authority is vested in the king and exercised through a cabinet, which is presided over by the prime minister, who is appointed by the king. The cabinet advises the king and is responsible to a bicameral parliament. The House of Assembly is composed of 50 members—40 of whom are elected by an electoral college and 10 appointed by the king. There is universal franchise for all men and women over the age of 21 who are citizens of Swaziland and who are not insane and who have not been convicted of serious crimes. The Senate has 20 members, 10 elected by an electoral college and 10 nominated by the king. The civil service is controlled by a public service commission appointed by the king on the advice of the Judicial Service Commission.

Local urban government operates through town councils and urban area advisory committees. A majority of the council and committee members are elected and a minority appointed by the Minister of Local Administration. Rural administration is conducted in the four districts of Shiselweni, Lebombo, Manzini, and Hhohho by district commissioners responsible to the Minister of Local Administration.

Justice. The judicial system comprises a court of appeal, a high court, subordinate courts, and Swazi (traditional) courts. The court of appeal comprises a judge president and three justices of appeal appointed by the king. The high court, presided over by the chief justice, hears appeals in civil and criminal matters from subordinate courts and the Higher Swazi Court of Appeal. Subordinate courts, presided over by magistrates, have wide criminal jurisdiction but not over murder, sedition, or treason. There are 14 Swazi courts, two courts of appeal, and the Higher Swazi Appeal Court; these may hear cases only if all those involved are Africans and if the charges fall within a restricted list of criminal and civil offenses.

Land ownership is the most sensitive issue in Swazi national life. Under the British colonial administration about 44 percent of the land was either sold as concessions to Europeans or alienated by the government. Traditionally, however, all land is vested in trust in the king as Swazi national land, and all Swazis are entitled to a share, which is allocated by the king through the chiefs. The Swazi National Council adopted a policy in 1946 of repurchasing for the nation as much as possible of the 1,873,000 acres still held out of communal ownership by Europeans.

Educational services

Education. Half of the school-age children attend primary schools, of which most are mission (primarily Protestant) schools. A small proportion also attend secondary schools. Schooling is free except at the Waterford School at Mbabane, which is a multiracial institution for pupils from all over southern Africa. There are two teacher-training colleges. Higher education is provided at the Swaziland campus of the University of Botswana and Swaziland.

Health and welfare. Medical services are provided by the government, missions, industrial concerns, and private practitioners. The country has hospitals and rural clinics. Chief causes of illness are schistosomiasis (a parasitic infestation of the bladder or intestines), food-deficiency diseases, typhoid and paratyphoid, and respiratory diseases. Malaria has been virtually eliminated.

CULTURAL LIFE

Major national celebrations

Traditional ceremonies and music play an important part in Swazi social life. The two major cultural events are the *Incwala* (National Ceremony) held in December or January, and the reed dance (*umhlanga*) held in June or July, both at the royal village of Lobamba. The *Incwala*, the most sacred of the ceremonies, represents the king as the source of fertility and the symbol of power and unity. It lasts for six days and is a highly ritualized festival of song, dance, folklore, and martial display. The reed dance, which lasts for one week, encourages girls of the same age group to work in harmony under discipline and to preserve their virginity.

Traditional music accompanies the Swazi from birth to death, beginning with the first lullaby. There are songs for children and adults; for courting, weddings, and mourning; for working; for specific tribes and clans; and for insult. The main musical instruments are rattles, which are tied to the wrists or ankles or shaken by hand. The shield is used for percussion; buckhorn whistles and long reed flutes are the principal wind instruments.

For statistical data on the land and people of Swaziland, see the *Britannica World Data* section in the BRITANNICA WORLD DATA ANNUAL. (Co.L.)

History

According to the traditional account of the origin of the Swazi, when in the 16th century the main Bantu tribes were moving gradually southward along the coast of what is now Mozambique, a group diverged and crossed the Lebombo Range into the area that lies between the Pongola and Great Usutu rivers. Under pressure from the Zulus, the group moved northward to the Little Usutu River. Further attacks by the Zulus took place from time to time, and when King Sobhuza I died in 1836 his successor, Mswazi (Mswati) II, who reputedly gave his name to the tribe, moved farther north to the Pigg's Peak area. By that time, the Swazi had a well-trained army and conquered the tribes to the north, but in 1846 their ruler ceded any claims that he had to the country north of the Crocodile River in favour of the Lydenburg Republic. It was at that time also that Mswazi was compelled to seek British aid against the Zulus.

During the 1880s large numbers of Europeans visited the

Swazi king Mbandzeni in search of concessions. It may be said that these concessions were only intended to allow the concessionaires to share the use of the concession property with the Swazi.

In 1888 a charter of self-government was granted to the Europeans. Two years later, under a convention between the British government and the South African republic, a provisional government consisting of representatives of the two powers and a representative of the Swazi people was set up with the consent of the latter. In 1893 the British government signed a new convention permitting the South African republic to negotiate with the Swazi regent and her council for a proclamation allowing it to assume powers of jurisdiction, legislation, and administration, without incorporation in the republic. The Swazi refused to sign the proclamation, but in 1894 another convention was signed by the two powers, virtually giving effect to its terms. After the conquest of the Transvaal all rights and powers of the republic passed to Great Britain, and in June 1903 by an order in council under the Foreign Jurisdiction Act the governor of the Transvaal was empowered to administer Swaziland and to legislate by proclamation. In 1906 these powers were transferred to the high commissioner for Basutoland, Bechuanaland, and Swaziland. A commission was appointed in 1907 to deal with the problem of the concessions, and as a result of its report, land and grazing concessions were reduced by one-third. The regent did not accept this settlement, and a deputation of protest was sent to London in 1908, but it achieved little.

In 1963 the office of resident commissioner was replaced by that of Her Majesty's commissioner (equal in status to governor); in 1964 the office of high commissioner was abolished. A constitution providing limited self-government was promulgated in 1963. In 1967, under the new constitution of that year, the country became a protected state under the name of the Kingdom of Swaziland, and in the following year another and greater change took place when Swaziland became fully independent. (Jo.W.H.)

King Sobhuza II of Swaziland celebrated his diamond jubilee on September 4, 1981. Born on July 22, 1899, he was installed as the *ngwenyama* of the Swazi nation in 1921. Although he received a modern education at the Zombodze National School and at South Africa's famous missionary-run Lovedale College, the King jealously cherished and preserved Swazi traditions. Five years after Swaziland became independent in 1968, the King tore up the constitution designed by the British and restored the more ancient system of government in which all effective power remains in the royal capital, while a system of village government, known as the *tinkundlu,* operates at the grass roots. His one concession to modern government was to establish a Cabinet system and a prime minister, but all ministers were chosen by the King himself.

With a kingdom wedged between white supremacist South Africa and Marxist Mozambique, Sobhuza trod a careful path to maintain good relations with both his neighbours. He was a strong supporter of the Organization of African Unity but refused to allow Swaziland to be used as a base for guerrilla attacks against South Africa. Vehemently anti-Communist, the King maintained his strong Western ties. His was among the very few African countries that maintained close diplomatic and economic ties with Israel and Taiwan. (Co.L.)

Under the King's firm rule, Swaziland enjoyed a remarkable degree of stability and a measure of economic growth. Swaziland joined the Southern African Development Coordination Conference but maintained close diplomatic and economic ties with South Africa.

The King's death in 1982 was followed by a power struggle. The queen mother, Dzeliwe, who had assumed the regency, and her prime minister and advisory council were unseated by a faction loyal to the teenage heir, Prince Makhosetive. The Prince's mother, Queen Ntombi, was made regent in 1983, and in April 1986 the Prince was crowned King Mswati III.

For later developments in the history of Swaziland, see the *Britannica Book of the Year* section in the BRITANNICA WORLD DATA ANNUAL. (Ed.)

ZAMBIA

Zambia is a landlocked republic in south central Africa. It has an area of 290,586 square miles (752,614 square kilometres). One of the largest producers of copper in the world, Zambia is one of the richest black-controlled states on the continent. It is bordered to the west by Angola; to the northwest by Zaire; to the northeast by Tanzania; to the east by Malaŵi; to the southeast by Mozambique; to the south by Zimbabwe and a small, ill-defined border with Botswana; and to the southwest by a thin strip of South West African territory called the Caprivi Strip. A wedge of Zaire territory—known as the Pedicle—thrusts so deeply into Zambia that the northern and eastern parts of the country are united to the western and central region only by a narrow corridor about 120 miles wide. The name Zambia is derived from the Zambezi River, which forms a common border with Zimbabwe. The capital is Lusaka.

Physical and human geography

THE LAND

Relief. The greater part of Zambia forms a plateau 3,000 to 5,000 feet above sea level. The landscape has a flat or gently undulating surface, resulting from long and uninterrupted erosion of the underlying crystalline rocks. The plateau occasionally retains remnants of earlier surfaces, and there are isolated hills and ranges. In the north, the general level rises toward the Muchinga Mountains, which form the Congo–Zambezi watershed, and toward the plateau at Mbala, near the Tanzanian border, which is 5,000 to 6,000 feet (1,500 to 1,800 metres) high. In the east, toward the Malaŵian edge of the East African Rift Valley, the Nyika Plateau is generally over 6,000 feet high, rising to about 7,000 feet in the Mufinga Hills in the northeast.

From the Congo–Zambezi divide in the north, the elevation falls southward toward the Zambezi Valley. The lowest point in the country, about 1,200 feet (360 metres) above sea level, is near the confluence of the Zambezi and Luangwa rivers. Most of the country consequently drains south and east to the Indian Ocean. Only in the Luapula River area does drainage flow to the west.

Several rift valleys are wholly or partly located within Zambia. The southern tip of Lake Tanganyika, where it intrudes into Zambia's northern border, is some 2,000 feet below the plateau. Nearby, the Kalambo River, which forms part of the border with Tanzania, flows into the lake through a narrow gorge cut by the Kalambo Falls (726 feet), the highest in the country. Lake Mweru, on the Zambia–Zaire border, lies in a rift about 1,500 feet below the plateau. The largest of the rifts, however, is the valley of the Luangwa River, extending more than 350 miles (560 kilometres) from the northeastern tip of Zambia to the Luangwa's confluence with the Zambezi in the south and constituting a serious barrier to communication. In the northwest region of the plateau lies Bangweulu, a shallow depression containing the lake of the same name and the Bangweulu Swamp, approximately 4,000 square miles (10,360 square kilometres) in area and one of the largest inland swamps in the world. Bangweulu is fed by the Chambeshi River and drained by the Luapula. The Luapula joins Lake Mweru, which in turn is drained by the Luvua, one of the headstreams of the Congo. Mweru Swamp, east of Lake Mweru, is another shallow lake in a swampy region. Extensive floodplains surround the Bangweulu area, the Chambeshi and Kafue rivers, and the Barotse Plain.

The Bangweulu region

Soils. Geologically, Zambia is dominated by a basement complex of ancient sediments and widespread volcanic flows that have been folded, distorted, and metamorphosed by granitic intrusions. The oldest rocks, found along the Kafue River near Kapiri Mposhi, are about 2,600,000,000 years old. It is in the Katanga sediments (550,000,000 to 620,000,000 years old), deposited on the basement rocks, that all the minerals of economic importance so far discovered have been found. To the west is part of the Kalahari region, consisting of a blanket of unconsolidated windblown sand and gravel partly derived from, and possibly overlying, the younger sedimentary rocks of the Karroo System. These Karroo rocks, between 200,000,000 and 300,000,000 years old, have largely disappeared elsewhere in the country, except in the Luangwa and Middle Zambezi valleys.

The porous Kalahari sands contrast with the generally well-watered plateau areas. The plateaus are drained by numerous seasonal streams, as well as by *dambos,* spongy, grass-covered depressions that act as water reservoirs. The commonest soil types in Zambia are the sand to sandy clay soils of the plateau, which are generally acid, low in humus, and underlain by laterite (red, leached, iron-bearing soil). These soils form the so-called sandveld, which is of indifferent agricultural value. The soils below the main plateau, being less acid and lacking the laterite layer, are more fertile. In the bottoms of the river valleys there are areas of highly fertile alluviums, and the lake basin soils, such as occur in the Chambeshi–Bangweulu area, are similarly fertile.

Climate. Although Zambia lies within the tropics, its altitude ensures a climate that, varying widely in both season and location, is generally favourable to both human settlement and comfort. There are four seasons.

The cool season. This season lasts from May to August. Temperatures tend to decline during May, and the weather is fair to fine, with occasional ground frosts on calm nights in wind-sheltered areas. Invasions of cold air from the south or southeast may bring overcast conditions. Instability in the upper air leads occasionally to light rain or showers. A rapid increase of temperature begins about the middle of August.

MAP INDEX

Political subdivisions

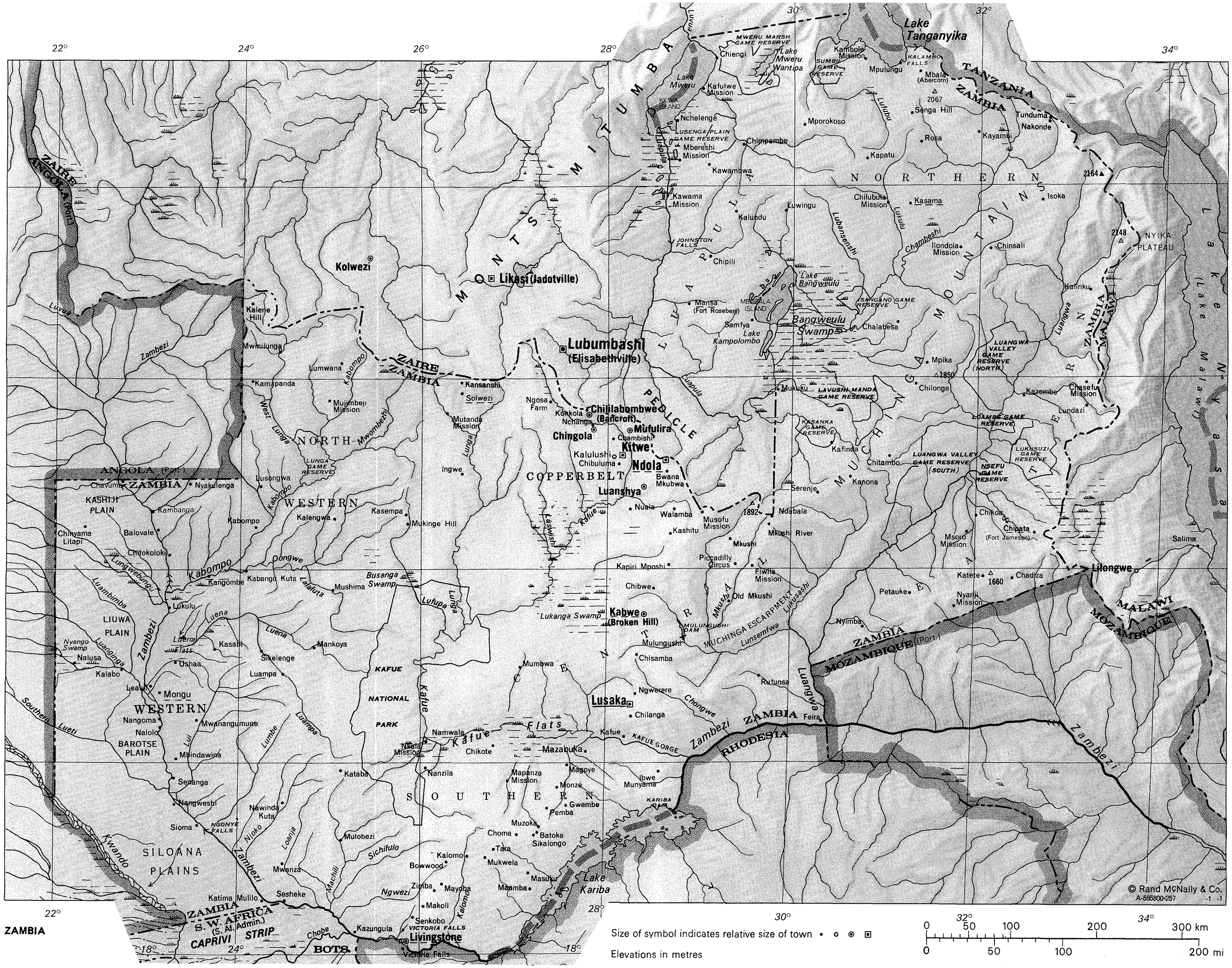
Central	15·00s	29·00e
Copperbelt	13·00s	28·00e
Eastern	13·00s	32·15e
Luapula	10·55s	29·00e
Northern	11·00s	31·00e
North-Western	13·00s	25·00e
Southern	16·00s	27·00e
Western	16·00s	24·00e

Cities and towns

Abercorn, see Mbala		
Balovale	13·33s	23·06e
Batoka	16·47s	27·15e
Bancroft, see Chililabombwe		
Bowwood	17·07s	26·17e
Broken Hill, see Kabwe		
Bwana Mkubwa	13·01s	28·42e
Chadiza	14·05s	32·28e
Chalabesa	11·22s	31·01e
Chambishi	12·40s	28·03e
Chasefu Mission	11·55s	33·08e
Chavuma	13·05s	22·40e
Chibuluma	12·51s	28·05e
Chibwe	14·12s	28·31e
Chiengi	8·39s	29·10e
Chikoa	13·24s	32·07e
Chikote	15·52s	26·54e
Chilanga	15·34s	28·17e
Chililabombwe (Bancroft)	12·18s	27·43e
Chilonga	12·03s	31·21e
Chilubula Mission	10·09s	31·00e
Chimpembe	9·31s	29·33e
Chingola	12·32s	27·52e
Chinsali	10·34s	32·03e
Chinyama Litapi	13·31s	22·21e
Chipata (Fort Jameson)	13·39s	32·40e
Chipili	10·44s	29·04e
Chisamba	14·58s	28·23e
Chitambo	12·55s	30·39e
Chitokoloki	13·50s	23·13e
Choma	16·48s	26·59e
Feira	15·37s	30·25e
Fiwila Mission	13·58s	29·36e
Fort Jameson, see Chipata		
Fort Rosebery, see Mansa		
Gwembe	16·30s	27·35e
Ibwe Munyama	16·09s	28·34e
Ilondola Mission	10·42s	31·47e
Ingwe	13·02s	26·25e
Isoka	10·10s	32·35e
Kabangu Kuta	14·00s	24·03e
Kabompo	13·15s	24·10e
Kabwe (Broken Hill)	14·27s	28·27e
Kafinda	12·39s	30·20e
Kafue	15·47s	28·11e
Kafulwe Mission	9·00s	29·02e
Kalabo	14·57s	22·40e
Kalene Hill	11·11s	24·10e
Kalengwa	13·30s	25·00e
Kalinku	11·12s	33·12e
Kalomo	17·02s	26·30e
Kalulushi	12·50s	28·03e
Kalundu	10·16s	29·24e
Kamapanda	12·00s	24·10e
Kambanga	13·23s	23·03e
Kambole Mission	8·46s	30·46e
Kangombe	14·03s	23·40e
Kanona	13·04s	30·38e
Kansanshi	12·05s	26·26e
Kapatu	9·43s	30·42e
Kapiri Mposhi	13·58s	28·41e
Kasabi	14·48s	23·42e
Kasama	10·13s	31·12e
Kasempa	13·27s	25·50e
Kashitu	13·42s	28·40e
Kataba	16·05s	25·10e
Katete	14·05s	32·07e
Katima Mulilo	17·27s	24·14e
Kawama Mission	10·04s	28·37e
Kawambwa	9·47s	29·05e
Kayambi	9·27s	31·58e
Kazembe	12·11s	32·37e
Kazungula	17·45s	25·20e
Kitwe	12·49s	28·13e
Konkola	12·17s	27·47e
Lealui	15·10s	23·02e
Livingstone	17·50s	25·53e
Luampa	15·03s	24·28e
Luanshya	13·08s	28·24e
Lukulu	14·25s	23·12e
Lumwana	11·50s	25·10e
Lundazi	12·19s	33·13e
Lusaka	15·25s	28·17e
Lusongwa	12·58s	24·16e
Luwingu	10·15s	29·55e
Maamba	17·23s	27·10e
Magoye	16·00s	27·37e
Makoli	17·27s	26·05e
Mankoya	14·47s	24·48e
Mansa (Fort Rosebery)	11·12s	28·53e
Mapanza Mission	16·15s	26·55e
Masuku	17·12s	27·07e
Mayoba	17·13s	26·16e
Mazabuka	15·51s	27·46e
Mbala (Abercorn)	8·50s	31·22e
Mbereshi Mission	9·45s	28·46e
Mbindawina	15·57s	23·18e
Mkushi	13·38s	29·23e
Mkushi River	13·32s	29·45e
Mongu	15·15s	23·09e
Monze	16·16s	27·28e
Mpika	11·54s	31·26e
Mporokoso	9·23s	30·05e
Mpulungu	8·46s	31·07e
Msoro Mission	13·36s	31·55e
Mufulira	12·33s	28·14e
Mujimbeji Mission	12·11s	24·57e
Mukinge Hill	13·29s	25·52e
Mukuku	12·09s	29·49e
Mukwela	17·02s	26·39e
Mulobezi	16·48s	25·09e
Mulungushi	14·40s	28·50e
Mumbwa	14·59s	27·04e
Mushima	14·13s	25·05e
Musofu Mission	13·31s	29·02e
Mutanda Mission	12·24s	26·16e
Muzoka	16·41s	27·19e
Mwanangumune	15·31s	23·30e
Mwanza	17·02s	24·27e
Mwinilunga	11·44s	24·26e
Nakonde	9·20s	32·42e
Nalolo	15·35s	23·07e
Nalusa	14·55s	22·13e
Namwala	15·45s	26·26e
Nangoma	15·30s	23·08e
Nangweshi	16·26s	23·17e
Nanzila	16·05s	26·07e
Nawinda Kuta	16·25s	24·28e
Nchanga	12·30s	27·53e
Nchelenge	9·20s	28·50e
Ndabala	13·28s	29·50e
Ndola	12·58s	28·38e
Ngosa Farm	12·18s	27·28e
Ngwerere	15·18s	28·20e
Nkala Mission	15·55s	26·00e
Nuala	13·27s	28·16e
Nyakulenga	13·03s	23·29e
Nyanji Mission	14·23s	31·48e
Nyimba	14·35s	30·52e
Old Mkushi	14·22s	29·22e
Pemba	16·31s	27·22e
Petauke	14·15s	31·20e
Piccadilly Circus	13·56s	29·24e
Rosa	9·38s	31·21e
Rufunsa	15·05s	29·40e
Samfya	11·21s	29·32e
Senanga	16·06s	23·16e
Senga Hill	9·22s	31·12e
Senkobo	17·38s	25·58e
Serenje	13·11s	30·52e
Sesheke	17·29s	24·18e
Sikalongo	16·46s	27·07e
Sikelenge	14·50s	24·14e
Sioma	16·39s	23·30e
Solwezi	12·11s	26·25e
Tara	16·56s	26·47e
Tunduma	9·19s	32·47e
Ushaa	14·55s	23·18e
Walamba	13·29s	28·45e
Zimba	17·19s	26·13e

Physical features and points of interest

Bangweulu, Lake	11·05s	29·45e
Bangweulu Swamp	11·30s	30·15e
Barotse Plain	15·40s	23·10e
Busanga Swamp	14·10s	25·50e
Chambeshi, *river*	11·21s	30·37e
Chongwe, *river*	15·43s	29·20e
Dongwe, *river*	13·58s	23·53e
Isangano Game Reserve, *wildlife refuge*	11·10s	30·40e
Johnston Falls, *waterfalls*	10·35s	28·40e
Kabompo, *river*	14·10s	23·11e
Kafue, *river*	15·56s	28·55e
Kafue Flats	15·40s	27·25e
Kafue Gorge	15·54s	28·34e
Kafue National Park	14·30s	26·10e
Kalambo Falls, *waterfalls*	8·36s	31·14e
Kalomo, *river*	17·57s	26·24e
Kampolombo, Lake	11·37s	29·42e
Kariba, Lake, *reservoir*	17·00s	28·00e
Kariba Dam	16·30s	28·50e
Kasanka Game Reserve, *wildlife refuge*	12·35s	30·12e
Kashiji Plain	13·20s	22·30e
Kilwa Island	9·20s	28·33e
Kwando, *river*	17·35s	23·20e
Lalafuta, *river*	13·57s	24·41e
Lavushi Manda Game Reserve, *wildlife refuge*	12·20s	30·50e
Liuwa Plain	14·30s	22·40e
Loanja, *river*	17·22s	24·48e
Luambe Game Reserve, *wildlife refugé*	12·25s	32·15e
Luambimba, *river*	15·00s	22·48e
Luampa, *river*	14·33s	24·10e
Luanginga, *river*	15·11s	22·56e
Luangwa, *river*	15·40s	30·25e
Luangwa Valley Game Reserve (North)	11·50s	32·15e
Luangwa Valley Game Reserve (South)	12·50s	31·45e
Luapula, *river*	9·26s	28·33e
Lubansenshi, *river*	11·21s	30·35e
Luena, *river*	14·45s	23·25e
Luena Flats, *swamp*	14·50s	23·20e
Lufubu, *river*	8·36s	30·47e
Lufupa, *river*	14·37s	26·12e
Lui, *river*	16·21s	23·18e
Lukanga Swamp	14·25s	27·45e
Lukulu, *river*	10·56s	31·05e
Lukusashi, *river*	14·38s	30·00e
Lukusuzi Game Reserve, *wildlife refuge*	12·50s	32·35e
Lumbe, *river*	16·42s	23·42e
Lunga, *river*	14·34s	26·25e
Lunga Game Reserve, *wildlife refuge*	12·55s	25·10e
Lungwebungu, *river*	14·19s	23·14e
Lunsemfwa, *river*	14·54s	30·12e
Lusenga Plain Game Reserve, *wildlife refuge*	9·30s	29·10e
Luswishi, *river*	13·55s	27·24e
Machili, *river*	17·26s	25·02e
Mbabala Island	11·18s	29·44e
Mkushi, *river*	14·40s	29·07e
Muchinga Escarpment	13·40s	31·00e
Muchinga Mountains	12·00s	31·45e
Mulungushi Dam	14·40s	28·50e
Mweru, Lake	9·00s	28·45e
Mweru Marsh Game Reserve, *wildlife refuge*	8·45s	29·30e
Mweru Wantipa, Lake	8·45s	29·40e
Mwombezhi, *river*	12·52s	25·00e
Ngonye Falls, *waterfalls*	16·40s	23·35e
Ngwezi, *river*	17·40s	25·07e
Njoko, *river*	17·10s	24·05e
Nsefu Game Reserve, *wildlife refuge*	13·07s	32·10e
Nyengo Swamp	14·51s	22·07e
Sichifulo, *river*	17·26s	25·02e
Siloana Plains	17·00s	23·15e
Southern Lueti, *river*	16·14s	23·13e
Sumbu Game Reserve, *wildlife refuge*	8·50s	30·25e
Tanganyika, Lake	8·30s	30·50e
Victoria Falls, *waterfalls*	17·55s	25·51e
West Lunga, *river*	13·06s	24·39e
Zambezi, *river*	15·40s	30·25e



ZAMBIA

© Rand McNally & Co.
A-565800-257 -1 -1

Size of symbol indicates relative size of town • ○ ◉ ▣

Elevations in metres

0 50 100 200 300 km

0 50 100 200 mi

The hot season. This season, from August to October, consists of rising temperatures that continue through September, interrupted by occasional invasions of cold air. Rain is unlikely. In October, temperatures rise more slowly, reaching the highest levels of the year in mid-month, after which there is a tendency for cooler oceanic air to move in, increasing humidity and cloud formation. On the northwestern border, moist air from the Atlantic often causes thunderstorms. Storms often occur in early November, but rains are frequently scattered and are undependable.

The rainy season. The rainy season lasts from November to April. A westerly airstream flowing in from the Atlantic converges with the southeast trade winds (which usually cross Zambia as northeast winds), giving rise to the Congo air boundary that moves into Zambia from the northwest in November, bringing the first rains. The intertropical convergence zone (continental hot, dry air coming southeast from the Sahara to meet maritime moist, relatively cool air coming northeast) moves down from the north, entering Zambia in early December to bring the main rains as it proceeds south. In January or February there is a temporary decrease of rainfall in the north and northwest, followed by an increase as the intertropical convergence zone passes back to the north.

The post-rainy season. This season falls in April and May. In this transitional period, the air is still fairly moist, and clouds form daily. Showers occur occasionally in April and exceptionally in May. In April, day temperatures tend to rise and night temperatures to fall as cloud cover decreases, while in May both day and night temperatures are falling. Since out-of-season rainfall in Zambia is negligible, the mean annual rainfall is virtually identical with the seasonal average and ranges from about 58 inches (1,475 millimetres) in the north to about 28 inches in the south. Places with an average annual rainfall of 35 inches may expect rain to fall on half the days from November to March. The highest daily rainfalls do not usually exceed four to five inches. The highest daily rainfall on record is 12.11 inches.

empera- re ranges

Temperatures are generally moderate, due partly to the altitude and partly to the cooling effect of showers and of cloud shade during what would otherwise be the hottest time of the year. The extremes of temperature recorded are 111° F (44° C) in the Luangwa Valley in the east, at 1,230 feet above sea level; and 19° F (−7° C) at Sesheke in the southwest, at 3,120 feet. Throughout most of the country, however, extreme temperatures seldom exceed the range of 30° to 100° F (−1° to 38° C).

Average annual hours of sunshine range from over 3,000 in the southwest to under 2,500 on the eastern border. Relative humidity ranges from under 40 percent in the dry season to over 80 percent in the rainy season.

Plant and animal life. The plateau vegetation is dominated by wooded savanna (grassy parkland). These woodlands consist mostly of small leguminous trees of the *Brachystegia, Isoberlinia,* and *Julbernardia* species. Since they rarely form a dense leaf canopy, they permit the growth of tall perennial grasses or herbs. These grasses readily catch fire in the dry season, causing the trees to develop a corky, fire-resistant bark.

Low regions, such as the Zambezi and Luangwa valleys, are dominated by the mopane tree and short-lived annual grasses and by a few perennial herbs. In the sandy Kalahari area of the southwest, there are economically valuable forests of evergreens, notably Rhodesian teak. Few other specific species are commercially exploited, though mukwa (*Pterocarpus angolensis*), a good furniture timber, is found in the Bangweulu Depression. The ordinary woodlands, nevertheless, provide mining timber, cordwood, charcoal, building material, forage, and supplementary foods. More than 8 percent of the country has been set aside as forest reserve or protected forest area.

There are extensive areas of poor drainage that do not support tree growth; some of the largest of these are found in the major swamps and floodplains. These grasslands support livestock as well as wildlife. The resumption of the traditional practice of burning off the vegetation after crops have been harvested has led to an increase in the area of secondary grassland and has tended to increase the danger of soil erosion, as well as to lower the productivity of the land.

The mammals of Zambia are more notable for their variety than for their numbers. The predominant woodlands have never supported many large animals, though large concentrations occur on the major floodplains. While no species is in danger of extinction, a depletion of wildlife has occurred because of hunting, the increasing use of land for agriculture as a result of population increase, and the elimination of game as a measure to help eradicate tsetse flies.

About one-third of Zambia is either infested by tsetse flies or consists of marginal lands. A number of areas unsuitable for human settlement are game reserves. These include the Kafue National Park (8,650 square miles) and the Luangwa Valley game management areas (5,040 square miles).

Elephants are numerous; other animals found in Zambia include the genet, a small carnivorous animal related to the civet cat; hyena; wild dog; jackal; and ratel, a nocturnal animal resembling a badger; and such cats as the serval, leopard, cheetah, and lion. Hoofed animals include black rhinoceros, zebra, hippopotamus, warthog, bush pig, giraffe, buffalo, and several species of antelope. Two species of baboon, four species of monkey, and two species of galago (known as bush babies) are also found. There are many smaller mammals.

Bird life is varied and numerous; about 700 species have been recorded. These include shoebills, secretary birds, wood hoopoes, and guinea fowl—families unique to Africa—as well as many other tropical species. Lochinvar National Park on the Kafue Flats is the best bird-viewing area, with almost 400 recorded species. The other national parks also have a wealth of bird life, outstanding being the water birds of the Busanga Plain and the carmine bee eater of the Luangwa Valley. The fish eagle, Zambia's national emblem, is common on large areas of water.

Fish are found not only in Lake Tanganyika in the far north but also in the Luapula Region, which includes Lake Mweru and Lake Bangweulu, and in the Zambezi River system, which includes the Kafue River and Lake Kariba. Commercial fisheries catch bream, barbel, and tiger fish in these waters. The Lake Tanganyika sardine, commonly called *kapenta,* is considered a national dish. The best game-fishing areas for tiger fish are the Kafue River, Lake Kariba, the upper Zambezi, and Lake Tanganyika.

Reptiles include crocodiles, tortoises, terrapins (freshwater turtles), a variety of lizards (including geckos, agamas, and skinks), and many poisonous and nonpoisonous snakes. Insects of most orders are prevalent.

THE PEOPLE

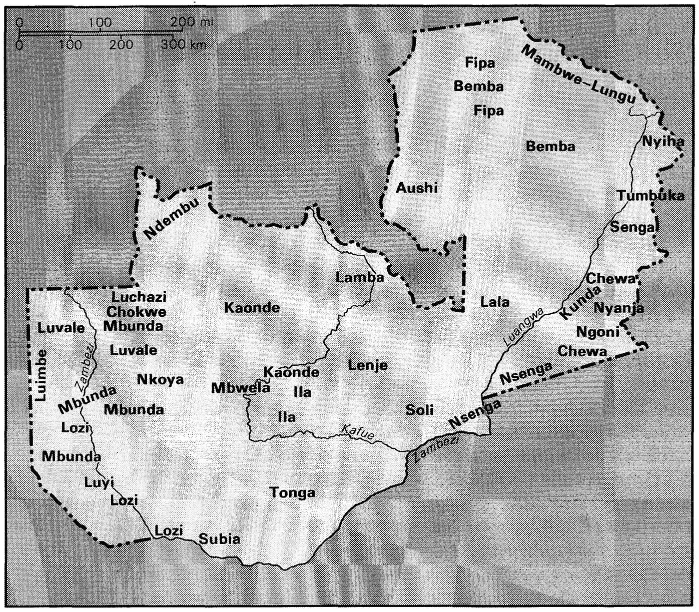
Zambia's population is small in relation to its area. Within the population, the small white (European) minority exercises a disproportionate influence on the social, economic, and—until recently—political development of the country. Two-thirds of the African population is still engaged in traditional subsistence agriculture, although the introduction of a money economy has encouraged many young men to move to the towns.

Tribal peoples

The African peoples of Zambia—with the exception of a few bands of San (Bushman) peoples living on the fringes of the Kalahari and the Twa found in the Bangweulu Swamp, the middle Kafue Valley, and the Lukanga Swamp—are all of Bantu origin and for the most part descendants of immigrants from the Congo region. Within this grouping, however, there is wide cultural and linguistic variety. Tribal affiliations have not been recorded since 1962, but the census of 1969 included a question on mother tongues, from which linguistic groupings can be tentatively derived. The census indicated that about 34 percent of Zambia's population belonged to the Bemba or near-Bemba group, 17 percent to the Nyanja group, 15 percent to the Tonga group, 11 percent to the North-Western Province group, 9 percent to the Lozi group, 5 percent to the Mambwe group, and 5 percent to the Tumbuka group. There are also some Zambians belonging to the Swahili-speaking and other minorities.

The larger aggregations contain numerous small, separately identified groups. In 1962 more than 90 tribal names with membership ranging from a few hundred to 250,000 were listed, but many of these small groups would meet only a few of the accepted criteria for classification as a separate tribe. Similarly, mutually intelligible languages or dialects cover large areas and populations. Radio broadcasting employs seven languages—Bemba, Tonga, Nyanja, Lozi, Kaonde, Lunda, and Luvale—of which the first four are understood far beyond the mother tongue boundaries. The extended ties of kinship involved in the traditional system of society have continued to exert a powerful influence even in the urban areas.

The traditional regions within the country approximately coincide with the administrative division into provinces. The Western (formerly Barotse) Province is dominated by the Lozi, who live chiefly on the floodplain of the Zambezi River. They have an unusually mixed economy, based on garden culture, cattle ownership, fishing, and hunting. They also have a strong tradition of government and law and a distinct aristocracy.



General ethnic composition of Zambia.

The North-Western Province, adjoining the Angolan and Zaire borders, is occupied by tribes descended from the great Luba-Lunda Kingdom of Zaire. These include the southern Lunda and the Luvale, Chokwe, Luchazi, Mbunda, Ndembu, and Kaonde. The area is thinly populated except along the Zambezi, and the economy is based on the cultivation of cassava and millet and on hunting.

The Ila-Tonga group and the Bemba

The Southern Province contains the Ila-Tonga group of about 12 tribes, speaking closely related dialects. The third most numerous group in the country, it was probably among the earliest established. The Ila-Tonga have no chiefs. They are traditionally cattle owning but also cultivate grain crops.

The Northern Province is dominated by the Bemba, who trace their origins to the Luba-Lunda Kingdom and have a strongly centralized chieftainship, led by a paramount chief. Bemba is the lingua franca of the Copperbelt, and Bemba language and culture have spread throughout the province, as well as into nearby areas of Zaire. The Northern Province is generally poor in soil, and there is no tradition of cattle owning. Agriculture is of the *chitemene* system (*i.e.,* clearing land for temporary cultivation by cutting and burning vegetation), which is thought to enhance fertility; the usual crop is millet. The different tribes dwelling in the northeastern area of the Northern Province are, with perhaps the exception of the Senga, for the most part connected with the peoples of Tanzania and Malaŵi.

The Eastern Province is inhabited by four tribes, of which three—the Nsenga, Chewa, and Kunda—originated in Zaire. These three were conquered by the fourth, the warlike Ngoni, who are related to the cattle-owning Zulus of South Africa. Many Ngoni customs have, however, disappeared, and the Ngoni language has been almost entirely replaced by Nyanja, a tongue of neighbouring Malaŵi.

The ethnic boundary between the Ila-Tonga and the Lala-Lamba groups runs approximately through the Central Province, with the Lenje-Soli peoples occupying a buffer area between the two. The Lenje are related to the Ila-Tonga and the Soli to the Lala-Lamba, who, in turn, are connected with the Kaonde of the North-Western Province.

The Copperbelt (formerly Western) Province is the location of the mining industry of the Copperbelt and its surrounding districts. There the population, both densely concentrated and diverse, is composed primarily of Africans from all parts of Zambia, with about a tenth of the total from neighbouring countries. This pattern is repeated among the populations living along the railway that stretches from the Copperbelt to Livingstone. Zambian policy nevertheless discourages the employment of alien Africans and restricts their entry.

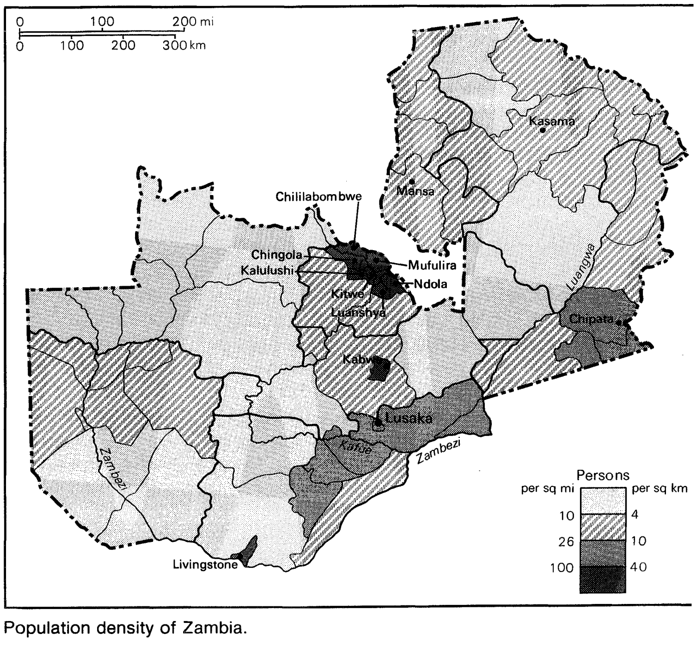
The decline in the populatior

Most of the white population lives in Kitwe, Ndola, and Lusaka or other towns of the railway belt. The European population is multinational in origin but includes notable South African and British elements. Since independence was gained in 1964, restrictions on permanent residence by aliens and the introduction of a contract system of employment for new residents have engendered instability among the white population. The government's policy of Zambianization, aided by the rapid development of education and training, has brought about a gradual decline in the number of whites. Nearly all whites are nominally Christian, with Jews probably the next largest group.

There are several thousand Indians, nearly all of whom are Gujarātī speakers from western India. About two-thirds are Hindu and the rest Muslim. Most of them arrived in Zambia between 1945 and 1954, after which Indian immigration was severely restricted. Most Indians live in Lusaka, Ndola, Livingstone, Kabwe, or the town and district of Chipata, and the majority are shop owners or employees.

There are also a few thousand persons of mixed race, mostly European-African, who are chiefly settled in Lusaka, Ndola, and Chipata.

In religion, the majority of the African population is animist. About a quarter of the remainder are nominally Christian; of these, more than half are Catholic, and the remainder divided among other denominations, which include Jehovah's Witnesses and African separatist churches, and nativistic movements.



Population density of Zambia.

THE ECONOMY

Resources. Zambia has been one of the few African countries that, since independence, has had sufficient income to finance most of its economic and social development. As the fourth largest copper-producing country in the world, Zambia's copper output is exceeded only by that of the United States, the Soviet Union, and Chile. Most economic growth takes place in the Copperbelt or along the railway to the south, despite government efforts to divert development resources to the countryside.

he
sources
' the
opperbelt

The Copperbelt, an area about 70 miles long by 30 miles wide, flanks the border of the Zaire Pedicle and virtually adjoins the industrialized mining area of Zaire's Shaba (formerly Katanga) Province, which belongs to the same geological field. There are 10 producing mines—Luanshya, Mufulira, Chibuluma, Chambishi, Rokana (near Kitwe), Nchanga (near Chingola), Bwana Mkubwa, Konkola, Lufubu, and Baluba—with ore reserves of over 882,000,000 tons, which constitute about one-eighth of the world's known and exploitable copper reserves.

Expansion plans for the Copperbelt area include development of the Baluba ore body and expansion of several other mines. Full production began in 1970 at Kalengwa, a small, high-grade copper-ore deposit about 200 miles west of the Copperbelt; and the old workings at Kansanshi, 100 miles northwest of Chingola, resumed production in 1973. About 100 miles north of Kalengwa, at Lumwana, large tonnages of low-grade copper mineralization are being evaluated. Mkushi Mines Limited is redeveloping the copper deposit at Mkushi, which lies southeast of the Copperbelt.

The Broken Hill mine at Kabwe produces zinc and lead. Coal was first produced in 1966 from deposits in the Zambezi Valley. Extensive iron-ore deposits west of Lusaka and another ore body on the Zaire border near Chililabombwe may provide the basis for an iron and steel industry.

Limestone, underlying large areas in the Lusaka and Copperbelt districts, is quarried for stone, lime, and cement; amethyst is produced from mines near the Zambezi Valley; and emeralds have been produced from a small deposit near Luanshya.

Power comes from the hydroelectric stations on the Kariba Dam, at Victoria Falls, and at the Kafue River project. The Copperbelt mines account for 70 percent of consumption.

Agriculture, forestry, and fishing. While mining and quarrying account for more than 90 percent of the value of domestic exports, and about 20 percent of gross domestic product, its contribution to paid employment is only about 15 percent. Agriculture provides the livelihood of about 60 percent of the population, with most of the production being absorbed by domestic consumption; exports of tobacco, peanuts (groundnuts), and corn (maize) account for less than 2 percent of export earnings. About half the total output is produced by the subsistence sector and half by commercial farmers, mostly European. The decline of commercial farming, which produces about three-quarters of agricultural exports, has caused concern. The difficulty of transforming subsistence agriculture into commercial farming is another major problem. The staple food crops grown by villagers are millet, maize, and sorghum, together with cassava. Sorghum and millet may be grown on relatively poor soils and under fairly low rainfall. Maize matures relatively quickly and can be eaten green in the so-called hunger months (approximately January and February). The cultivation of cassava continues to spread, despite its poor nutritional value. Secondary food crops include sweet potato, taro, yams, peanuts, peas, beans, pumpkins, sugarcane, bananas, rice, and small quantities of other common fruits and vegetables.

ıe
ecline in
ımmer-
al
rming

Four statutory boards are concerned with the marketing of tobacco, maize, peanuts, cotton, cattle, and milk. Maize production has fluctuated, and the government's aims are to achieve self-sufficiency, to establish a reserve, and thereafter to reduce costs and stimulate exports. Tobacco, mostly Virginia flue-cured, is an important export crop, but production has declined. Although there are training schemes for African tobacco farmers, the continued emigration of white growers, the competition of livestock farming, and an apparent shift to maize production have hindered the government in its efforts to promote an increased tobacco output. It is, nevertheless, hoped to raise Virginia production by means of a price support policy.

Commercial herds of beef and dairy cattle have declined with the departure of white farmers, but the state has invested in ranching, and African owners are encouraged to rear and sell cattle through grazier (rancher) schemes. Zambia is self-sufficient in poultry and eggs but not in beef or dairy products.

About half of the peanut (groundnut) crop, most of which comes from the Eastern Province, is exported for confectionery manufacture. Large amounts of peanuts and peanut oil are imported to meet domestic oil demand.

Improved pest control has resulted in an expansion of cotton growing. Zambia's first textile mill was opened at Kafue in 1968. Sugarcane is grown under irrigation near Mazabuka, but efforts to attain self-sufficiency have met difficulty in keeping pace with the growing demand for sugar.

Timber
resources

About 26,000 square miles of Zambia is reserved as forest estate, and products include sawlogs, poles, fuel, and charcoal wood. The principal timber resources are the vast forests of Rhodesian teak in the Western Province, which have been exploited since 1912 and which produce railway sleepers (ties), timber, and unfinished parquet. The government also has established vast plantations of exotic trees; about one-quarter of the acreage has been planted with tropical pines and the remainder with hardwoods, mainly eucalyptus. Much of the production of exotic trees is used by the mining industry, and the most extensive plantations are consequently located in the Copperbelt Province.

The major fisheries are in Lakes Mweru, Bangweulu, Tanganyika, and Kariba, in the Lukanga and Mweru swamps, and in the Kafue River. The largest contributor is Lake Bangweulu, followed by Lake Tanganyika and Lake Mweru. Production from Lake Kariba has declined—a phenomenon that is being investigated by the Central Fisheries Research Institute.

Administration of the economy. In 1968, in the Mulungushi Declaration, the government of Zambia indicated its intention of fostering a mixed economy in which both public and private enterprise would operate and government intervention would occur where private initiative was absent or inadequate or where developments were not in conformity with public policy—for example, in the Africanization of personnel, in price restraint, and in disengaging economically from Rhodesia and South Africa. To this end, the government acquired equity participation—usually of 51 percent or more—in a number of foreign-owned firms, most of which were engaged in wholesale and retail trade or in transport. These acquisitions were followed in 1969 by a proclamation that all rights of ownership or partial ownership of minerals should revert to the state. Subsequently the two copper-mining companies—the Anglo American Corporation and Roan Selection Trust groups—transferred 51 percent of their shares in Zambian mining enterprises to the state under a compensation agreement.

All the government's equity participation is acquired and managed through state-owned companies. These are the Zambia Industrial and Mining Corporation Ltd. (Zimco) and its two wholly owned subsidiaries—Indeco Ltd., which administers the government's interests in industrial and commercial activities, and Mindeco Ltd., which is the holding company for Zambia's equity participation in mining ventures. The mining interests of the Anglo American Corporation group are consolidated in Nchanga Consolidated Copper Mines Ltd. and those of the Roan Selection Trust group in Roan Consolidated Mines Ltd.

Industrial
development
projects

Indeco Ltd. has also been the vehicle for accelerated industrial development since 1965. By 1970, completed projects included a nitrogenous fertilizer factory, a textile mill, a copper-fabricating plant, the organization of hessian and grain-bag production, an explosives factory, the Tanzania–Zambia oil products pipeline, a vehicle-tire plant and two luxury hotels—all of which include the participation of technical partners.

Transportation. Before Rhodesia's Unilateral Declaration of Independence (UDI) in November 1965, most of Zambia's imports and exports were moved on the three-foot six-inch gauge single-line railway system. The line was constructed along a watershed route, entering the country through Livingstone and reaching Ndola in 1909. Southward, the railway connects with the Zimbabwean, Mozambique, and South African railways. Northwest of Ndola, it enters Zaire to connect with the Benguela railway system, which reaches the sea at the Angolan port of Lobito.

After UDI, however, in accordance with the sanctions policy of the British government and the United Nations and in an effort to reduce its dependence on a single exit route, Zambia took measures to develop alternative routes to the sea. Over half of Zambia's exports and imports are now moved westward by the Benguela railway system or eastward along the Great North Road to Dar es Salaam in Tanzania.

Copper-belt–Dar es Salaam railway

The leaders of both Zambia and Tanzania long favoured the construction of a rail link between the Copperbelt and the port of Dar es Salaam. Following failure to obtain Western support for the project, an agreement was signed in Peking in 1970 under which the People's Republic of China loaned Zambia and Tanzania $400,000,000 including labour and equipment to cover the costs of a 1,060-mile rail link with Dar es Salaam.

Since independence in 1964, efforts have been made to improve the national road system. Internal roads have also been improved; official policy is to bituminize all roads from provincial centres to Lusaka as finance becomes available. Most roads are of semipermanent construction—usually gravel or unsurfaced.

An eight-inch petroleum-products pipeline from Dar es Salaam to Ndola was completed in 1968.

Despite the existence of several major rivers, waterways are of little economic importance. Few tributaries have a perennial flow, and there are numerous natural obstacles in the Kafue, Zambezi, and Luangwa rivers. The three lakes in the north of the country are used for some local transport, and Mpulungu, a small port on Lake Tanganyika, which connects with the East African Railway system, handles a small quantity of goods.

Zambia Airways

Zambia Airways Corporation was established in 1967. Internal services connect provincial centres with the capital, and there are landing strips for light aircraft at most administrative posts and at some mission stations. There is an international airport at Lusaka, airports at Ndola and Livingston, and landing grounds at all the Copperbelt towns.

ADMINISTRATIVE AND SOCIAL CONDITIONS

Government. In August 1973 Zambia's second constitution since independence came into effect. In accordance with the dissolution of all opposition political parties in 1972, it laid the foundations for a one-party state, the government structure paralleling that of the United National Independence Party (UNIP).

The president is the head of state and commander in chief of the armed forces. He is elected by universal adult suffrage to a five-year term of office. The president is empowered to appoint the prime minister, the chief justice and members of the high court, and the secretary general of the UNIP. In the case of the president's absence, his duties are assumed by the secretary general of the party. The president also appoints his cabinet from elected members of the National Assembly. The cabinet is subordinate to the UNIP Central Committee; most of the committee's 24 members are elected in separate elections every five years.

The legislature, called the National Assembly, includes 125 elected members and up to 10 members appointed by the president. Assemblymen are elected by universal adult suffrage from three candidates in each constituency who have been nominated by an electoral college of local party chiefs and approved by the UNIP Central Committee. The 27-member House of Chiefs may consider bills but may not block their passage.

Central government is represented throughout Zambia by the provincial government system, according to which the president appoints a resident Cabinet minister to each of eight provinces. (Lusaka is excluded). Each minister is the president's direct representative and is responsible for the coordination of policy and for liaison with local political parties. Provincial administration is carried out by Central Committee members who each have a senior civil servant as permanent secretary.

The nine provinces are divided into 53 districts, each of which has a district governor responsible to the provincial Cabinet minister; the district governor is particularly concerned with political and economic developments. His civil-service counterpart is the district secretary.

Local government is controlled by the Minister of Provincial and Local Government and Culture. Lusaka, Kitwe, and Ndola have city councils, and there are five municipal councils and 24 township councils. All of these bodies have elected majorities, although the minister may appoint up to three additional councillors to each council.

There are also 34 rural councils with a majority of elected members and up to three members (mainly chiefs and government officials in the district) nominated by the minister. Finally, there are eight mine township management boards that advise mine managements on the needs of residents in the mine areas. They have very limited powers and their revenues and services are provided by the mining companies. Government policy is ultimately to combine these townships into the adjoining local government areas. Local government in the urban areas is wealthy and comparatively powerful, but not in the rural areas, where sources of revenue are slender. Most important public services in the rural areas are provided by various ministries of the central government, under the coordination at district level of district governors and other officials.

Justice. The court system consists of the Supreme Court, the High Court, subordinate magistrate courts, and local courts. The law administered by all except the local courts being based on English common law, decisions of the higher British courts are of persuasive value; a few statutes of the United Kingdom Parliament that were declared by ordinance (decree) to apply to Zambia are in force so far as circumstances permit. Most of the laws on the statute book, however, have been locally enacted by ordinance or, since independence, by Zambian acts.

The Supreme Court consists of the chief justice, and four other justices; it is the court of last resort. The High Court, presided over by the chief justice, has 12 puisne judges and is basically an appellate court. There are four classes of magistrates' court, with progressive degrees of criminal and civil jurisdiction. Local courts consist of a president sitting alone or with other members, all appointed by the Judicial Service Commission. Jurisdiction is conferred by the minister of justice and may encompass any written law, but punishment powers are limited. Local courts also deal with civil cases of a customary nature. Customary law is followed when it is not repugnant to justice or equity and when it is not incompatible with other legislation.

The independence of the judiciary is maintained under the constitution. The chief justice is appointed by the president, and the judges are appointed by the president on the advice of the Judicial Service Commission.

Education. Since independence, Zambia claims to have maintained the highest rate of educational expansion in Africa. The country is now beginning to attain its objective of a seven-year primary-education course for all children. All districts have secondary schools. Enrollments for adult-education classes have increased.

Universit of Zambi

The University of Zambia was opened in 1966 in Lusaka. The first students received their degrees in 1969. There are four-year courses in humanities, social sciences, natural sciences, education, law, social work, business and public administration, and agricultural sciences. There is a school of medicine with a seven-year course and a five-year engineering course.

Technical training is being encouraged to meet the demand for skilled workers. The Commission for Technical Education and Vocational Training is supervising the expansion of institutions to accommodate more trainees and to provide a wider choice of courses. Teacher training has also been much expanded, but large numbers of white teachers are still being recruited.

laria, harziasis, d other pical eases

Health and welfare. Malnutrition, caused by poverty, lack of dietary knowledge, and overcrowding, is widespread in Zambia. The most prevalent tropical diseases are malaria, bilharziasis (schistosomiasis), and other parasitic infections, such as hookworm, and leprosy. With the development of out-patient treatment, the number of lepers is declining. Trypanosomiasis (sleeping sickness, transmitted by the tsetse fly) occurs in some sparsely populated areas, but it is more an economic than a medical threat in reducing the land available for cattle raising. Smallpox and typhoid fever have been successfully controlled through immunization programs. The largest single cause of hospital admissions and deaths is the respiratory group, followed by accidents and injuries (Zambia has one of the world's highest rates of motor accidents), and gastroenteric disorders.

The government has devoted large sums to the extension of the hospital system, and training establishments—apart from the university medical school—include a school of hygiene and nurse and medical-assistant training schools. A government-financed "flying doctor service" and a voluntary organization called "Mission Medic-Air" provide medical services to some remote rural areas and to mission hospitals. Psychiatric services are based on a central hospital at Lusaka, to which are linked small psychiatric units at other centres. There is a specialist pediatric hospital at Ndola. Generally, however, Zambia suffers from a shortage of trained medical staff.

In traditional Zambian society, welfare services are still largely part of the obligations of kinship groups. A rapidly growing urban population, in which family ties are weakened, has, however, necessitated the development of government welfare services concerned with juvenile delinquency, adoption, refugee control, and care for the aged, the indigent, and the handicapped. A contributory National Provident Fund provides retirement benefits.

A shortage of good housing is a pressing problem. Dwellings range from rural huts and urban shanties at one extreme to the large brick and tile homes of the wealthier town dwellers at the other. Zambia is, nevertheless, among the most urbanized countries south of the Sahara. The Department of Community Development is responsible for the improvement of both urban and rural housing and for self-help development projects.

CULTURAL LIFE

Traditional Zambian art consists chiefly of wood carving, pottery, basket weaving, and, in some areas, house-wall paintings. Among musical instruments, drums are the most widely used, but there are also stringed bows, flutes, horns and pipes, xylophones, bells, rattles, and the sansa, or "African piano," made of strips of steel attached to a small board at one end and vibrated by the fingers at the other.

Music, dancing, and song are used in tribal rituals and celebrations, as well as for entertainment, varying in form from tribe to tribe. In urban settings, traditional music has often been modified by the use of Western instruments, such as the guitar. The introduction of Western popular music and dance has had a strong influence in the urban areas.

The Zambia Department of Cultural Services was formed in 1966, with a mandate to coordinate the activities of cultural institutions; to arrange participation in national and international cultural festivals, conferences, and exhibitions; and to disseminate information on cultural life and activities.

The department also records and analyzes Zambian oral traditions and folklore, and the Kenneth Kaunda Foundation was established in 1967 to encourage the publication of Zambian literature, as well as the preparation of Zambian-oriented school textbooks. There is a national museum at Livingstone, a small museum on the Copperbelt, and a Commission for the Preservation of Natural and Historical Monuments and Relics.

For statistical data on the land and people of Zambia, see the *Britannica World Data* section in the BRITANNICA WORLD DATA ANNUAL. (Ri.H.)

History

The discovery in Zambia of split and crudely trimmed pebbles similar to the tools made by *Homo habilis* and found associated with his remains at the Olduvai Gorge in Tanzania lends support to the possibility that earliest man may also have roamed over Zambia between 2,000,000 and 1,000,000 years ago. Early Stone Age sites have been located in a number of places, notably in the vicinity of the Victoria Falls and the Kalambo Falls. During the Middle Stone Age there appears to have been a considerable increase in the population of Zambia and the presence of artifacts similar to those used by *Homo rhodesiensis* indicates that Broken Hill man lived in Zambia about 25,000 years ago. Evidence of the activities of the San (Bushman) hunters of the Late Stone Age is to be seen in the paintings and engravings found in caves and rock shelters such as the Chifubwa Stream cave near Solwezi and the Nsalu cave in the Serenje district, while 20 San skeletons were unearthed during the Cwisho excavations. The earliest Iron Age people entered Zambia from the north about 2,000 years ago, bringing with them an economy based upon agriculture and the rearing of domestic animals. These invaders either married the earlier hunter inhabitants of Zambia or drove them into the less suitable areas of the country. Another group of farmers, quite distinct from the first Iron Age men, inhabited the high plateau country of the southern region between the 8th and 12th centuries; while about AD 800, traders from the east brought beads and cloth to the Kariba Valley and traded them for ivory.

The ancestors of the modern Tonga tribe reached Zambia early in the 2nd millennium AD, but the other modern peoples of Zambia only reached the country from the Congo and Angola in the 17th and 18th centuries. In 1798 a Portuguese trading mission from Tete reached the capital of Chief Kazembe near Lake Mweru, and then in 1835 the Ngoni, a Bantu people in flight from Zululand, crossed the Zambezi in a northeasterly direction. One section of this tribe finally settled on the Lake Nyasa–Luangwa watershed near the future site of Fort Jameson (Chipata). In the west, the Kololo (Makololo), a Sotho people, under their chief Sebituane crossed the Upper Zambezi and made themselves masters of Barotseland.

European exploration

David Livingstone also reached the Upper Zambezi in 1851 and encountered the Victoria Falls in 1855. He subsequently explored the whole Zambezi Basin and the plateau to the south of Lake Tanganyika, dying at Chitambo's village near Lake Bangweulu on May 1, 1873. During the same period Arab slave traders from Kilwa began to penetrate into Zambia in the region of Lake Nyasa and with the support of some of the tribes from east of the lake continued to terrorize the region until very close to the end of the century.

THE BRITISH SOUTH AFRICA COMPANY

The Christian missionary François Coillard established himself in Barotseland in 1884, but the only other Europeans to visit the country were a handful of hunters and prospectors until the emissaries of Cecil Rhodes arrived in Barotseland in 1890 to make a treaty with the paramount chief, Lewanika. The latter had sought Queen Victoria's protection in 1887, and, as the charter of the British South Africa Company included, though vaguely, responsibility for the lands north as well as south of the Zambezi, it was with the company that treaties were concluded by most of the chiefs in Zambia during the 1890s. The company administered the country until it became a protectorate in 1924. Only the warlike Bemba refused to negotiate, and it took several years to drive the Arabs out of the Bemba country and to pacify the Bemba. By orders in council of 1899 and 1900 the western and eastern halves of the region were administered by the company as separate protectorates called Northwestern and Northeastern Rhodesia respectively. In 1911 they were amalgamated as Northern Rhodesia, and Livingstone, near the Victoria Falls, became the first capital. Meanwhile, missionaries, miners, settlers, and traders had begun to move in, and by 1911 there were about 1,500 Europeans in the country.

The lead–zinc–vanadium mine at Broken Hill (now Kabwe) had been discovered in 1902, and copper was found at Bwana Mkubwa soon after the railway from Bulawayo (in Southern Rhodesia) reached the border of the Belgian Congo in 1909. During World War I Northern Rhodesian troops fought in the long campaign against German colonial forces, and many thousands of porters were recruited to keep them supplied.

By 1924, when the development of the Copperbelt was about to begin, there were 4,000 Europeans in the country. Under its agreement with the crown the British South Africa Company was able to retain its mineral rights, and between 1924 and 1960 it reaped a rich reward for its early and completely unprofitable endeavours. During this period, entirely as a result of its output of copper, Northern Rhodesia became a relatively progressive and prosperous country. In 1935 the capital was moved from Livingstone to Lusaka. By 1952 the annual revenue of the country had reached Z£25,000,000 and its European population (living mostly in the Copperbelt) was 43,000.

Political development had followed the pattern usual in British African dependencies with European settlers. By exerting continuous pressure over the years the Europeans had achieved considerable power in the central government, but the Africans were barely represented and did not yet have the vote. After 1930, under the system of indirect rule, Africans had been given increasing powers in local government. However, though territorial self-government was the aim, Africans were not yet thought ready to play a responsible part in national affairs.

FEDERATION

In 1953 the Federation of Rhodesia and Nyasaland was established. This move was welcomed by the European settlers in Northern Rhodesia, who had long advocated closer union with Southern Rhodesia; but it was generally opposed by Africans, who feared domination by the Europeans in Southern Rhodesia.

Constitutional progress in Northern Rhodesia continued, and gradually more power was conceded to the Africans. The nationalist movement was divided between the United National Independence Party (UNIP) and the African National Congress (ANC), and during the 1950s there was a good deal of friction between the parties. By 1962 the Africans had an elected majority in the Legislative Council. Various constitutional devices were tried in order to protect the position of the European minority, but the Africans continued to insist on the principle of "one man, one vote." The federation was dissolved on December 31, 1963.

INDEPENDENCE

In January 1964 full internal self-government was granted. UNIP, led by Kenneth Kaunda, was returned to power with a large majority; and in October 1964 Northern Rhodesia became the independent Republic of Zambia within the Commonwealth of Nations, with Kaunda as president.

In January 1965 proposals for a transitional development plan were announced, emphasizing education, defense, and agriculture. In November 1965 the Ian Smith government in Rhodesia (as Southern Rhodesia was renamed) illegally declared Rhodesia independent, and sanctions were in consequence imposed on that country by Britain. Zambia was placed in a very difficult position, since all of its communications were through Rhodesia and most of its consumer goods were supplied from there. In the following months Britain, the United States, and other friendly countries gave considerable economic aid to Zambia, including an airlift of essential oil supplies. Active steps were also taken to begin the reorientation of Zambia's communications northward to the coasts of East and West Africa. Meanwhile, Zambia had to continue to buy essential consumer goods from Rhodesia in spite of sanctions.

Zambia led the rising demands made on Britain by African countries throughout 1966 to take more effective steps, including force if necessary, to end the Rhodesian rebellion. In December the UN Security Council adopted a resolution proposed by Britain to impose mandatory sanctions on Rhodesia. Diplomatic and trade relations were not reopened until Rhodesia gained independence as Zimbabwe in 1980.

Interparty violence preceded and followed the general elections of December 1968 in which Kaunda and his party were returned to power. He won popular support in June 1969 for his proposal to amend entrenched constitutional clauses on a two-thirds majority vote and announced far-reaching political and economic measures in August. These were, perhaps, the most significant for he proposed that the state should take over all mineral rights, grant to existing companies a 25-year exploitation lease while inviting them to sell 51 percent of their shares to the state, and offer similar terms to other companies in the future. Pending a new constitution Kaunda appointed himself secretary-general of an interim executive committee and assumed emergency powers. (K.Br./K.In.)

Independent Zambia

When, in 1964, Zambia achieved its independence, it was, despite the Copperbelt, a poverty-stricken and backward country with severe class, regional, ethnic, and rural–urban divisions. During the period of the transference of power, Lozi separatism was a serious problem, and in 1964 open warfare broke out between the millenarian Lumpa (Lenshina) Church and the state.

It was not until 1964 that the British South Africa Company ceded its concession rights to copper royalties, which had, over the years, gone to enrich the company and the British treasury. The result was that very little capital had been made available for investment in the increasingly impoverished rural areas or for welfare and educational services. Despite attempts, since independence, at reform, the countryside continues to be drained of manpower as thousands of hopeful young people make their way to the towns. In the towns, however, neither housing nor welfare is adequate, and unemployment grows. Fluctuations in the world price of copper, the oil crisis, food shortages, and the war that ravaged southern Africa in the 1970s greatly aggravated Zambia's weakness. In 1972 Zambia became a one-party state under the UNIP, and in 1973 a new constitution was introduced that bolstered the power of the party. A huge para-statal corporation, the party's response to the need for industrial diversification, appeared by 1980 to have simply increased the possibilities for inefficiency and corruption. The popularity of the President declined, and dissidents were suppressed with an increasingly heavy hand. Kaunda's government sought the advice of the International Monetary Fund in the 1980s to salvage its shattered economy.

For later developments in the history of Zambia, see the *Britannica Book of the Year* section in the BRITANNICA WORLD DATA ANNUAL. (Sh.M./Ed.)

ZIMBABWE

Zimbabwe is a republic in southern Africa. It achieved majority rule and internationally recognized independence in April 1980.

Zimbabwe shares a 125-mile (200-kilometre) border on the south with the republic of South Africa, and is bounded on the southwest and west by Botswana, on the north by Zambia, and on the northeast and east by Mozambique. Its total area is 150,873 square miles (390,759 square kilometres). The capital of the nation is Harare (formerly called Salisbury).

Physical and human geography

THE LAND

Relief. Zimbabwe lies almost entirely over 1,000 feet (300 metres) above sea level. Its principal physical feature is the broad ridge running 400 miles from southwest to northeast across the whole country, from Plumtree through Gweru (formerly Gwelo) and Marondera (formerly Marandellas) to Inyanga. About 50 miles wide, this ridge ranges in altitude from 4,000 to 5,000 feet, until it eventually rises to 8,503 feet (2,592 metres) at Mt. Inyangani, the highest point in Zimbabwe, in the eastern highlands. This ridge is known as the Highveld and comprises about 25 percent of the country's total area of 150,873 square miles. On each side of this central spine, sloping down northward to the Zambezi River and southward to the Limpopo River, lies the wider plateau of the Middle Veld, which, at an altitude between 3,000 and 4,000 feet, makes up about 40 percent of the area of Zimbabwe. Beyond this again, and mostly in the south, where the Sabi, Lundi, and Nuanetsi rivers drain from the plateau into the Limpopo, lies the Lowveld, which comprises about 23 percent of the total area. The lowest point in Zimbabwe lies at an altitude of 660 feet near Dumela, where the Limpopo flows down toward Mozambique. There are no parts of Zimbabwe that can properly be called desert, although a sector northwest of Plumtree and a lengthy belt across the Lowveld in the south are severely arid.

eology

A characteristic of the landscape is its Precambrian rock, which is between 570,000,000 and 4,600,000,000 years old. The most ancient part of this rock formation, known as the Basement complex, covers the greater part of the country. About four-fifths of the Basement complex consists of granite; the Matopo Hills south of the city of Bulawayo are formed of a hard outcrop of granite and gneiss. These outcrops, known as balancing rocks, have been severely weathered by wind and water, leaving some blocks precariously balanced upon others. Elsewhere are found innumerable rounded hillocks known as kopjes and composed of ball granite. Belts of schist in the Basement complex contain the veins and lodes of most of the country's gold, silver, and other minerals.

The Great Dyke, which is up to eight miles wide and about 320 miles long, is another notable feature. The Alkali Ring complexes are igneous intrusions of volcanoes; they occur in the Sabi Valley and near Beitbridge. The Karroo System—a thick layer of sedimentary rocks consisting of shale, sandstone, and grit—covers the Zambezi Valley and the valleys of its tributaries from Hwange (formerly Wankie) southward to Bulawayo and spreads across parts of the southern Lowveld from Tuli, near the southern border, to the Sabi River.

Drainage and soils. Major faulting from southwest to northeast formed the middle Zambezi trough, which is now partially flooded by Lake Kariba. Other faulting affected the depressions of the Sabi and Limpopo rivers. As a result of subsidiary rifting a few sizable rivers drain north and south from the Plumtree–Inyanga watershed.

ɔw
rtility of
ils

The light, sandy soils developed on granite rocks are highly weathered and leached, even in the areas of lower rainfall, and do not easily retain water. Since the bulk of the rain occurs in heavy showers during a few months of the year, problems of drainage and of retaining nutrient reserves also occur. The meagre mineral reserves in the soils imply an inherently low fertility; under cultivation, productivity drops rapidly. The difficulty of cultivating these lighter soils is greatest in the black farming areas, where population pressure no longer allows land to be temporarily abandoned after cultivation; black farmers, because of a lack of capital, are also less able than white farmers to maintain the mineral fertility with manure and chemical fertilizers.

Climate. Zimbabwe lies north of the Tropic of Capricorn but enjoys subtropical conditions because of its altitude. Toward the end of the hot, dry months, which last from August to October, monsoon winds that have crossed the Indian Ocean and Mozambique cause rainfall when they meet the rampart formed by the eastern highlands. The eastern districts consequently receive the heaviest rainfall and have a more prolonged rainy season (lasting from October into April) than the rest of Zimbabwe. The altitude of the broad plateau of western Zimbabwe helps to guarantee fine weather for most of the country during the cool, dry winter months of May to August. June is generally the coolest month and October the warmest; temperature variations correspond closely to altitude. Inyanga, at about 5,500 feet in the eastern highlands, varies in temperature from a mean of 52° F (11° C) in July to one of 65° F (18° C) in October. Harare, at about 4,800 feet, has temperatures varying from 57° F (14° C) to 70° F (21° C), and Bulawayo, at 4,400 feet, has temperatures varying from 57° F (14° C) to 70° F (21° C). On a 20-year average, Harare and Bulawayo have had a year-round daily mean of eight hours of sunshine, and this average does not drop below six hours during the rainy season. (For rainfall patterns, see below *Settlement patterns.*)

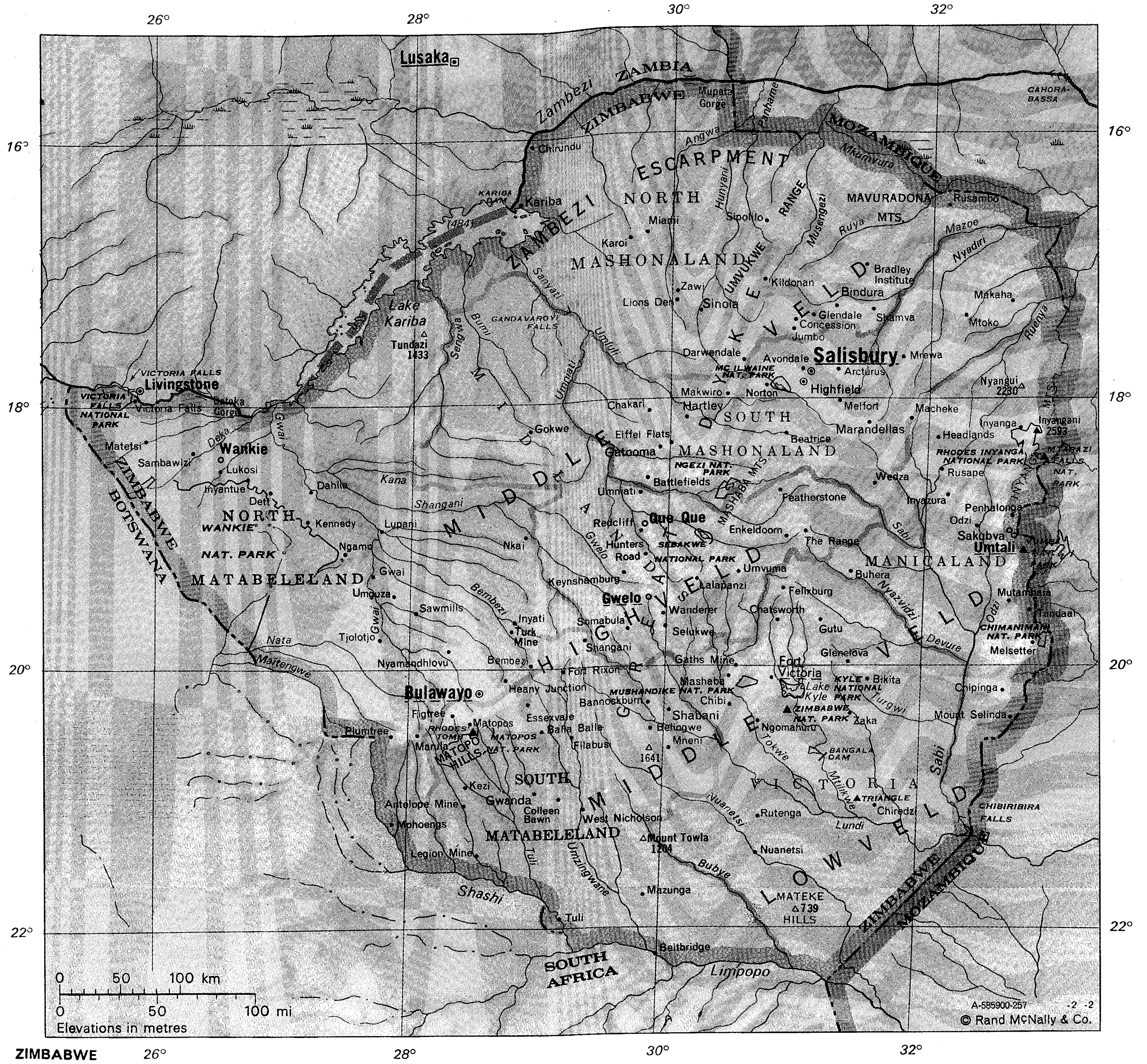
Savanna
country

Plant and animal life. Zimbabwe is predominantly savanna (tropical grassland) country, with a generous tree growth encouraged by the wet summers. The only true forests, however, are the evergreen forests of the eastern border and the savanna woodland, which includes teak, northwest of Bulawayo. Various species of *Brachystegia* (a hardwood tree up to 90 feet high with pale reddish-brown wood) are dominant in the Middle Veld and Highveld. Other common varieties include the mohobohobo (a medium-sized tree with large spadelike leaves) and the thorn tree. In the valleys of the Zambezi and Limpopo rivers the mopani, which resembles the mohobohobo, is common, together with baobab and the knobby thorn tree. Australasian eucalyptus trees have been widely introduced, predominantly on white-owned farms, where they are used as windbreaks and for fuel; Australian wattle has been planted in the eastern districts as a source of tannin. Pure grassland is uncommon but occurs particularly along the eastern border around Mandidzudzure (formerly Melsetter). Swamps are widespread on the Botswana border.

Hwange
National
Park

Cultivation of the land has resulted in the disappearance of many forms of animal life from large areas. Hwange (formerly Wankie) National Park has an area of more than 5,000 square miles and stretches from the Bulawayo–Victoria Falls railway line westward to the Botswana border. Among the flesh-eating animals found there, and occasionally elsewhere, are the lion, leopard, cheetah, serval, civet, aardvark, spotted and brown hyena, black-backed and side-striped jackal, zorilla, ratel, bat-eared fox, ant bear, and scaly anteater. Elephants are found in the northern region and giraffes in the western bushland; hippopotamuses and crocodiles live in the larger rivers. Among a great variety of hoofed and horned ruminant animals are the eland (which is immune to the deadly tsetse fly), greater kudu, blue duiker, impala, klipspringer, steenbok and grysbok, and sable and roan antelope. Snakes include mambas, boomslangs, and the black-necked cobra. Baboons, which are the bane of farmers whose crops they damage, include the Rhodesian and yellow species, as well as the chacma, the largest known baboon species. Notable among the birdlife are the martial eagle, the bateleur eagle, and the little hammerhead, which builds enormous nests and is revered as a bird of omen.

Settlement patterns. Zimbabwe may be divided into six different regions, with the amount of rainfall constituting the determining factor in land use. Of the nation's 96,500,000 acres, (39,000,000 hectares), some 1,500,000 acres in the eastern highlands, with more than 25 inches of rainfall annually, are suitable for diversified farming with cattle and plantation and orchard crops. A further 18,000,000 acres sweeping west along the central spine past Harare and to the midlands receive 20 to 25 inches of rain and are used for intensive farming of corn (maize) and tobacco and the raising of livestock. An almost equal area to the southwest, enclosing Bulawayo, receives 16 to 20 inches of rain a year; it is suitable for mixed farming and for raising livestock on a semi-intensive scale. One-third of the country, lying farther outward from the spine of Zimbabwe, mostly to the south, and receiving 14 to 18 inches of rain a year, is used for semi-extensive farming, while 25,000,000 acres in the Lowveld toward the Limpopo and Zambezi rivers, receiving less than 16 inches a year, are fit only for ranching. Finally, some 3,000,000 acres, mostly toward the Zambezi River, are unsuitable for either agriculture or forestry.



ZIMBABWE

MAP INDEX

Political subdivisions

Manicaland	19·30s	32·15e
Midlands	19·00s	29·45e
North Mashonaland	16·30s	30·00e
North Matabeleland	19·30s	28·00e
South Mashonaland	18·15s	30·45e
South Matabeleland	20·00s	28·15e
Victoria	20·54s	31·21e

Cities and towns

Antelope Mine	21·02s	28·27e
Arcturus	17·47s	31·20e
Avondale	17·43s	30·58e
Balla Balla	20·26s	29·02e
Bannockburn	20·16s	29·51e
Battlefields	18·31s	29·52e
Beatrice	18·15s	30·55e
Beitbridge	22·13s	30·00e
Belingwe	20·30s	29·53e
Bembezi	20·00s	28·56e
Bikita	20·06s	31·41e
Bindura	17·19s	31·20e
Bradley Institute	17·02s	31·27e
Buhera	19·18s	31·29e
Bulawayo	20·09s	28·36e
Chakari	18·05s	29·51e
Chatsworth	19·38s	31·13e
Chibi	20·19s	30·30e
Chipinga	20·12s	32·38e
Chiredzi	21·03s	31·45e
Chirundu	15·59s	28·54e
Colleen Bawn	21·00s	29·13e
Concession	17·22s	30·57e
Dahlia	18·35s	27·08e
Darwendale	17·43s	30·33e
Dett	18·38s	26·50e
Eiffel Flats	18·15s	29·59e
Enkeldoorn	19·01s	30·53e
Essexvale	20·18s	28·56e
Featherstone	18·42s	30·49e
Felixburg	19·29s	30·51e
Figtree	20·24s	28·21e
Filabusi	20·34s	29·20e
Fort Rixon	20·01s	29·18e
Fort Victoria	20·05s	30·50e
Gaths Mine	20·00s	30·31e
Gatooma	18·21s	29·55e
Glenclova	19·59s	31·26e
Glendale	17·21s	31·04e
Gokwe	18·07s	28·58e
Gutu	19·38s	31·10e
Gwai	19·15s	27·42e
Gwanda	20·57s	29·01e
Gwelo	19·27s	29·49e
Hartley	18·10s	30·14e
Headlands	18·14s	32·03e
Heany Junction	20·06s	28·54e
Highfield	17·50s	30·59e
Hunters Road	19·09s	29·48e
Inyanga	18·13s	32·46e
Inyantue	18·32s	26·41e
Inyati	19·39s	28·54e
Inyazura	18·43s	32·10e
Jumbo	17·28s	30·55e
Kariba	16·30s	28·45e
Karoi	16·50s	29·40e
Kennedy	18·52s	27·10e
Keynshamburg	19·15s	29·39e
Kezi	20·58s	28·32e
Kildonan	17·21s	30·37e
Lalapanzi	19·16s	30·15e
Legion Mine	21·23s	28·33e
Lions Den	17·16s	30·02e
Lukosi	18·30s	26·30e
Lupani	18·54s	27·44e
Macheke	18·05s	31·51e
Makaha	17·17s	32·37e
Makwiro	17·58s	30·28e
Marandellas	18·10s	31·36e
Marula	20·26s	28·06e
Mashaba	20·02s	30·29e
Matetsi	18·16s	25·56e
Matopos	20·24s	28·28e
Mazunga	21·45s	29·52e
Melfort	17·59s	31·19e
Melsetter	19·48s	32·50e
Miami	16·40s	29·46e
Mneni	20·38s	30·03e
Mount Selinda	20·25s	32·43e
Mphoengs	21·10s	27·51e
Mrewa	17·39s	31·47e
Mtoko	17·24s	32·13e
Mutambara	19·36s	32·33e
Ngamo	19·08s	27·32e
Ngomahuru	20·26s	30·43e
Nkai	19·00s	28·54e
Norton	17·53s	30·42e
Nuanetsi	21·22s	30·45e
Nyamandhlovu	19·50s	28·16e
Odzi	18·58s	32·23e
Penhalonga	18·54s	32·40e
Plumtree	20·30s	27·50e
Que Que	18·55s	29·49e
Redcliff	19·02s	29·46e
Rusambo	16·35s	32·12e
Rusape	18·32s	32·07e
Rutenga	21·08s	30·45e
Sakubva	18·57s	32·39e
Salisbury	17·50s	31·03e
Sambawizi	18·21s	26·16e
Sawmills	19·31s	28·02e
Selukwe	19·40s	30·00e
Shabani	20·20s	30·02e
Shamva	17·18s	31·34e
Shangani	19·47s	29·22e
Sinoia	17·22s	30·12e
Sipolilo	16·39s	30·42e
Somabula	19·41s	29·41e
Tandaai	19·36s	32·48e
The Range	19·00s	31·04e
Tjolotjo	19·47s	27·46e
Tuli	21·59s	29·15e
Turk Mine	19·45s	28·50e
Umguza	19·25s	27·51e
Umniati	18·39s	29·49e
Umtali	18·58s	32·40e
Umvuma	19·19s	30·35e
Victoria Falls	17·56s	25·50e
Wanderer	19·37s	29·59e
Wankie	18·22s	26·29e
Wedza	18·35s	31·35e
West Nicholson	21·06s	29·25e
Zaka	20·20s	31·29e
Zawi	17·13s	30·02e

Physical features and points of interest

Angwa, *river*	15·51s	30·35e
Bangala Dam	20·40s	31·15e
Batoka Gorge	17·56s	26·15e
Bembezi, *river*	18·57s	27·47e
Bubye, *river*	22·20s	31·07e
Bumi, *river*	16·50s	28·26e
Chibiribira Falls, *waterfalls*	21·14s	32·20e
Chimanimani National Park	20·00s	32·20e
Deka, *river*	18·03s	26·44e
Devure, *river*	19·50s	31·45e
Gandavaroyi Falls *waterfalls*	17·17s	29·07e
Great Dyke, *hills*	20·00s	30·00e
Gwai, *river*	17·59s	26·52e
Gwelo, *river*	18·45s	28·36e
Highveld, *physical region*	19·30s	29·50e
Hunyani, *river*	16·00s	30·35e
Inyanga Mountains	18·00s	33·00e
Inyangani, *mountain*	18·20s	32·50e
Kana, *river*	18·30s	27·22e
Kariba, Lake	17·00s	28·00e
Kariba Dam	16·30s	28·50e
Kyle, Lake, *reservoir*	20·14s	31·00e
Kyle National Park	20·14s	31·00e
Limpopo, *river*	25·15s	33·30e
Lowveld, *physical region*	21·30s	31·30e
Lundi, *river*	21·43s	32·34e
McIlwaine National Park	17·55s	30·50e

MAP INDEX (continued)

Maitengwe, *river*	19·57s	26·40e
Mashaba Mountains	18·45s	30·32e
Mateke Hills	21·48s	31·00e
Matopo Hills	20·36s	28·28e
Matopos National Park	20·33s	28·20e
Mavuradona Mountains	16·30s	31·20e
Mazoe, *river*	16·32s	33·25e
Middle Veld, *physical region*	18·30s	28·30e
	20·30s	30·30e
Mkumvura, *river*	15·55s	31·07e
Mtarazi Falls National Park	18·28s	32·47e
Mtilikwe, *river*	21·09s	31·30e
Mupata Gorge	15·38s	30·05e
Musengezi, *river*	15·55s	31·05e
Mushandike National Park	20·10s	30·38e
Nata, *river*	20·14s	26·10e
Ngezi National Park	18·40s	30·28e
Nuanetsi, *river*	22·40s	31·50e
Nyadiri, *river*	16·44s	32·33e
Nyangui, *mountain*	17·53s	32·44e
Nyazvidzi, *river*	20·00s	32·17e
Odzi, *river*	19·45s	32·24e
Rhodes Inyanga National Park	18·12s	32·45e
Rhodes Tomb	20·30s	28·30e
Ruenya, *river*	17·15s	32·58e
Ruya, *river*	16·35s	33·13e
Sabi, *river*	21·00s	35·02e
Sanyati, *river*	16·49s	28·45e
Sebakwe National Park	19·00s	30·14e
Sengwa, *river*	17·07s	28·05e
Shangani, *river*	18·41s	27·10e
Shashi, *river*	22·14s	29·20e
Tokwe, *river*	21·09s	31·30e
Towla, Mount, *mountain*	21·22s	29·52e
Triangle, *ranch*	21·00s	31·26e
Tuli, *river*	21·48s	29·04e
Tundazi, *mountain*	17·33s	28·05e
Turgwi, *river*	20·27s	32·17e
Umfuli, *river*	17·30s	29·23e
Umniati, *river*	17·30s	29·23e
Umvukwe Range	17·20s	30·40e
Umzingwane, *river*	22·12s	29·56e
Victoria Falls, *waterfalls*	17·55s	25·21e
Victoria Falls National Park	17·55s	25·40e
Vumba National Park	19·05s	32·45e
Wankie National Park	19·00s	26·35e
Zambezi, *river*	18·50s	36·17e
Zambezi Escarpment	16·20s	30·00e
Zimbabwe National Park	20·17s	30·57e

The Rhodesian Front came to power in 1962 with the promise to maintain land segregation, despite the evidence gathered in 1960 by a Rhodesian parliamentary committee that the black reserves could not support all the families with land rights under the Native Land Husbandry Act of 1951. The Front put aside the implementation of this act and attempted to establish racial division "for all time" with the Land Tenure Act of 1969. In consequence, the nationalist struggle focussed sharply upon the issue of land ownership, and a major concern for the Zimbabwe government after independence was to carry through land reform and launch large-scale settlement of black families on former white farms.

The Land Apportionment Act made no provision for blacks who chose an urban life, for towns were designated as white areas. As a result, though urban blacks now outnumber whites by four to one, blacks mostly live in rented homes in townships located some miles from city centres. The cities of Harare and Bulawayo therefore constitute studies in contrast, with impressive office buildings and quiet white suburbs partially ringed by crowded black townships. The Land Tenure Act was amended, while the civil war was still being fought, to allow blacks to purchase white farms and urban property, and after the end of hostilities residential segregation began to be significantly breached.

THE PEOPLE

he
ıona and
debele

The majority of the population of Zimbabwe speaks Shona; these are nearly four times the number of those who speak Ndebele as their first language. Both Shona and Ndebele are Bantu languages; from the time of their great southward migration, Bantu-speaking tribes have populated what is now Zimbabwe for more than 10 centuries. Those who speak Ndebele are concentrated in a circle radiating from Bulawayo, with Shona-speaking peoples beyond them on all sides—the Kalanga to the southwest, the Karanga to the east around Nyanda (formerly Fort Victoria), the Zezuru to the northeast, and the Rozwi and Tonga to the north. Generations of intermarriage have to a degree blurred the linguistic division between the Shona and Ndebele peoples.

Among the whites in Zimbabwe at independence were the descendants of the country's first European immigrants. Only about one-quarter of the adult white population, however, was born in Zimbabwe. Since World War II the white population has been trebled by heavy immigration, and more than two-thirds of the present white population have their origins in Europe, the great majority from Britain. The rest have come from South Africa.

There are several thousand Asians, forming a community that is predominantly concerned with trade. There are also Zimbabweans of mixed race, in Zimbabwe called Coloureds, who are mainly skilled and semiskilled workers. Of the whites living in rural areas, about one-quarter are Afrikaners. English is the language of government; teaching in schools is also conducted in English, except for the instruction of the youngest children in black schools.

Traditional religion

The great majority of the black population adheres to traditional religion based on reverence for ancestors. The Shona have preserved their ancient reputation for prophecy, divination, and rainmaking; they believe in Mwari, a supreme being. The stone ruins of Great Zimbabwe are regarded as a shrine of deep religious significance, as also are parts of the Matopo Hills. In the last 50 years mission schools have exercised much influence, and most of the members of the first Cabinet of independent Zimbabwe were graduates of these schools. The strongest Christian influence has been that of the Roman Catholic Church. The Anglican, Methodist, Presbyterian, Baptist, and Dutch Reformed churches are also represented. Because the Catholic Church supported nationalist aspirations, it held a position of influence in the post-independence period.

White migration patterns

Immigration and emigration. Migration has been the most important factor influencing the size and composition of the white population. Net migration figures have fluctuated in reaction to political events. In the years immediately preceding the breakup of the Federation of Rhodesia and Nyasaland, there was a net emigration of 13,000 whites; this was followed during the first 10 years of UDI by a net immigration of 40,000 (with 111,270 immigrants and 71,330 emigrants). As warfare spread after 1976, the pendulum swung again from a peak white population of 260,000 to fewer than 200,000 after independence. These net figures obscure, however, the gross turnover during 1965–79 of 132,560 immigrants and 133,864 emigrants. Even when allowance is made for the subsequent return of some emigrants, it is probable that at least half of the country's adult whites were newcomers after 1965.

Distribution of population. About one-fifth of the total population lives in Zimbabwe's 19 urban centres, nearly two-thirds of them in either Harare or Bulawayo. Among urban blacks, there is a disproportionately high number of males of working age, leaving an excess of older people, women, and children in rural areas. At least half of the black households are partly or wholly dependent on incomes earned in the wage economy.

THE ECONOMY

Agriculture, forestry, and fisheries. Agriculture contributed nearly one-fifth of Zimbabwe's national income until about 1973, when the war caused its contribution to decline. Even in years of drought, agriculture provided one-third of the country's total foreign exchange and about 35 percent of total employment in the wage economy.

Tobacco production

Economic sanctions over the 14 years of Ian Smith's Unilateral Declaration of Independence (UDI) had a broad effect on patterns of commercial agriculture. Tobacco, primarily the Virginia variety that is flue-cured and grown on white-owned farms, accounted for more than half of the gross value of output in 1965 but fell below 20 percent during UDI. There were 1,750 white farmers still growing tobacco at the time of independence, however, and the flue-cured production in 1979 was the highest since 1965.

Livestock—mainly beef and milk production—became more valuable than tobacco, but the cattle industry was hurt in the later stages of the war. Disease spread because of breakdown in dipping services, and the national beef breeding herd was reduced by more than 30 percent. Recovery was hampered somewhat by cattle rustling.

Soy bean and cotton production have increased rapidly, and cotton has become an important crop for smallholders in the Middle Veld.

Corn (maize) production was high enough in the mid-1970s to permit export but slumped so low in the late 1970s that the country imported maize from South Africa. Although the collapse of tribal agriculture during the war, one reason for the slump, was remedied in the early 1980s, commercial farmers found maize to be an unprofitable crop and cut their acreage. Wheat was a minor crop in 1965, but by the late 1970s the country was able to export to neighbouring territories. Sugar output in the Lowveld toward the Limpopo increased with the lifting of sanctions and the rise in world prices. Other crops that thrived during the diversification prompted by the sanctions were coffee and various vegetables.

The country remained dependent upon a small group of white farmers. During UDI, their numbers decreased by 900, and at independence only 5,400 whites were actively farming. In comparison, there were some 700,000 black farming families.

Mining. Although mining accounted for about 5 percent of the gross domestic product (GDP) and provided work for only an equally small percentage of the employed labour force in 1979, its significance in the economy was considerable as a major earner of foreign exchange. Direct mineral exports account for about one-third of total value.

It was the prospect of great mineral wealth—comparable to the gold deposits of the Witwatersrand in neighbouring South Africa—that attracted the first permanent European settlers in the 1890s. These great expectations faded for many years after the peak of gold production was reached in 1915. By the 1950s, however, production of the chrome mines along the Great Dyke was significant, as was that of asbestos and copper. During UDI, the value of mining output increased. The rise in gold prices in the 1970s revived gold as the country's leading export and led to the reopening in 1979–80 of more than 100 dormant mines. Nickel mining, which began in the late 1960s, continues; the Great Dyke may indeed contain more nickel than chrome ore. The known coalfields in Zimbabwe contain proved reserves of 500,000,000 tons of salable coal.

Industry. Manufacturing has been the fastest growing sector of the economy. From 1954 to 1963, Southern Rhodesia was able to rely on the resources and larger market of the Federation of Rhodesia and Nyasaland for a 150 percent increase in manufacturing output. Then, after UDI in 1965, hundreds of new manufacturing projects were begun in the effort to defeat economic sanctions by import substitution. By 1979 manufacturing contributed one-quarter of the gross national product.

Energy. Electricity and water production contributes only about 3 percent of the gross domestic product. Its principal users are industries, mines, and farms. Electrification of the railways was begun in 1980, and there has also been considerable electrification of low-cost housing in urban townships. Less than half the black homes in Bulawayo and Salisbury, however, had their own electricity at the time of independence.

During the 14 years of UDI, the consumption of power trebled, and the increase was covered by drawing heavily upon the Central African Power Corporation grid (capacity: 11,500,000 kilowatt-hours per year), with its twin stations on the Zimbabwean–Zambian shores of Kariba Gorge. The manufacturing, transportation, and construction industries consumed nearly 50 percent of the power distributed in the grid, and the mining industry about 20 percent.

Finance and trade. Finance and insurance services contribute about 6 percent of the gross domestic product. In the later stages of the civil war, the demand for commercial bank loans fell and, after 1974, the total assets of financial institutions began to decline. This decline reflected a fall in demand for hire-purchase and lease-hire facilities and a reluctance by the general public and the business sector to commit themselves to long-term credit. The principal cause of growth in the money supply has been government borrowing for increased expenditure.

Economic sanctions, which had been imposed by stages from 1966 to 1968 on both imports and exports, were lifted in December 1979. They had been widely breached, particularly in mineral exports and in the supply of petroleum, but they nevertheless strongly affected certain commodities, such as tobacco exports. Although the country's trade surplus was diminished in 1979 by the rise in oil prices, the value of exports still outpaced that of imports. As a landlocked country, Zimbabwe faces heavy freight and travel costs, as well as outflows on investment account and transfers such as pensions and migrants' remittances; these payments turn the trade surplus into a small deficit.

Private enterprise and socialism

Administration of the economy. *Private and public sectors.* The government of independent Zimbabwe moved cautiously to alter the pattern of management that it inherited from the white minority regime. The first budget of July 1980 was described by the finance minister as "conservative [with] a mild and pragmatic application of socialism." But the whites had passed on government machinery that included many levers of economic power. While the whites by inclination were wedded to a system of private enterprise, they had evolved a system of government intervention to support infant industries and maintain agricultural prices through marketing boards. The need to cushion the blows dealt by economic sanctions during UDI brought acceptance of the imposition of exchange and import controls.

Prime Minister Robert Mugabe was swift to use this apparatus in 1980 to guarantee farmers a pre-planting price for maize more than 40 percent higher than the price of 1979–80, while subsidizing the price to consumers of this staple food. To benefit foreign investors, foreign exchange controls were gently applied in the 1980 budget, 50 percent of after-tax profits were allowed to be remitted to non-resident shareholders, and new foreign investment was to be freely repatriated. The government machinery by which the Smith regime controlled the economic affairs of rural blacks was altered after independence by the establishment of district councils that took on the major task of grass roots economic development.

Taxation. The government raises nearly half of its revenue from personal and corporate income taxes which, since 1966, have been collected on a pay-as-you-earn system. About two-fifths of government revenue comes from customs and excise duties and sales taxes, and much of the rest from government borrowing and, since independence, international aid.

The independent Zimbabwe government removed sales taxes on the staple items of food and fuel for the poorest people and extended sales taxes to travel, hotel accommodations, taxis, telecommunications, and other services. It continued the former rates of personal income tax.

Consolidation of trade unions

Trade unions and employers' associations. The evolution of the trade-union movement was some two years behind the pattern of political change by 1980. The Mugabe government dealt with immediate labour problems, such as strikes for a higher minimum wage, rather than institute a thorough revision of the basic Industrial Conciliation Act of 1959. The government seemed to favour the strengthening, by mergers or amalgamation, of small unions in the same industry; the strengthening of the whole movement by the formation of a single trade-union congress from the five or six existing confederations of unions; and an arm's-length relationship of government with such a congress. Despite the large number of unions in existence, the largest sections of the labour force—the agricultural workers and domestic servants—remained outside the system.

Employers' groups, such as the Associated Chambers of Commerce of Zimbabwe and the Association of Rhodesian Industries, remained influential.

Transportation. The main road system, which is excellent, generally follows the line of white settlement along the spine of the country, with two branches north to Victoria Falls and Kariba and a network fanning out from Nyanda (formerly Fort Victoria), close to the Great Zimbabwe ruins. Wartime operations brought an improvement in certain areas, including the construction of strategic roads in the eastern highlands and near the Zambian border. The roads in white farming areas and the gravel and earth roads in the Tribal Trust Lands received barely adequate maintenance, however.

Railways

The railway closely follows the main road network; its single track has a gauge of three feet six inches. Economic sanctions, followed by the closure of the Zambian border in 1973, drastically cut rail traffic and transit revenues from Zambia's copper exports. The southern routes, however, became the Smith regime's lifeline after the British

Navy imposed an oil blockade on the Mozambique port of Beira and effectively cut the Beira–Mutare (formerly Umtali) pipeline supply in 1966. Rail links with South Africa remained intact after independence and the link with Zambia Railways was reopened. Two lines connect with lines through Mozambique to give landlocked Zimbabwe access to the ports of Maputo and Beira.

Air Zimbabwe replaced Air Rhodesia, a government-backed company that had operated only within Rhodesia and to and from South Africa. The international airport at Harare has one of the longest civil runways in the world. Seven other airports—at Bulawayo, Kariba, Gweru, Masvingo, Hwange, Buffalo Range, and Victoria Falls—can accommodate medium-sized jet aircraft.

ADMINISTRATIVE AND SOCIAL CONDITIONS

he 1979 onstitu-on

The independence constitution of Zimbabwe, which was written in London during September–December 1979, secured majority rule for Zimbabweans. Under the 1979 constitution, white voters, registered on a separate roll, elect 20 of the 100 members of the National Assembly. Although these members no longer can veto constitutional amendments, a unanimous vote is required during the first 10 years to alter the Declaration of Rights, which stipulates (among other matters) that, if land is acquired for settlement schemes, there must be "prompt payment of adequate compensation . . . remittable within a reasonable time to any country outside Zimbabwe." The British insisted that there be a constitutional head of state, a president elected by the National Assembly, and an executive prime minister, and that citizenship of Zimbabwe be automatically available to anyone who was (or had the qualifications to be) a citizen of Rhodesia immediately before independence.

There is a Senate of 40 members, half of whom are either nominees of the white Assembly members or of the Council of Chiefs. The Senate has the power to delay ordinary legislation for 90 days and constitutional amendments for 180 days.

Local governments. At the time of independence, whites controlled the municipal councils, but legislation was soon introduced to amalgamate each municipal council with the council of its surrounding township and, for the first time, black mayors were elected in 1981. Local government elections in rural areas replaced the old apparatus of district commissioners with a party-based council structure.

The political process. In the elections held in February 1980 under the 1979 constitution, Ian Smith's Rhodesian Front took all 20 seats on the white voters' roll. Of the nine parties that contested the 80 common roll seats, Bishop Abel Muzorewa's United African National Council (UANC) secured only three seats; Joshua Nkomo's Zimbabwe African People's Union (ZAPU) won 20 seats, representing nearly all in Matabeleland, and Mugabe's Zimbabwe African National Union (ZANU) won 57 seats, nearly all in Mashonaland. A total of 2,649,529 valid votes were cast, and the Commonwealth Observer Group drawn from 11 nations declared their satisfaction that the elections had been "free and fair."

Justice. The Ministry of Law and Order of Smith's period was renamed the Ministry of Justice by the independent government. The government voted to extend the state of emergency, first declared at UDI, because of unsettled conditions, mainly in rural areas.

Under the constitution of 1979, a four-member Judicial Service Commission advises the president on the appointment of judges to the High Court. High Court judges may not be removed from office except for misconduct or incapacity. The first black lawyer was appointed a High Court judge in 1980. In addition to magistrates who preside over criminal and civil litigation, other courts adjudicate on matters of African law and custom.

Education. The dismantling of Rhodesia's segregated system of schooling began less than two years before independence. The minority government had concentrated upon providing compulsory (and virtually free) education to white children between the ages of five and 15, and had left the schooling of black children in the hands of missionaries. In 1950 there were only 12 government schools for blacks, compared with 2,230 mission and independent schools. The disparities continued through the 1970s. Facilities were sharply reduced after the lower primary stage, and the civil war cut opportunities for blacks even further; from 1978, at least one-quarter of all black primary schools were closed. Enrollment fell at the University of Zimbabwe, where blacks had become a majority of the student body by 1979. Priority has been given to reopening and reequipping schools and to providing new schools in the drive toward free primary education. Because of the many opportunities in higher education offered Rhodesians by Commonwealth and other countries during UDI, the leadership of Zimbabwe was particularly well-educated.

Health and welfare. Health services were biased toward curative medicine in central hospitals. Before 1980, missionaries had the major responsibility for running rural clinics and small hospitals. After independence health allocations were increased.

A severe housing shortage in the main urban centres was aggravated by the influx of rural refugees and by the trebling of the cost of building materials during UDI. The government appealed to private contractors and individual employers to aid in the development of housing for lower paid workers.

Infant mortality

As in other Third World countries, the burden of disease is heaviest on Zimbabwe's youngest children. The infant mortality rate for the black population in malarial parts of the Zambezi Valley has been as high as 300 per 1,000, and the rate is thought to lie between 120 and 220 per 1,000 for the black population as a whole. Measles and pneumonia are major causes of death; a tuberculosis control scheme of the mid-1970s was relatively effective. Improved nutrition is increasingly seen as the most important health need.

CULTURAL LIFE

The year-round temperate climate of the Highveld has combined with the natural inclinations of the white population to produce an outdoor society. Tennis—whether on farms or at urban clubs—and bowling have many more followers than any ballet group. Happily for the cause of reconciliation, the first sport heroes after independence were the members of the all-white team that was awarded the first gold medal for women's field hockey in Olympic history at Moscow in 1980. The most famous of Rhodesian-bred writers, Doris Lessing, settled in England in 1949. In some contrast, the nationalist struggle prompted a renaissance of Shona culture. A forerunner of this renaissance (and a victim of the liberation struggle) was Herbert Chitepo, both as abstract painter and epic poet. Stanlake Samkange's novels reconstruct the Shona and Ndebele world of the 1890s, while those of the much younger Charles Mungoshi explore the clash of Shona and Western cultures in both the Shona and English languages. Folk traditions have survived in dance and pottery. The revival of sculpture has drawn on tribal religion and totems to produce some remarkable works, particularly those of Takawira and the Tengenenge school of craftsmen who sculpt in hard serpentine.

Cultural renaissance

For statistical data on the land and people of Zimbabwe, see the *Britannica World Data* section in the BRITANNICA WORLD DATA ANNUAL. (C.W.S.)

History

The remains of Stone Age cultures dating back 500,000 years have been found in Zimbabwe, and it is thought that the San, who still survive mostly in the Kalahari Desert of Botswana, are the last descendants of these original inhabitants of southern and central Africa. They were driven into the desert by the Bantu-speaking tribes during the long migrations from the north in the course of which the Bantu-speaking peoples populated much of Africa from Lake Chad to Zululand. The first Bantu are thought to have reached Zimbabwe between the 5th and 10th centuries AD. The stone ruins of Zimbabwe date mainly from about the 9th century, although the most elaborate belong to a period after the 15th century and are of Bantu origin.

PORTUGUESE EXPLORATION

The Portuguese, who arrived on the east coast of Africa at the end of the 15th century, dreamed of opening up the interior and establishing an intercoastal route to connect their eastern settlements with Angola in the west. The first European to enter Zimbabwe was probably António Fernandes, who tried to cross the continent and reached the neighbourhood of Que Que (now Kwekwe). Nearly 50 years afterward the "emperor" Mwene Matapa was baptized by a Jesuit father, and in 1569 an abortive military expedition entered the interior in search of gold.

A second great movement of the Bantu peoples began in 1830, this time from the south. To escape from the power of the great Zulu chief Shaka (Chaka) three important tribes fled northward, one of them the Ndebele, who carved out a kingdom for themselves in what is now Zimbabwe. The Ndebele were warriors and pastoralists, in the Zulu tradition, and under their formidable chief Mzilikazi they mastered and dispossessed the weaker tribes, known collectively as Shona (Mashona), who were sedentary, peaceful tillers of the land. For more than half a century, until the coming of European rule, the Ndebele continued to enslave and plunder the Shona. During this period, however, British and Afrikaner hunters, traders, and prospectors had begun to move up from the south, and with them came the missionaries. Robert Moffat visited Mzilikazi in 1857, and this meeting led to the establishment in 1861 of the first mission to the Ndebele by the London Missionary Society.

THE BRITISH SOUTH AFRICA COMPANY

In South Africa, Cecil Rhodes formed the British South Africa Company, which received its charter in October 1889. Its objects were: (1) to extend the railway from Kimberley northward to the Zambezi; (2) to encourage emigration and colonization; (3) to promote trade and commerce; and (4) to secure all mineral rights, in return for guarantees of protection and security of rights to the tribal chiefs.

European settlement

In 1890 a pioneer column set out from Bechuanaland and reached the site of the future capital of Rhodesia without incident on September 12. There, the new arrivals settled and began to lay claim to prospecting rights. The Ndebele resented this European invasion, and in 1893 they took up arms, being defeated only after months of strenuous fighting. Lobengula fled, and the company assumed administrative control of Matabeleland. In 1895 many of the pioneers were persuaded to take part in the Jameson Raid into the Transvaal and were captured and sent to England for trial. In the same year the company-administered territories, which had previously been loosely known as Zambesia, were formally named Rhodesia, by proclamation. In 1896 the Ndebele rose again. Returning from London, Rhodes met with the Ndebele chiefs and persuaded them to make peace. The Shona had at first accepted the Europeans, but they too became rebellious and the whole country was not pacified until 1897.

Economic and political development. By 1892 about 1,500 settlers from the south had arrived in Rhodesia. The railway reached Bulawayo in 1896 and the Victoria Falls in 1904. By the following year there were 12,500 settlers in the country, and in 1909 gold exports were worth more than £2,500,000. Agricultural development, however, was slower, and it was not until 1907 that steps were taken to facilitate the acquisition of land. By 1911 tobacco to the annual value of nearly £35,000 was being exported and the European population had risen to 23,600.

From the earliest years the settlers had demanded representation on the Legislative Council, which in 1903 comprised seven company officials and seven elected representatives of the settlers. In 1907 the settlers were given a majority of seats. In 1914, when the 25-year term of the company's charter was due to expire, the settlers, faced with the alternative of joining the Union of South Africa, asked for the continuation of the charter pending the grant of self-government. The British government therefore extended the charter for 10 years, with the proviso that self-government could be granted earlier if the settlers showed themselves capable of administering the country unaided.

Self-government. Immediately after World War I the pressure for self-government was resumed and a royal commission was appointed to consider the future of the territory. As a result of the commission's report a referendum of the electors among the 34,000 Europeans in the country was held in 1922, on the choice between entry into the Union of South Africa as its fifth province and full internal self-government. In spite of the offer of generous terms by the Union prime minister, Gen. J.C. Smuts, a majority voted for self-government. On September 12, 1923, Southern Rhodesia was annexed to the crown and became a self-governing colony. The British government retained control of external affairs and a final veto in respect of legislation directly affecting Africans.

The interwar period was one of material progress, with the development of a reasonably prosperous economy based on copper, gold, and other minerals, corn (maize), tobacco, and cattle. By 1953 Southern Rhodesia had a European population of 157,000 and an annual revenue of more than £28,000,000.

Lord Malvern's governing principles

The policy of Sir Godfrey Huggins (later Lord Malvern), who served as prime minister of Southern Rhodesia for 20 years, was to build a society in accord with Rhodes's dictum of "equal rights for all civilized men," one in which merit and not colour should be the test of political and economic advancement. He believed that political power should not be given to the Africans until they were sufficiently experienced to know how to exercise it in cooperation with the Europeans and thus to maintain the economic development built up over the years.

A second principle in which Lord Malvern and most other Europeans in Southern Rhodesia and Northern Rhodesia (later Zambia) profoundly believed was that the two countries should be joined together, both for their mutual economic benefit and to ensure the establishment of a powerful state based on British culture and traditions. Malvern failed to secure their amalgamation, but he supported the federation of Southern Rhodesia, Northern Rhodesia, and Nyasaland (later Malaŵi) when that solution was eventually accepted by the British in 1953.

FEDERATION

In 1957 a new electoral law was passed providing for a common roll of voters with a special roll for those with lower qualifications. At the same time there was growing political consciousness among the African population, together with increasing hostility to the idea of federation. In 1960 disturbances broke out after the arrest of some African nationalist leaders belonging to the National Democratic Party (NDP). In the following year proposals for a new constitution were strongly criticized by the NDP, which considered them insufficient to meet African demands. The NDP was subsequently banned for acts of intimidation. After that, the African nationalist movement split into two rival organizations, the Zimbabwe African People's Union (ZAPU)—led by Joshua Nkomo—and the Zimbabwe African National Union (ZANU)—led initially by Ndabaningi Sithole and later by Robert Mugabe—both of which were later banned and their leaders placed under restriction. In June 1962 the UN General Assembly called for a more liberal constitution for the territory.

The election of December 1962, when the 1961 constitution came into force, was boycotted by the African nationalists. It resulted in the defeat of the ruling United Federal Party by the more conservative Rhodesian Front, and Winston Field became prime minister. At the end of 1963 the federation was dissolved, and Southern Rhodesia reverted to its former status as colony. (K.Br./K.In.)

FROM RHODESIA TO ZIMBABWE

In Southern Rhodesia majority rule was not won easily. The white response to the collapse of federation and to African independence in the north was a massive swing to the right. In 1961 the Rhodesian Front party had been formed and voted into power on a platform of immediate white independence. This Britain was unwilling to concede, and in 1965 the Rhodesian Front, under the leadership of Ian Smith, unilaterally declared Rhodesian independence.

Despite international pressure, Britain refused to use force against the illegal regime. Economic sanctions initially boosted the Rhodesian economy because of the activities of South Africa, Portugal, and the multinational oil companies. The sanctions did, however, put considerable strain on Zambia, especially after Smith closed the Zambia–Rhodesia border in 1973 and the Benguela railway in Angola was sabotaged in 1975. In an attempt to find alternatives, new roads were constructed and in 1974 the Tanzania–Zambia railway was completed.

The banning of successive nationalist organizations and the detention and exile of their leadership led to fierce criticism of Nkomo, and by 1964 the splinter group ZANU had formed from ZAPU. Internal divisions were matched by mass disillusion, as the exiled leaders continued to appeal to Britain and the United Nations while their followers attacked each other inside the country. With the Frelimo liberation of northeastern Mozambique in 1972, a new guerrilla strategy began to make headway. At the same time, Smith agreed on a settlement with the British government that required African agreement in a referendum to be held under the supervision of the British Pearce Commission. This allowed for fresh political organization, and at the end of 1971 Bishop Abel Muzorewa formed the United African National Council (UANC) to campaign against the constitutional proposals, which were overwhelmingly rejected in May 1972.

he
nternal
ttlement"

Over the next few years, as the war began to affect the Rhodesian economy, Smith came under heavy pressure from South Africa and the West to negotiate with imprisoned black leaders. They, in turn, were under pressure from their hard-pressed neighbours to come to terms. Between 1974 and 1978 several attempts at talks broke down, and in 1976 the war was resumed with vigour by a ZANU and ZAPU alliance, the Patriotic Front. The Smith regime pursued the war with increasing ferocity, recruiting large numbers of unemployed Africans to fight in the rural areas, while concentrating white troops on defending white farming and industrial areas. The havoc of an unwinnable war and fear of Communist intervention led the Western powers and South Africa to seek to get rid of minority rule in the extreme form represented by the Rhodesian Front, and in 1979 Smith agreed to an "internal settlement": multiracial elections were held and won by the UANC. The foundations of white power were unaltered by Muzorewa's victory, however, and the war escalated.

In 1979 fresh negotiations in London ultimately led to peace, and new elections were supervised by the British in 1980. A landslide victory was won by Robert Mugabe and the ZANU wing of the Patriotic Front, and Zimbabwe was declared independent. (Sh.M./Ed.)

Mugabe's new government moved deliberately to redress inequalities of race and class, redistribute land, and promote economic development, with a one-party socialist state as its long-term goal. Nkomo and ZAPU broke with Mugabe's government, chiefly over the issue of one-party rule.

Hostility to South Africa increased markedly in independent Zimbabwe, especially in response to South African attacks against alleged African National Congress operatives on the Zimbabwe border. Zimbabwe joined the Southern African Development Coordination Conference and established diplomatic ties with several nations of the West and with China and the Soviet Union.

For later developments in the history of Zimbabwe, see the *Britannica Book of the Year* section in the BRITANNICA WORLD DATA ANNUAL. (Ed.)

BIBLIOGRAPHY

Traditional cultures. The classic reference work is I. SCHAPERA, *The Khoisan Peoples of South Africa* (1951), a review and synthesis of the literature up to 1928 with a full bibliography. J.A. ENGELBRECHT, *The Korana* (1936), is an account of a now virtually extinct Hottentot people. ELIZABETH MARSHALL THOMAS, *The Harmless People* (1959), is a popular account that conveys better than any other book in English the quality of !Kung Bushman life. A series of papers based upon sustained anthropological fieldwork among Bushmen by LORNA MARSHALL were published in various numbers of *Africa:* vol. 29, no. 4 (1959); vol. 30, no. 4 (1960); vol. 31, no. 3 (1961); and vol. 32, no. 3 (1962). Also useful are GEORGE B. SILBERBAUER, *Report to the Government of Bechuanaland on the Bushman Survey* (1965), which describes a six-year survey of the Bushmen of Botswana; RICHARD B. LEE, "What Hunters Do for a Living, or How to Make Out on Scarce Resources," in R.B. LEE and I. DEVORE (eds.), *Man the Hunter* (1968); and "!Kung Bushman Subsistence: An Input-Output Analysis," in D. DAMAS (ed.), *Contributions to Anthropology: Ecological Essays* (1969).

History. *Southern Africa to 1800: (Early man, the Stone Age, and the Iron Age)*: For an overview, see the relevant chapters in J. DESMOND CLARK (ed.), *The Cambridge History of Africa,* vol. 1, *From Earliest Times to c. 500 B.C.* (1982). R.R. INSKEEP, *The Peopling of Southern Africa* (1978); and D.W. PHILLIPSON, *The Later Prehistory of Eastern and Southern Africa* (1977), which deal with both the Late Stone Age and the Iron Age; P.S. GARLAKE, *The Kingdoms of Africa* (1978), an excellent introduction to the Iron Age in southern Africa, and *Great Zimbabwe* (1973); M. HALL and J. VOGEL, "Recent Radio-Carbon Dates from Southern Africa," *Journal of African History,* vol. 21, no. 4 (1980). (*The Portuguese in west central and east central Africa*): E.A. ALPERS, *Ivory and Slaves in East Central Africa* (1975); M.D.D. NEWITT, *Portuguese Settlement on the Zambesi* (1973); A.F. ISAACMAN, *Mozambique: The Africanization of a European Institution: The Zambesi Prazos, 1750–1902* (1972); DAVID BIRMINGHAM, *Trade and Conflict in Angola: The Mbundu and Their Neighbours Under the Influence of the Portuguese, 1483–1790* (1966); J.C. MILLER, *Kings and Kinsmen: Early Mbundu States in Angola* (1976); C.R. BOXER, *Race Relations in the Portuguese Colonial Empire, 1415–1825* (1963); J. VANSINA: *Kingdoms of the Savanna* (1966); and P.M. MARTIN, *The External Trade of the Loango Coast, 1576–1870* (1972). (*The Dutch at the Cape in the 17th–18th century*): R. ELPHICK and H. GILIOMEE (eds.), *The Shaping of South African Society, 1652–1820* (1979), an authoritative overview; R. ELPHICK, *Kraal and Castle, Khoikhoi and the Founding of White South Africa* (1977); ROBERT ROSS, *Cape of Torments: Slavery and Resistance in South Africa* (1983); S.D. NEUMARK, *Economic Influences on the South African Frontier* (1957).

Southern Africa, 1800–75: (The Mfecane and its effects): J.D. OMER-COOPER, *The Zulu Aftermath* (1966), the standard account covering southern, Central, and East Africa, the findings of which have been modified by detailed regional research. (*North of the Limpopo*): BEACH, ALPERS, ISAACMAN, ROBERTS, VAIL, and WHITE, as cited above; W.G. CLARENCE-SMITH, *Slaves, Peasants and Capitalists in Southern Angola, 1840–1926* (1979): RENÉ PÉLISSIER, *Les guerres grises: résistance et révolte en Angola, 1845–1961,* 3 vol. (1977–78). (*South of the Limpopo*): For British policy and the expansion of European settlement, see J.S. GALBRAITH, *Reluctant Empire: British Policy on the South African Frontier, 1834–54* (1963); W.M. MACMILLAN, *Bantu, Boer and Briton,* 2nd rev. ed. (1963), and *The Cape Colour Question: A Historical Survey* (1927); C.W. DE KIEWIET, *British Colonial Policy and the South African Republics, 1848–1872* (1929); and D. WELSH, *The Roots of Segregation: Native Policy in Colonial Natal, 1845–1910* (1971). For some attempt to look at the African side in this period, see L. THOMPSON, *Survival in Two Worlds: Moshoeshoe of Lesotho, 1786–1870* (1975); JEFFREY B. PEIRES, *The House of Phalo* (1981); R. ROSS, *Adam Kok's Griqua* (1976); and N. ETHERINGTON, *Preachers, Peasants and Politics in Southeast Africa, 1835–80* (1978). (*The era of mineral discoveries and the scramble for southern Africa*): C.W. DE KIEWIET, *The Imperial Factor in South Africa* (1937, reprinted 1965); T.R.H. DAVENPORT, *The Afrikaner Bond, 1880–1911* (1966); J. GUY, *The Destruction of the Zulu Kingdom* (1979), a major study of the aftermath of the Zulu War (1879) that is also invaluable for the earlier history of the Zulu kingdom; PHILIP BONNER, *Kings, Commoners and Concessionaires: The Evolution and Dissolution of the Nineteenth Century Swazi State* (1983); R. ROBINSON and J. GALLAGHER with A. DENNY, *Africa and the Victorians* (1966), has several chapters dealing with the scramble in southern Africa; D.M. SCHREUDER, *The Scramble for Southern Africa, 1877–1895* (1980), a synthesis of the subject; J. FLINT, *Cecil Rhodes* (1976), probably the best of the many biographies of this key figure; J.S. MARAIS, *The Fall of Kruger's Republic* (1961); A. JEEVES, "The Rand Capitalists and the Coming of the South African War, 1896–99," in *Canadian Historical Papers* (1973); T. PAKENHAM, *The Boer War* (1979); and P. WARWICK (ed.), *The South African War* (1980), both dealing in different ways with the causes, course, and aftermath of the war; and D. DENOON, *A Grand Illusion: The Failure of Imperial Policy in the Transvaal Colony during the Period of Reconstruction 1900–1905* (1973). For a critique of the literature on British policy in South Africa, see A. ATMORE and S. MARKS, "The Imperial Factor in South Africa: Towards a Reassessment," in *Journal of Imperial and Commonwealth History* (October 1974); and S. MARKS and S. TRAPIDO, "Lord Milner and the South Africa State," in *His-*

tory Workshop (November 1979). For the social and economic consequences of industrialization, see S. MARKS and RICHARD RATHBONE, *Industrialization and Social Change in South Africa: African Class Formation, Culture, and Consciousness, 1870–1930* (1982). (*The scramble in central Africa and the establishment of colonial society*): P. MASON, *The Birth of a Dilemma: The Conquest and Settlement of Rhodesia* (1958); J.S. GALBRAITH, *Crown and Charter: The Early Years of the British South Africa Company* (1974); I.R. PHIMISTER, "Rhodes, Rhodesia and the Rand," in *Journal of Southern African Studies,* vol. 1, no. 1 (1974), a radical reinterpretation of the connections; T.O. RANGER, *Revolt in Southern Rhodesia, 1896–97* (1967), a classic study to be read with J. COBBING, "The Absent Priesthood: Another Look at the Rhodesian Risings of 1896–7," in *Journal of African History,* vol. 18, no. 1 (1977), and, with D.N. BEACH, " 'Chimurenga': The Shona Rising of 1896–7," in *Journal of African History,* vol. 20, no. 3 (1979); L. GANN, *The Birth of a Plural Society under the British South Africa Company: Northern Rhodesia 1894–1914* (1958); R.I. ROTBERG, *Christian Missionaries and the Creation of Northern Rhodesia 1880–1924* (1965); A.D. ROBERTS, *A History of the Bemba: Political Growth and Change in North-eastern Zambia before 1900* (1973); GWYN PRINS, *The Hidden Hippopotamus: Reappraisal in African History, The Early Colonial Experience in Western Zambia* (1980); J. MCCRACKEN, *Politics and Christianity in Malawi, 1875–1940* (1977); and I. LINDEN with J. LINDEN, *Catholics, Peasants, and Cewa Resistance in Nyasaland, 1889–1939* (1974). (*Angola and Mozambique*): M.D.D. NEWITT, *Portugal in Africa: The Last Hundred Years* (1981), a lucid overview with a synopsis of the earlier period of Portuguese rule. For a critique of the literature, see W.G. CLARENCE-SMITH, "The Myth of Uneconomic Imperialism: The Portuguese in Angola, 1836–1926," in *Journal of Southern African Studies,* vol. 5, no. 2 (1979); and A.F. ISAACMAN, *The Tradition of Resistance in Mozambique: Anti-Colonial Activity in the Zambesi Valley, 1850–1921* (1976). (*The Germans in South West Africa*): P. GIFFORD and W.R. LOUIS (eds.), *Britain and Germany in Africa* (1967); H. BLEY, *Kolonialherrschaft und Sozialstruktur in Deutsch-Südwestafrika 1894–1914* (1968; Eng. trans., *South-West Africa Under German Rule, 1894–1914,* 1971); and HORST DRECHSLER, *Let Us Die Fighting: The Struggle of the Herero and Nama Against German Imperialism* (*1884–1915*) (1980; originally published in German, 1966).

Southern Africa, c. 1900 to the present: M. CHANOCK, *Unconsummated Union: Britain, Rhodesia, and South Africa, 1900–1945* (1977), a masterly account of interregional politics. For coverage of more recent events from differing viewpoints, see B. DAVIDSON, J. SLOVO, and A. WILKINSON, *Southern Africa: The New Politics of Revolution* (1976); R.W. JOHNSON, *How Long Will South Africa Survive?* (1977); and CATHOLIC INSTITUTE OF INTERNATIONAL RELATIONS (CIIR), *South Africa in the 1980s* (1980). (*South Africa and Namibia*): L.M. THOMPSON, *The Unification of South Africa, 1902–10* (1960); H.J. and R.E. SIMONS, *Class and Colour in South Africa, 1850–1950* (1969); H. ADAM and H. GILIOMEE, *Ethnic Power Mobilized* (1979); and T.D. MOODIE, *The Rise of Afrikanerdom; Power, Apartheid and the Afrikaner Civil Religion* (1975). Notable among the growing revisionist literature on South Africa's 20th-century political economy are F.A. JOHNSTONE, *Class, Race and Gold* (1976); C. BUNDY, *The Rise and Fall of the South African Peasantry* (1979); WILLIAM BEINART, *The Political Economy of Pondoland, 1860–1930* (1982); R.H. DAVIES, *Capital, State and White Labour in South Africa 1900–1960* (1979); BELINDA BOZZOLI, *The Political Nature of a Ruling Class: Capital and Ideology in South Africa, 1890–1933* (1981); DAN O'MEARA, *Volkskapitalisme: Class, Capital, and Ideology in the Development of Afrikaner Nationalism, 1934–1948* (1983); D. YUDELMAN, *The Emergence of Modern South Africa: State, Capital, and the Incorporation of Organized Labour on the South African Gold Fields, 1902–1939* (1983); and MARIAN LACEY, *Working for Boroko: The Origins of a Coercive Labour System in South Africa* (1981). For the African political response, see P. WALSHE, *The Rise of African Nationalism in South Africa: The African National Congress, 1912–1952* (1970); P.L. WICKENS, *The Industrial and Commercial Workers Union of Africa* (1978); G.M. GERHART, *Black Power in South Africa: The Evolution of an Ideology* (1978); KEN LUCKHARDT and BRENDA WALL, *Organise or Starve!: The History of the South African Congress of Trade Unions* (1980); and T. KARIS and G.M. CARTER, *From Protest to Challenge: A Documentary History of African Politics in South Africa, 1882–1964,* 4 vol. (1972–77). (*Lesotho, Botswana, and Swaziland*): LORD HAILEY, *Native Administration in the British African Territories* (1950–53), and *The Republic of South Africa and the High Commission Territories* (1963); J.E. SPENCE, *Lesotho: The Politics of Dependence* (1968); B.M. KHAKETLA, *Lesotho 1970: An African Coup under a Microscope* (1972); PALMER and PARSONS, as cited above, have substantial essays of relevance; C. MURRAY, *Families Divided: The Impact of Migrant Labour in Lesotho* (1981); and H. KUPER, *Sobhuza II, Ngwenyama, King of Swaziland: The Story of an Hereditary Ruler and His Country* (1978). (*Central Africa: Zimbabwe, Zambia, and Malaŵi*): R. GRAY, *The Two Nations* (1960); C. LEYS, *European Politics in Southern Rhodesia* (1959); G. ARRIGHI, *The Political Economy of Rhodesia* (1967); R. PALMER, *Land and Racial Domination in Rhodesia* (1977); C. VAN ONSELEN, *Chibaro: African Mine Labour in Southern Rhodesia 1900–1933* (1976), a major pioneering study in social history; T.O. RANGER, *The African Voice in Southern Rhodesia, 1898–1930* (1970); R.I. ROTBERG, *The Rise of Nationalism in Central Africa: The Making of Malawi and Zambia, 1873–1964;* G. SHEPPERSON and T. PRICE, *Independent African: John Chilembwe and the Origins, Setting and Significance of the Nyasaland Native Rising of 1915* (1958); B. PACHAI, *Land and Politics in Malawi, c. 1875–1975* (1979); R. MACDONALD (ed.), *From Nyasaland to Malawi* (1975); E. BERGER, *Labour, Race and Colonial Rule: The Copperbelt from 1924 to Independence* (1974); C. PERRINGS, *Black Mineworkers in Central Africa* (1979); P. KEATLEY, *The Politics of Partnership* (1963); W.H. MORRIS-JONES, *From Rhodesia to Zimbabwe: Behind and Beyond Lancaster House* (1980); R.S. HALL, *The High Price of Principles: Kaunda and the White South* (1969; rev. ed., 1973); and W. TORDOFF *et al.* (eds.), *Politics in Zambia* (1974). For events of the 1970s, see the publications of the South African Institute of Race Relations, the Catholic Institute for International Relations, and C. LEGUM (ed.), *Africa Contemporary Record* (annual). (*Angola and Mozambique*): J. MARCUM, *The Angolan Revolution,* vol. 1, *The Anatomy of an Explosion (1950–1962)* (1969), and vol. 2, *Exile Politics and Guerrilla Warfare (1962–76)* (1978); and F.W. HEIMER, *Der Entkolonisierungskonflikt im Angola* (1979; Eng. trans., *The Decolonization Conflict in Angola, 1974–76: An Essay in Political Sociology,* 1979). For Mozambique, VAIL and WHITE, as cited above, replace the previous sparse literature.

General works: The Cambridge History of Africa (in progress) has long chapters on southern Africa by leading authorities and places southern Africa in the context of African history; for a more condensed attempt to do the same, see P. CURTIN, S. FEIERMAN, L. THOMPSON, and J. VANSINA, *African History* (1978). D. DENOON and B. NYEKO, *A History of Southern Africa* (1972), deals with the region as a whole; while R. PALMER and N. PARSONS (eds.), *The Roots of Rural Poverty in Central and Southern Africa* (1977), contains key essays on its political economy in the 19th and 20th centuries. For South Africa, C.W. DE KIEWIET, *A History of South Africa, Social and Economic* (1941), is still a useful, though dated, introductory essay; M. WILSON and L.M. THOMPSON (eds.), *The Oxford History of South Africa,* 2 vol. (1969–71), was the first general history of all the peoples of South Africa; E.A. WALKER, *A History of Southern Africa,* 3rd rev. ed. (1962); and T.R.H. DAVENPORT, *South Africa: A Modern History* (1977), provide basic factual material; and SHULA MARKS and ANTHONY ATMORE (eds.), *Economy and Society in Preindustrial South Africa* (1980), contains essays that challenge the existing historiography, especially for the 19th century. For Central Africa, including the Portuguese colonies, see DAVID BIRMINGHAM and PHYLLIS M. MARTIN (eds.), *History of Central Africa* (1983), containing essays by acknowledged experts and replacing most previous accounts. There is, as yet, no equivalent to A.D. ROBERTS, *A History of Zambia* (1976), for either Zimbabwe or Malaŵi. For early Zimbabwe, see D.N. BEACH, *The Shona and Zimbabwe 900–1850: An Outline History* (1980). For Malaŵi, see G.J. PIKE, *Malawi: A Political and Economic History* (1968), a somewhat inadequate introduction. For Mozambique and Angola there are similarly no satisfactory overall general accounts, but see D.L. WHEELER and R. PELISSIER, *Angola* (1971); and G.J. BENDER, *Angola under the Portuguese: The Myth and the Reality* (1978). Similarly, L. VAIL and L. WHITE, *Capitalism and Colonialism in Mozambique* (1980), covers that territory's 19th- and 20th-century history in masterly and original fashion, but there is no general introductory work. General works for the rest of the region are far less satisfactory. For Namibia and Lesotho, Botswana, and Swaziland (the former High Commission Territories): J.H. WELLINGTON, *South West Africa and Its Human Issues* (1967); and R.P. STEVENS, *Lesotho, Botswana and Swaziland: The Former High Commission Territories in Southern Africa* (1967), provide general introductory accounts, though both are dated; R. FIRST, *South West Africa* (1963); and J. HALPERN, *South Africa's Hostages: Basutoland, Bechuanaland, and Swaziland* (1965), are equally dated but were considered classics of investigative journalism in their time. Knowledge of the history of southern Africa was transformed in the 1970s, and general accounts usually did not keep pace with the explosion of research in that period. Some of the new material can be found in specialist journals such as the *Journal of African History,* the *Journal of Southern African Studies,* and *African Affairs.* There is also a growing literature available comparing different aspects of U.S. and South African history: STANLEY B. GREENBERG, *Race and State in Capitalist Development: Comparative Perspectives* (1980); GEORGE M. FREDRICKSON, *White Supremacy: A Comparative Study in American and South African History* (1981); HOWARD LAMAR and LEONARD THOMPSON (eds.), *The Frontier in History: North America and Southern Africa Compared* (1981); and JOHN W. CELL, *The Highest Stage of White Supremacy: The Origins of Segregation in South Africa and the American South* (1982).

Angola. JOHN A. MARCUM, *The Angolan Revolution,* 2 vol. (1969), a history of the Angolan liberation struggle up to 1970, sympathetic to FNLA; BASIL DAVIDSON, *In the Eye of the Storm: Angola's People* (1972), on the building of a national liberation movement by MPLA; WILLIAM MINTER, *Portuguese Africa and the West* (1972), a brief history of Western influences and actions on the liberation struggle; BASIL DAVIDSON, JOE SLOVO, and ANTHONY R. WILKINSON, *Southern Africa: The New Politics of Revolution* (1976), the struggle in the Portuguese colonies, South Africa, and Rhodesia, following the Portuguese coup of 1974; EDUARDO DE SOUSA FERREIRA, *Aspectos do Colonialismo Portugues* (1974), economic and political analysis of Portuguese colonies in Africa and of links with South Africa and Namibia; HENRIQUE GUERRA, *Angola, Estrutura Economica e Classes Sociais* (1975), an account of the colonial economic structures and class formation; CARLOS ROCHA DILOLWA, *Contribuição á Historia Economica de Angola* (1978), the colonial economic history of Angola, with many statistical tables, by an MPLA economist and planner; MPLA, *Historia de Angola* (1975), a first attempt by the liberation movement to write a history of the country, for use as a textbook in schools; JOSE REDINHA, *Distribuição Étnica de Angola* (1975), brief notes on ethnic and linguistic groups, plus map of distribution; C.R. BOXER, *Race Relations in the Portuguese Colonial Empire, 1415–1825* (1963), the reality of race relations in the colonies; GROUPE AFRIQUE CENTRALE DU CEDETIM, *Angola: La Lutte continue* (1977), account and analysis of the period from the start of the liberation struggle to February 1976, by a collective of African studies experts; ALLISON BUTLER HERRICK *et al., Area Handbook for Angola* (1967), though out of date, still useful in many respects.

Botswana. *Physical geography:* HARM J. DE BLIJ, *A Geography of Subsaharan Africa* (1964); ANTHONY SILLERY, *Africa: A Social Geography,* 2nd ed. (1972); ALAN BERTRAM MOUNTJOY and CLIFFORD EMBLETON, *Africa: A New Geographical Survey* (1967); WILLIAM A. HANCE, *The Geography of Modern Africa,* 2nd ed. (1975).

Land, vegetation, and water resources: I.B. POLE-EVANS, *Report No. 2: A Reconnaissance Trip Through the Eastern Portion of the Bechuanaland Protectorate in Search of Pasture Grasses* (1930); A.J.C. MOLYNEUX, "A Contribution to the Geology of the Bechuanaland Protectorate," *Rhodesian Scientific Association,* 6:73–86 (1906), and "Prospecting in Bechuanaland," *S. Afr. Min. Engng. J.,* 51:705–707 (1940); DARRELL RANDALL, *Factors of Economic Development in the Okavango Delta* (1957).

People, institutions, culture, and history: RICHARD P. STEVENS, *Historical Dictionary of the Republic of Botswana* (1975); JACK HALPERN, *South Africa's Hostages: Basutoland, Bechuanaland and Swaziland* (1965); VERNON G. SHEDDICK, *The Southern Sotho* (1953); C.W. DE KIEWIET, *A History of South Africa: Social and Economic* (1941); ERIC WALKER, *A History of Southern Africa,* 3rd ed. (1957); RICHARD STEVENS, *Lesotho, Botswana and Swaziland* (1967); I. SCHAPERA, *A Handbook of Tswana Law and Custom,* 2nd ed. (1955, reprinted 1970), and *The Tswana* (1953); ANTHONY SILLERY, *Sechele: The Story of an African Chief* (1954), and *Botswana: A Short Political History* (1974); CHIEF TSHEKEDI KHAMA, *Bechuanaland and South Africa* (1955), and *Political Change in African Society: A Study of the Development of Representative Government* (1956).

Administration, and social, political, and constitutional affairs: ZDENEK CERVENKA, *Republic of Botswana: A Brief Outline of Its Geographical Setting, History, Economy and Politics* (1970); COLIN LEGUM *et al.* (eds.), *Africa Contemporary Record* (annual); COLONIAL OFFICE, BECHUANALAND, *Annual Report* (HMSO, to 1966), useful official reports reviewing all aspects of policy and development; *Basutoland, the Bechuanaland Protectorate and Swaziland: History of Discussions with the Union of South Africa, 1909–1939,* Cmd. 8707 (HMSO, 1952); MARY BENSON, *Tshekedi Khama* (1960), a biography; RICHARD VENGROFF, *Botswana: Rural Development in the Shadow of Apartheid* (1977).

Economy: PENELOPE HARTLAND-THUNBERG, *Botswana: An African Growth Economy* (1978); *Basutoland, Bechuanaland Protectorate and Swaziland: Report of an Economic Survey Mission* (the Morse report) (HMSO, 1960); *National Development Plan, 1970–75: National Development Plan, 1979–85* (1980).

Lesotho. *Land and water resources:* JEFFARES and GREEN, *A Water Resources Survey of Basutoland* (1951); VERNON G. SHEDDICK, *Land Tenure in Basutoland* (HMSO 1954); G.M. STOCKLEY, *Report on the Geology of Basutoland* (1947).

Demography: R.R. KUCZYNSKI, *Colonial Population* (1937); a primary source of population information about Africa.

People, institutions, culture, and history: HUGH ASHTON, *The Basuto* (1952), the accepted standard work on the Basotho people; PETER SANDERS, *Moshoeshoe, Chief of the Sotho* (1975); D.F. ELLENBERGER (comp.), *History of the Basuto,* Eng. trans. by J.C. MACGREGOR (1912, reprinted 1969); GORDON HALIBURTON, *Historical Dictionary of Lesotho* (1977); H.V. MEYEROWITZ, *A Report on the Possibilities of the Development of Village Crafts in Basutoland* (1936); JACK HALPERN, *South Africa's Hostages: Basutoland, Bechuanaland and Swaziland* (1965); VERNON G. SHEDDICK, *The Southern Sotho* (1953); C.W. DE KIEWIET, *A History of South Africa: Social and Economic* (1941); ERIC WALKER, *A History of Southern Africa,* 3rd ed. (1957); RICHARD STEVENS, *Lesotho, Botswana and Swaziland* (1967); G.P. MURDOCK, *Africa: Its Peoples and Their Cultural History* (1959); C.G. SELIGMAN, *Races of Africa,* 4th ed. (1966).

Administration, and social, political, and constitutional affairs: LORD W.M. HAILEY, *The Republic of South Africa and the High Commission Territories* (1963), *An African Survey* (1938); COLIN LEGUM *et al.* (eds.), *Africa Contemporary Record* (annual); COLONIAL OFFICE, LESOTHO, *Annual Reports* (HMSO, to 1966), useful official reports reviewing all aspects of policy and development; *Basutoland, the Bechuanaland Protectorate and Swaziland: History of Discussions with the Union of South Africa, 1909–39,* Cmd. 8707 (HMSO 1952); BASUTOLAND CONSTITUTIONAL COMMISSION, *Report* (1963); J.E. SPENCE, *Lesotho: The Politics of Dependence* (1967).

Description: AUSTIN COATES, *Basutoland* (1966), a well-written useful introduction to the country; LESOTHO, DEPARTMENT OF INFORMATION, *Lesotho: Land of Rolling Mountains and Running Streams* (1960–70).

Economy: GABRIELE WINAI STRÖM, *Development and Dependence in Lesotho, the Enclave of South Africa* (1978), covering the period between 1960 and 1975; PERCY SELWYN, *Industries in the Southern African Periphery* (1975); G.M.E. LEISTNER, *Lesotho: Economic Structure and Growth* (1966); BASUTOLAND, *Development Plan, 1963/66* (1963); GREAT BRITAIN. HIGH COMMISSIONER FOR BASUTOLAND, THE BECHUANALAND PROTECTORATE AND SWAZILAND, *Basutoland, Bechuanaland Protectorate and Swaziland: Report of an Economic Survey Mission* (1960); INTERNATIONAL BANK FOR RECONSTRUCTION AND DEVELOPMENT, *Lesotho: A Development Challenge* (1975); LESOTHO NATIONAL DEVELOPMENT CORPORATION, *Annual Report.*

Bibliography: SHELAGH M. WILLET and DAVID P. AMBROSE, *Lesotho: A Comprehensive Bibliography* (1980), extensive, classified, and well indexed, with English-language annotations.

Malaŵi. The most extensive bibliography on Malaŵi is R. BOEDER, *Annotated Bibliography of Malawi* (1979). The following works are especially recommended: S. AGNEW and M. STUBBS, (ed.), *Malawi in Maps* (1972); EDWARD ALPERS, *Ivory and Slaves in East Central Africa: Changing Patterns of International Trade to the Later 19th Century* (1975); W.J. BARBER, *The Economy of British Central Africa* (1961); B. BINNS, *A First Checklist of the Herbaceous Flora of Malawi* (1968); E. DEAN, *The Supply Responses of African Farmers: Theory and Measurement in Malawi* (1966); H. DEQUIN, *Agricultural Development in Malawi* (1969), a study of agricultural development in Malaŵi between 1890 and 1967; F. and L.O. DOTSON, *The Indian Minority of Zambia, Rhodesia and Malawi* (1968); W.J.C. GERKE and C.J. VILJOEN, *Master Plan for Lilongwe: The Capital City of Malawi* (1968), an example of urban design, land use, and traffic planning in the context of traditional Malaŵi life; I. and J. LINDEN, *Catholics, Peasants and Chewa Resistance in Nyasaland 1889–1939* (1974); MALAWI GOVERNMENT, DEPARTMENT OF CENSUS AND STATISTICS, *Malawi Population Census 1966: Final Report* (1969), includes many detailed tables; MALAWI GOVERNMENT, ECONOMIC PLANNING DIVISION, *Statement of Development Policies, 1971–1980* (1971); MALAWI GOVERNMENT, NATIONAL STATISTICAL OFFICE, *Population Census, 1977: Preliminary Report* (1978); R.J. MACDONALD (ed.), *From Nyasaland to Malawi* (1975); J. MCCRACKEN, *Politics and Christianity in Malawi, 1875–1940: The Impact of the Livingstonia Mission in the Northern Province* (1977); B. PACHAI, *Malawi: The History of the Nation* (1973), and (ed.), *The Early History of Malawi* (1972); J.G. PIKE, *Malawi: A Political and Economic History* (1968), and with G.T. RIMMINGTON, *Malawi: A Geographical Study* (1965); T.O. RANGER (ed.), *Aspects of Central African History* (1968); M. READ, *The Ngoni of Nyasaland,* 2nd ed. (1970); N.R. HINTON and S. CARRENO, *Smiling Malawi: A Tenth Anniversary* (1976), with a 20-page introduction and 100 pages of photographs; and T.C. YOUNG, *Notes on the History of the Tumbuka-Kamanga Peoples of Northern Nyasaland* (1972). Periodicals and journals published in Malaŵi include the *Malawi Journal of Social Science,* and *The Malawian Geographer,* both from the University of Malaŵi; *Malawi Economic Report,* an annual publication

of the Malaŵi government; *Financial and Economic Review,* by the Reserve Bank of Malawi; and the *Society of Malawi Journal.*

Mozambique. D.M. ABSHIRE and M.A. SAMUELS (eds.), *Portuguese Africa: A Handbook* (1969), a good survey of Mozambique's history, government, and colonial economy; A.C.G. BEST and H.J. DE BLIJ, *African Survey* (1977), an examination of the political geography of Mozambique and its neighbours; JAMES DUFFY, *Portugal in Africa* (1962), a historical overview of Portuguese colonialism; MARVIN HARRIS, *Portugal's African "Wards"* (1958), a description of the forced labour and forced cropping system, written by a well-known anthropologist; and ADRIANO MOREIRA, *Portugal's Stand in Africa,* Eng. trans. (1962), a reply by a former Portuguese Cabinet minister to the criticism of Portuguese colonial policy. Three of the best analyses of Portuguese colonialism are A.F. ISAACMAN, *Mozambique: The Africanization of a European Institution—The Zambezi Prazos, 1750–1902* (1972), and *The Tradition of Resistance in Mozambique: The Zambezi Valley, 1850–1921* (1976); and M.D.D. NEWITT, *Portuguese Settlement on the Zambezi* (1973). I. KAPLAN (ed.), *Area Handbook for Mozambique* (1977), an excellent comprehensive survey of colonialism, the independence struggle, and contemporary development issues; EDUARDO MONDLANE, *The Struggle for Mozambique* (1969), a history of resistance in Mozambique and Frelimo's war against Portugal by Frelimo's first president; EDUARDO DE SOUSA FERREIRA, *Portuguese Colonialism in Africa: The End of an Era* (1974), a brief review of Portugal's politics and attitudes toward Mozambique; R.J. HAMMOND, *Portugal and Africa 1815–1910: A Study in Uneconomic Imperialism,* on Portuguese settlement and the colonial infrastructure; and MANFRED KUDER, *Moçambique: Eine geographische, soziale und wirtschaftliche Landeskunde* (1975), and MOZAMBIQUE GOVERNMENT, *Geografia Humana de Moçambique* (1975), two excellent surveys of Mozambique's physical landscapes and the economic and social geographies on the eve of independence.

Namibia. HEINRICH VEDDER, L. FOWIE, and C.H.L. HAHN, *The Native Tribes of South West Africa* (1928, reprinted 1966), with chapters on the Bushmen, Damara, Herero, and Nama; HEINRICH VEDDER, *Das alte Südwest-afrika* (1934; Eng. trans., *South West Africa in Early Times,* 1938, reprinted 1966); A.W. HOERNLE, "The Social Organization of the Nama Hottentots of Southwest Africa," *Am. Anthrop.* 27:1–24 (1925); H.R. MACCALMAN and B.J. GROBBELAAR, "Preliminary Report of Two Stoneworking Ovatjimba Groups in the Northern Kaokoveld of South West Africa," in *Cimbebasia,* no. 13 (1965); GORDON LE SUEUR, *Germany's Vanishing Colonies* (1915), concerned with the effects on colonialism of World War I; MARY E. TOWNSEND, *Origins of Modern German Colonialism, 1871–1885* (1921) and *The Rise and Fall of Germany's Colonial Empire, 1884–1918* (1930); WILLIAM O. HENDERSON, *Studies in German Colonial History* (1962), an important work; W.W. SCHMOKEL, *Dream of Empire: German Colonialism, 1919–45* (1964), an excellent book on the period between World Wars I and II; JON BRIDGMAN and DAVID E. CLARKE, *German Africa: A Select Annotated Bibliography* (1965); ODENDAAL COMMISSION, *Report on South West Africa* (1964), published by the South African government and therefore reflecting its views; J.H. WELLINGTON, *South West Africa and Its Human Issues* (1967), an authoritative work; and INTERNATIONAL COMMISSION OF JURISTS, *Apartheid in South Africa and South West Africa* (1967).

Swaziland. Descriptive works include DUDLEY BARKER, *Swaziland* (HMSO 1965); and DOUGLAS POTT, *Swaziland: A General Survey* (1965). HILDA KUPER, *An African Aristocracy: Rank Among the Swazi* (1947), is the accepted work on the social life and institutions of the Swazi; JACK HALPERN, *South Africa's Hostages: Basutoland, Bechuanaland and Swaziland* (1965), is a useful survey written by a political observer. For administration and political and constitutional affairs, see LORD HAILEY, *The Republic of South Africa and the High Commission Territories* (1963); COLONIAL OFFICE, SWAZILAND, *Annual Reports* (to 1966); RICHARD P. STEVENS, "Swaziland Political Development," *Journal of Modern African Studies,* 1:327–350 (1963); *Lesotho, Botswana and Swaziland* (1967); DENIS V. COWEN, *Swaziland: Report on Constitutional Reform* (1961); and COLIN LEGUM and JOHN DRYSDALE, *Africa Contemporary Record,* 3 vol. (1968–71). J.F. HOLLEMAN (ed.), *Experiment in Swaziland: Report of the Swaziland Sample Survey, 1960* (1964), is a detailed scientific study of the country's physical resources.

Zambia. *General references:* AMERICAN UNIVERSITY, WASHINGTON, *Area Handbook for Zambia* (1969), a handbook covering social, economic, political, and military institutions and practices, prepared by the university's Foreign Area Studies program; and W.V. BRELSFORD (ed.), *Handbook to the Federation of Rhodesia and Nyasaland* (1960), a compendium of information about the former Federation (Northern and Southern Rhodesia and Nyasaland) covering prehistory, history, and natural resources as well as many other subjects.

Land and people: D. HYWEL DAVIES, *Zambia in Maps* (1970); J.A. HELLEN, *Rural Economic Development in Zambia 1890–1964* (1968), an inventory of the land and people and a study of regional development; GEORGE KAY, *A Social Geography of Zambia* (1967), a study of the physical and historical setting of the population and its economic activities; CENTRAL STATISTICAL OFFICE, LUSAKA, *Census of Population and Housing, 1969: First Report;* W.V. BRELSFORD, *The Tribes of Zambia,* 2nd ed. (1965), a comprehensive historical survey.

Mines and minerals: J.A. BANCROFT, *Mining in Northern Rhodesia* (1961), a history of mining exploration and development; F. MENDELSOHN (ed.), *The Geology of the Northern Rhodesian Copperbelt* (1961), a technical description; and *Mining Year Book of Zambia* (annual).

Economy and development: UNITED NATIONS (ECA, FAO), *Economic Survey Mission on the Economic Development of Zambia* (1964), a report that paved the way for the transitional development program introduced after independence; OFFICE OF NATIONAL DEVELOPMENT AND PLANNING, *First National Development Plan 1966–70* (1966); and INTERNATIONAL LABOR OFFICE, *Report to the Government of Zambia on Incomes, Wages, and Prices in Zambia: Policy and Machinery* (1969).

Zimbabwe. PHILIP MASON, *The Birth of a Dilemma: The Conquest and Settlement of Rhodesia* (1958), the best account of the early days (up to 1918) of white settlement and race relations; and T.O. RANGER, *Revolt in Southern Rhodesia, 1896–97* (1967), a full-length study, drawing from African sources, of the risings against white rule in 1896–97, with significance in terms of the modern liberation movement. LAWRENCE VAMBE, *An Ill-Fated People: Zimbabwe Before and After Rhodes* (1972), a family history which portrays the humour and sadness of occupation; COLIN LEYS, *European Politics in Southern Rhodesia* (1959), an analysis of the flow of immigrants and of the continuously rightward trend in party politics before federation; CLYDE SANGER, *Central African Emergency* (1960), describing the first years of federation and the growth of the nationalist movement in territories that have since become Zimbabwe, Malaŵi, and Zambia; NDABANINGI SITHOLE, *African Nationalism* (1959) and *Letters from Salisbury Prison* (1976), a balanced political analysis and an exhortation of a father's concern for six children scattered by the civil war; NATHAN SHAMUYARIRA, *Crisis in Rhodesia* (1965), a broad description of the racial disparities and political collisions that culminated in the unilateral declaration of independence; ROBERT BLAKE, *A History of Rhodesia* (1977), including a commentary sympathetic to the white Rhodesian leaders; MARTIN MEREDITH, *The Past Is Another Country* (1979), a detailed and objective account of the political moves inside Rhodesia from 1965 to 1979; MARTIN BAILEY, *Oilgate: The Sanctions Scandal* (1979); ROGER RIDDELL *et. al., From Rhodesia to Zimbabwe,* 8 vol. (1977–79), a thorough investigation of the major problem areas for change; and D. MARTIN and P. KOHNSON, *The Struggle for Zimbabwe* (1981), an authoritative account of the liberation movement. DIANA MITCHELL and ROBERT CARY, *African Nationalist Leaders in Rhodesia* (1977), and DIANA MITCHELL, *Who's Who 1980,* provides invaluable background on the careers of the new leaders.

Of official reports published in Salisbury, the following retain more than historical interest: *Second Report of the Select Committee on the Resettlement of Natives* (1960), covering the heart of the dispute over landholdings and important recommendations on land apportionment reform; *Final Report of the 1962 Census of Africans in Southern Rhodesia* (1964), a basic source of information on socioeconomic conditions; *Census of Population, 1969* (1971); and *Report of the (Phillips) Advisory Committee on the Development of the Economic Resources of Southern Rhodesia, with Particular Reference to the Role of African Agriculture* (1962), containing details of the country's resources and potentialities. Three reports published in London cover more recent historical landmarks: *Report of the Commission on Rhodesian Opinion* (1972), known as the Pearce Report, a critically important state document; *Report on the Supply of Petroleum and Petroleum Products to Rhodesia* (1978), known as the Bingham Report, a painstaking but incomplete investigation of circumventions of sanctions by oil companies and connivance by British officials; and *Southern Rhodesian Elections* (1980), the report of the Commonwealth Observer Group, an illuminating picture of how the main actors behaved under strain in the final scenes before independence.